# The Workshop Programme

(Time assignment per presentation: 20min plus 5min for questions)

9:00 ***Welcome Note*** - *Anna Samiotou, ESTeam*

***Commercial Applications*** - *Chair: Gudrun Magnusdottir, ESTeam*

**Alignment of Parallel Texts for Populating MT & TM Databases**
Anna Samiotou, Lambros Kranias, George Papadopoulos, Marita Asunmaa, Gudrun Magnusdottir. *ESTeam AB, Sweden/Greece*
*Speakers*: Anna Samiotou, Lambros Kranias

**Comparing Rule-based and Statistical MT Output**
Gregor Thurmair. *Linguatec, Germany*
*Speaker:* Gregor Thurmair

***R&D Applications***

***MT Evaluation*** - *Chair: Yorick Wilks, University of Sheffield*

10:15 **Ground Truth, Reference Truth & "Omniscient Truth" – Parallel Phrases in Parallel Texts for MT Evaluation**
M.Vanni*****, C.R. Voss*****, C. Tate***** [§]. ******Multilingual Computing Group, Army Research Lab, Adelphi, MD, US.* [§]*Dept. of Mathematics, University of Maryland, US*
*Speaker:* Michelle Vanni

**Gathering Empirical Data to Evaluate MT from English to Portuguese**
Diana Santos*****, Belinda Maia[§], Luís Sarmento[¶]. **Linguateca, Oslo, SINTEF ICT, Norway.* [§]*Fac. de Letras, Univ. do Porto, Portugal.* [¶]*Linguateca, Porto, FLUP, Portugal*
*Speaker:* Diana Santos

**Compiling and Using a Shareable Parallel Corpus for MT Evaluation**
Debbie Elliot, Eric Atwell, Anthony Hartley. *School of Computing and Centre for Translation Studies, University of Leeds, UK*
*Speaker:* Debbie Elliot

11.05-11:25 *Coffee Break*

11:25 ***Alignment*** *- Chair: Lambros Kranias, ESTeam*

**Improving Word Alignment in an English-Malay Parallel Corpus for MT**
Suhaimi Ab Rahman, Normaziah Abdul Aziz. *Language Engineering Research Lab, MIMOS, Kuala Lumpur, Malaysia*
*Speaker:* Suhaimi Ab Rahman

**Alignment of Parallel Corpora Exploiting Asymmetrically Aligned Phrases**
Patrik Lambert, Núria Castel. *TALP Research Center, Universitat Politécnica de Catalunya, Barcelona, Spain*
*Speaker:* Patrik Lambert

**Sentence Alignment in Parallel, Comparable and Quasi-Comparable Corpora**
Percy Cheung, Pascale Fung. *Human Language Technology Center, Dept. of Electrical & Electronic Engineering, HKUST, Hong Kong*
*Speaker:* Pascale Fung

13:00 ***Semantic Annotations*** *- Chair: Nicoletta Calzolari, Istituto di Linguistica Computazionale del CNR*

**PropBanking in Parallel**
Paul Kingsbury, Nianwen Xue, Martha Palmer. *Dept. of Computer and Information Science, University of Pennsylvania, US*
*Speaker:* Paul Kingsbury

**Browsing Multilingual Information with the MultiSemCor Web Interface**
Marcello Ranieri, Emanuele Pianta, Luisa Bentivogli. *ITC-irst, Trento, Italy*
*Speaker:* Emanuele Pianta

13:30-15:00 *Lunch Break*

15:00 ***Surface Structure*** *- Chair: Gregor Thurmair, Linguatec*

**Application of Translation Corresponding Tree (TCT) Annotation Schema in Example-Based MT**
Wong Fai*****, Hu Dong Cheng*****, Mao Yu Hang*****, Tang Chiwai[§], Dong Mingchui[§]. *******Speech and Language Processing Research Center, Tsinghua University, Beijing, China.* [§]*Faculty of Science and Technology of University of Macao SAR*
*Speaker:* Fai Wong

**Improving Word Alignment Quality using Linguistic Knowledge**
Bettina Schrader. *Cognitive Science Doctorate Programme, Institute for Cognitive Science, University of Osnabruck, Germany*
*Speaker:* Bettina Schrader

**Using Comparable Corpora for Discovering Universals in Surface Structure**
John Elliott. *Computational Intelligence Research Group, School of Computing, Leeds Metropolitan University, UK*
*Speaker:* John Elliott

**Exploiting Parallel Corpora for Monolingual Grammar Induction – A Pilot Study**
Jonas Kuhn. *The University of Texas at Austin, Dept. of Linguistics, US*
*Speaker:* Jonas Kuhn

**A Multilingual Parallel Parsed Corpus as Gold Standard for Grammatical Inference Evaluation**
Menno van Zaanen*****, Andrew Roberts[§], Eric Atwell[§]. **Tilburg University, the Netherlands,* [§]*University of Leeds, UK*
*Speaker:* Menno van Zaanen

16:40-17:00 *Coffee Break*

17:00 ***Interactive Panel –*** *Chair: Anna Samiotou, ESTeam*

**True or False?: Every Serious Multilingual Application Needs a Parallel or Comparable Corpus**
*Panelists:* Nicoletta Calzolari, Gudrun Magnusdottir, Gregor Thurmair, Yorick Wilks and others to be announced

# Workshop Organisers

**Anna Samiotou**, R&D Manager ESTeam AB, Sweden/Greece
**Gudrun Magnusdottir**, Managing Director ESTeam AB, Sweden/Greece
**Lambros Kranias**, Chief Technical Manager ESTeam AB, Sweden/Greece

# Workshop Programme Committee

**Dr. Lambros Kranias**, Chief Technical Manager ESTeam AB, Sweden/Greece
**Dr. Nicoletta Calzolari**, Director, I stituto di Linguistica Computazionale del CNR, Italy
**Dr. Gregor Thurmair**, R&D Manager Linguatec, Germany
**Dr. Yorick Wilks**, Professor, University of Sheffield, UK
**Dr. Eduard Hovy**, Information Sciences Institute, University of Southern California, US
**Gudrun Magnusdottir**, Managing Director ESTeam AB, Sweden/Greece
**Anna Samiotou**, R&D Manager ESTeam AB, Sweden/Greece
**Dr. Khalid Choukri**, CEO ELDA/ELRA, France

# Table of Contents

*Commercial Applications*

**Anna Samiotou, Lambros Kranias, George Papadopoulos, Marita Asunmaa, Gudrun Magnusdottir,** *Exploiting of Parallel Texts for Populating MT & TM Databases……………………………………………………………………………………1*
**Gregor Thurmair**, *Comparing Rule-based and Statistical MT Output……………………5*

*R&D Applications*
*MT Evaluation*

**M.Vanni, C.R. Voss, C. Tate**, *Ground Truth, Reference Truth & "Omniscient Truth"– Parallel Phrases in Parallel Texts for MT Evaluation……………………………………… 10*
**Diana Santos, Belinda Maia, Luís Sarmento**, *Gathering Empirical Data to Evaluate MT from English to Portuguese…………………………………………………………………14*
**Debbie Elliot, Eric Atwell, Anthony Hartley**, *Compiling and Using a Shareable Parallel Corpus for MT Evaluation……………………………………………………………… 18*

*Alignment*

**Suhaimi Ab Rahman, Normaziah Abdul Aziz**, *Improving Word Alignment in an English-Malay Parallel Corpus for MT…………………………………………………………  22*
**Patrik Lambert, Núria Castel**, *Alignment of Parallel Corpora Exploiting Asymmetrically Aligned Phrases………………………………………………………………  26*
**Percy Cheung, Pascale Fung**, *Sentence Alignment in Parallel, Comparable and Quasi-Comparable Corpora……………………………………………………………………………… 30*

*Semantic Annotations*

**Paul Kingsbury, Nianwen Xue, Martha Palmer**, *PropBanking in Parallel…………  34*
**Marcello Ranieri, Emanuele Pianta, Luisa Bentivogli**, *Browsing Multilingual Information with the MultiSemCor Web Interface…………………………………………  38*

*Surface Structure*

**Wong Fai, Hu Dong Cheng, Mao Yu Hang, Tang Chiwai, Dong Mingchui**, *Application of Translation Corresponding Tree (TCT) Annotation Schema in Example-Based MT……………………………………………………………………………………   42*
**Bettina Schrader**, *Improving Word Alignment Quality using Linguistic Knowledge…46*
**John Elliott**, *Using Comparable Corpora for Discovering Universals in Surface Structure…………………………………………………………………………………………… 50*
**Jonas Kuhn**, *Exploiting Parallel Corpora for Monolingual Grammar Induction – A Pilot Study…………………………………………………………………………………………  54*
**Menno van Zaanen, Andrew Roberts, Eric Atwell**, *A Multilingual Parallel Parsed Corpus as Gold Standard for Grammatical Inference Evaluation……………………    58*

# Exploitation of Parallel Texts for Populating MT & TM Databases

**Anna Samiotou, Lambros Kranias, George Papadopoulos, Marita Asunmaa, Gudrun Magnusdottir**

ESTeam AB
Sikelianou 8, 146 71 Athens, Greece
esteam@otenet.gr

**Abstract**

Parallel texts are an important resource for applications in multilingual natural language processing and human language technology. This paper presents a method for exploiting available parallel texts, both human translated and revised machine translated texts in order to populate machine translation and translation memory databases.

## 1. Introduction

Parallel texts play an important role in Machine Translation (MT) and multilingual natural language processing. They are rich resources for development of monolingual, bilingual and multilingual resources both for new language pairs and for existing language pairs for a specific domain to be used in a number of natural language processing applications, for automatic lexical acquisition (e.g. Gale and Church, 1991; Melamed, 1997), etc.

This paper presents a method for populating MT and Translation Memory (TM) databases by exploiting parallel texts. The method deploys selected legacy data from the domain(s) under investigation and available parallel texts both human translated and revised machine translated texts. The software used is the ESTeam Translator© (ET) software[1] (ESTeam AB, 2004), a data-driven multilingual translation software product which integrates MT and TM technology to produce a full translation in one or multiple languages.

Current applications using the presented method include the creation of new LRs for the languages of the new members of the EU and their linking to all the existing EU languages as well as the creation of LRs for the translation needs of the Athens Organising Committee for the Olympic Games 2004.

The paper is organised as follows: Section 2 gives an overview of the methodological approach for processing parallel texts. Section 3 provides a brief outline for the application of the method in a commercial project. Finally, Section 4 concludes the paper.

## 2. Processing Parallel Texts

Translated data is a rich resource to solving translation problems. This resource has yet to be explored to its full extent (Isabelle et al., 1993). ESTeam applies pre-processing on a monolingual level as well as alignment in order to domain-tune lexical resources as well as extract translation equivalents on multiple levels.

### 2.1 Pre-processing

The monolingual data is structured into domains and analysed in three processing levels, that is, sentences, sub-sentences and words (tokenisation). The segmented text is sorted per level according to the frequency of occurrence and then, words and frequent collocations can be imported into the MT lexicon and sentences and sub-sentences into the TM database.

The processing on the monolingual level of the parallel texts is important since the monolingual data provides a resource for extensive multilingual linking. ET uses monolingual data to map to any other language once the data becomes available through resources or translation interaction.

Any general purpose lexical resource lacks information about domain. The information on the frequency of the units gives indications within the domains on which units have to be translated with priority (i.e. the high frequent ones). It also indicates which units are likely to be incorrect such as misspellings coming from wrong typing or scanning errors (i.e. the very low frequent ones) and this is judged on both frequency and similarity criteria. This information is used to automatically structure the lexical data for any domain when building the multilingual lexica (see example in *Figure 1*).

| *Language* | *Unit* | *Domain* | *Frequency* |
|---|---|---|---|
| French | fils | Computer/Textiles | 1011/741 |
| English | threads | Computer/Textiles | 14/573 |
| English | yarns | Computer/Textiles | 1/620 |
| English | wires | Computer/Textiles | 994/0 |

Figure 1. Example of Domain Tuning

In MT any monolingual data is deployed as a target language resource. When a source unit has multiple translations into another language, frequency information relating to the context of the target units gives indication on which translation alternatives MT automatically selects, i.e. the stronger the statistical indication is, the more likely it is to be selected (see example in *Figure 1*). ET also calculates the frequency of the translation links by combining source and target frequencies per domain (see example in *Figure 2*).

| *French⇔English* | *Domain* | *Link Frequency* |
|---|---|---|
| fils ⇔ threads | Computer/Textiles | 1025/1314 |
| fils ⇔ yarns | Computer/Textiles | 1012/1361 |
| fils ⇔ wires | Computer/Textiles | 2001/741 |

Figure 2. Example of Statistical Disambiguation

[1] http://www.esteam.gr

Context statistics are also calculated on the monolingual data, in order to assign weights on the co-occurrence of words and contribute to the word sense disambiguation within the same domain. In the examples in *Figures 1 & 2*, the English units *threads* & *yarns* win over *wires* as translations of the French unit *fils* in the Textiles domain. If the input French unit is: *fils de coton* and the context statistic model run on the monolingual data had calculated:

- *cotton threads* (100)
- *cotton yarns* (5)

then the *cotton threads* wins.

The more correct legacy monolingual data in the TM the better when using the ET, because it serves for target language verification (TLV), i.e. the machine translation result is automatically post-edited by the target language TM data (sentences and/or sub-sentences) based on a number of criteria permitting actions such as deletion or addition of functional units, changing word order and morphological variations. Example:

- input French source unit for translation:
    - *fils de coton*
- suggested translations in English:
    - *threads of cotton*
    - *yarns of cotton*
- existing units in the English TM:
    - *cotton threads*

=> TLV disambiguates and post-edits

## 2.2 Alignment

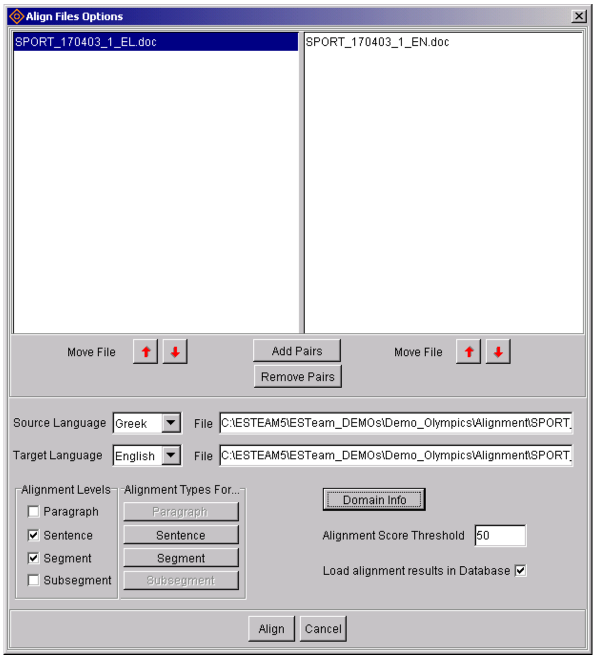
ET Aligner aligns parallel texts in different languages and at sentence and sub-sentence level (Kranias, 1995). The ET Aligner requires file, paragraph or sentence aligned parallel text. Assuming that a document is a hierarchical structure where the top level is the document itself and the deeper levels are paragraphs, sentences, sub-sentences and finally words, the ET Aligner, takes as input two parallel documents and automatically aligns them at the aforementioned deeper levels.

Alternatively, the ET Aligner processes pre-aligned documents at a given level and aligns them at a deeper level (e.g. if the given level is paragraph then it further processes at sentence and sub-sentence level). *Figure 3* displays the ET Aligner user interface. The supported format of the input documents is plain text in UTF-8 encoding, html, Microsoft Word document and TMX. The output results are in TMX format (see *Figure 4*)

At each level the ET Aligner uses a Dynamic Programming algorithm in order to detect the optimal text unit correspondences. The Aligner evaluates a number of criteria, mainly statistical and lexical information, in order to produce corresponding text unit pairs at each level, such as:

- the number of words per unit
- the number of characters per unit
- existence of strings such as numbers and dates
- special treatment of non-content words such as articles and prepositions
- special treatment of characters such as parenthesis and square brackets
- advanced lexicon look-up

The user can specify the criteria to be used and assign a weight to each criterion. Based on the previous, the Aligner assigns a reliability score to each produced aligned pair.



SPORT_170403_1_EL.doc

SPORT_170403_1_EN.doc

Source Language Greek File C:\ESTEAM5\ESTeam_DEMOs\Demo_Olympics\Alignment\SPORT_

Target Language English File C:\ESTEAM5\ESTeam_DEMOs\Demo_Olympics\Alignment\SPORT_

Alignment Levels: □ Paragraph ☑ Sentence ☑ Segment □ Subsegment

Alignment Score Threshold 50

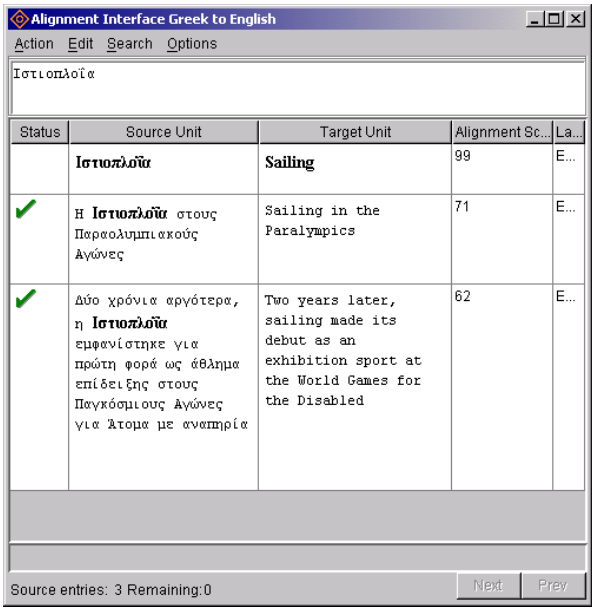
Load alignment results in Database ☑

Figure 3. ET Aligner User Interface

The alignment results are imported in a separate database, the ALIGN database. High quality alignment results are directly imported in the TM and/or MT databases. Medium and possibly low quality alignment results can be browsed and edited through the user-friendly ET Alignment Browser & Editor (see *Figure 5*) and the accepted and/or edited by the user results are imported in the TM.

```
<tu tuid="1-1" segtype="sentence">
    <prop type="x-ORGN">EL.doc_EN.doc.tmx</prop>
    <prop type="x-DOMN">0~Olympics~*/</prop>
    <prop type="x-ALGN">,01,,.</prop>
    <prop type="x-VALD">71</prop>
    <tuv lang="EL">
        <seg>Η Ιστιοπλοΐα στους Παραολυμπιακούς Αγώνες</seg>
    </tuv>
    <tuv lang="EN">
        <seg>Sailing in the Paralympics</seg>
    </tuv>
</tu>
```

Figure 4. Example of Alignment Results in TMX



Ιστιοπλοΐα

| Status | Source Unit | Target Unit | Alignment Sc... | La... |
|---|---|---|---|---|
| | **Ιστιοπλοΐα** | **Sailing** | 99 | E... |
| ✓ | Η **Ιστιοπλοΐα** στους Παραολυμπιακούς Αγώνες | Sailing in the Paralympics | 71 | E... |
| ✓ | Δύο χρόνια αργότερα, η **Ιστιοπλοΐα** εμφανίστηκε για πρώτη φορά ως άθλημα επίδειξης στους Παγκόσμιους Αγώνες για Άτομα με αναπηρία | Two years later, sailing made its debut as an exhibition sport at the World Games for the Disabled | 62 | E... |

Source entries: 3 Remaining:0

Figure 5. ALIGN Database Browser & Editor

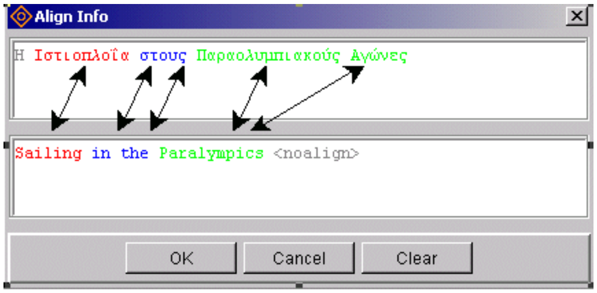
The ET Alignment Browser & Editor offers to the user full control over the alignment results which are stored in the ALIGN database. Its main features include:

- Various search modes (full/fuzzy match, word(s) in context, dynamic searches, combined source-target searches) for browsing the database contents
- Insert, Modify, Delete actions on selected contents
- Controlled global text replacements
- Logging of all user actions
- Dynamic Import/Export of the database contents to the TM

## 2.3 Word-Alignment Information

In the ET system, word-alignment information is available, through the alignment process (Meyers 1998, Ahrenberg et al, 2000) by the use of an MT lexicon of words and phrases. Word-alignment information defines the translation links between words of reference-SL and reference-TL text units (the TM pair), in other words it defines which word/phrase of the $S_{ref-SL}$ translates to which word/phrase of the $S_{ref-TL}$ (and can, in general, include phrases with non-consecutive words).

The MT lexicon defines the relevance of two text units being compared, by defining translation links between their words, and then puts a marker on the corresponding word-alignment information to be later used for the application of Fuzzy Match Post Editing (Kranias & Samiotou, forthcoming). Of course, as referred to in (Melamed, 2000): "bitext correspondence is typically only partial – many words in each text have no clear equivalent in the other text." *Figure 6* shows an example of word-alignment information which is originally displayed in different colours but due to the black and white printing we provide it with arrows.



Η Ιστιοπλοΐα στους Παραολυμπιακούς Αγώνες

Sailing in the Paralympics <noalign>

Figure 6. Example of Word-Alignment Information

## 2.4 Multilingual Linking

Multilingual linking is a unique feature of the ET which automatically connects, under defined conditions, entries through a common language entry. More specifically, it generates multilingual indirect links (i.e. links that are not imported as such through the user interface) for entries of a language that have no direct links (i.e. links that have been imported as such either manually or from a file or as alignment results) to other languages. This is possible due to the fact that the ET system is not pair-based but fully multilingual. For example, if (1) and (2) translation links are imported in the database then (3) translation link is automatically generated:

(1) Greek ⇔ English
Ολυμπιακοί Αγώνες ⇔ Olympic Games

(2) French ⇔ English
Jeux Olimpiques ⇔ Olympic Games

(3) Greek ⇔ French
Ολυμπιακοί Αγώνες<->Jeux Olympiques

## 2.5 Machine Translation

The units that have been left untranslated due to low alignment scores can be exported and machine translated by ET, with both MT and TM activated, using at least the already imported alignment results. If fuzzy matches are located then the system suggests its target language equivalent as the translation of the input unit. When no fuzzy match can be located for all or part of the input units, MT processing is activated to contribute in the translation of the remaining untranslated input unit. The MT results are automatically post edited by the TLV feature (see section 2.1) and imported, by filtering out the units that do not exist in the target pool of TM data.

## 3. The Olympic Games 2004 Project

The translation department of the Athens Organising Committee for the Olympic Games 2004 (ATHOC), selected ET to build TM data for Greek, English and French from the parallel texts they had previously translated, in order to generate as much feedback as possible form their legacy data. The legacy parallel texts were first processed on a monolingual level. Sentences and sub-sentences in English, French and Greek where imported in the TM together with their frequency of appearance information. A statistical repetition analysis on these texts indicated that the texts were quite repetitive, with a rate of 46% on both sentence and sub-sentence level.

Then, the legacy parallel texts were aligned. There were approximately 1,000 parallel document text pairs which resulted in approximately 15,000 TM sentences and 10,000 sub-sentences for each language pair. Source units were linked to one or multiple translations. More links where automatically generated through the multilingual linking feature of the ET.

ATHOC has been using ET in production since July 2003.

## 4. Conclusions

Parallel texts are a valuable resource for processing and extracting information for the translation process. ESTeam has proven that these resources are fully exploitable to improve any translation scenario where data is available. ESTeam has yet to explore the potential of parallel i.e. TM data as organisational and multilingual resource for knowledge representation.

## References

Ahrenberg L., Andersson M. & Merkel M. (2000), A knowledge-lite approach to Word Alignment. In Veronis, J., Parallel Text Processing: Alignment and Use of Translation Corpora.  (Kluwer Academic, 2000).

ESTeam AB. (2004). ESTeam Translator© White Paper. *URL: www.esteam.gr*

Gale, William A. and Kenneth W. Church. (1991). Identifying word correspondences in parallel texts. In Fourth DARPA Workshop on Speech and Natural Language, Asilomar, California.

Isabelle, Pierre, M. Dymetman, G. Foster, J-M. Jutras, E. Macklovitch, F. Perrault, X. Ren and M. Simard. (1993). Translation Analysis and Translation Automation. In Proceedings of the Fifth International Conference on Theoretical and Methodological Issues in Machine Translation, Kyoto, Japan

Kranias L. (1995). A New Optimal Algorithm for the Solution of a Generalised Assignment Problem - Application in Automatic Text Alignment. Proceedings of the International Conference on Systems, Man & Cybernetics, Vancouver, Canada

Kranias Lambros, Samiotou Anna. (2004). Automatic Translation Memory Fuzzy Match Post-Editing: A Step beyond Traditional TM/MT Integration. Proceedings of LREC2004, Lisbon, Portugal.

Melamed, I. Dan. (1997). Automatic discovery of non-compositional compounds in parallel data. In Proceedings of the 2nd Conference on Empirical Methods in Natural Language Processing (EMNLP-97), Brown University.

Meyers A., Kosaka M. and Grishman R. (1998). A Multilingual Procedure for Dictionary-Based Sentence Alignment. Proceedings of ACL-COLING-98, Montreal, Canada.

# Comparing Rule-based and Statistical MT Output

**Gregor Thurmair**

linguatec
Gottfried Keller Str. 12
D 81245 Munich
g.thurmair@linguatec.de

**Abstract**

This paper describes a comparison between a statistical and a rule-based MT system. The first section describes the setup and the evaluation results; the second section analyses the strengths and weaknesses of the respective approaches, and the third tries to define an architecture for a hybrid system, based on a rule-based backbone and enhanced by statistical intelligence.

This contribution originated in a project called “Translation Quality for Professionals” (TQPro)[1] which aimed at developing translation tools for professional translators. One of the interests in this project was to find a baseline for machine translation quality, and to extend MT quality beyond it. The baseline should compare state-of-the-art techniques for both statistical packages and rule-based systems, and draw conclusions from the comparison. This paper presents some insights into the results of this work.

## 1 Baseline

The experiment was to compare the state-of-the-art quality of MT, and it used a current statistical MT package and a commercial rule-based MT system.
The material was provided by SAP; it consisted of Translation Memory material, German to English, more than 100.000 segments in the domain of the R/3 system, to have sufficient training data for a statistical package.

### 1.1 Statistical MT

The statistical analysis and translation was done by the team of RTH Aachen; this team had the best results in the Verbmobil project (Vogel et al. 2000) and is a leading center of statistical MT in Europe (Och et al. 2003).

**Setup**

The data were processed as follows: After a preprocessing step, the material was split into a training corpus (with 1.068 mio German and 1.128 mio English tokens, representing 44.400 German and 26.600 English types, respectively). This was used as input for the alignment template SMT system to train the MT.
A test corpus (5% of the corpus) was then analysed, of which all sentences of (randomly) of 14 tokens of length and containing no unknown words were selected. This resulted in 68 sentences.

**Evaluation**

These sentences were evaluated by splitting them into three categories:

- **grammatical**: This means the sentences are syntactically correct, and convey the content.
- **understandable**: This means the sentences are incorrect but still convey the content (without reference to the source text).
- **wrong**: This means that the sentences cannot be understood without reference to the source text.

Such an evaluation scheme is a common standard in commercial MT development, often used for quality assessment[2].
About 10% of the resulting 68 sentences contain ill-formed input (incorrect German sentences: segmentation, agreement, and syntactic errors), which is a realistic figure. With the translations, a reference human translation (resulting from the SAP memory production) is available.
The resulting translation quality is as follows:

| | | |
|---|---|---|
| grammatical | 16 | 23,5% |
| understandable | 31 | 45,6% |
| wrong | 21 | 30,9% |

It can be seen that there is a significant amount of understandable results, while the really good and really bad sentences are less frequent. This underlines the robustness of such an approach. Together the good + understandable sentences are close to 70%. It should be noted, however, that from a practical point of view, understandable sentences need to be post-edited, while for grammatical sentences this is not necessarily the case.

**Improvements**

The authors propose some improvements to these results like: morphological analysis of German noun compounds, special treatment of variable and product names, lookup of (manual) lexicon (cf. Nießen/Ney 2000).
Such improvements point into the direction of creating a hybrid system, with statistical basis and additional linguistic features to improve the statistical machinery.

[1] This project (IST-1999-11407) has as partners: SAP, Lotus Ireland, SailLabs, and CST on the development side, and CAT technologies and Logoscript on the user and testing side. Details are given in (Thurmair, 2000).

[2] Note that the notion of a “word error rate” as used in teh NIST evaluations (NIST 2001) is not a suitable evaluation concept for translation as there is not such a thing as a ‘canonical’ or ‘reference translation’ from which deviations could be computed: Three human translators produce four different versions of a text, all of which they claim to be correct.

## 1.2 Rule-based MT

In a second evaluation step, the output of the statistical MT was compared to a commercial rule-based MT system (linguatec's "Personal Translator" German-to-English).

**Setup**

The system was basically used as a raw MT system, with no specific tuning towards the domain.

The only action was to add some of the unknown words to the system dictionary. The 68 test sentences contained about 860 words, mainly very specialised database terminology. About 60 were not in the system dictionary. Of those, 20 were coded, using the system's coding tool. This was done to match the requirement that all words should be known (as it holds for the statistical MT).

Coding took less than 10 minutes as only 1:1 transfers were added to the dictionary. No further tuning was done.

**Evaluation**

The same evaluation measure was taken as for the statistical MT. The result can be given in the following table:

| | | |
|---|---|---|
| grammatical | 30 | 44,1% |
| understandable | 24 | 35,3% |
| wrong | 14 | 20,6% |

This result shows that the system is less strong in the middle category; either it finds a parse, and then produces good and grammatical results, or it fails. This fact shows that rule-based systems are less robust than alternative approaches.

However, the rule-based system produces significantly more grammatical results, and significantly better overall results (close to 80%) than the statistical MT system, under the same conditions (14 words sentences, no unknown words).

**Improvements**

Of course there is plenty of room to improve the translation quality of the rule-based system; mainly by tuning translation alternatives; this can easily be done, e.g. by assigning subject area codes to translations and choosing the right subject areas in translation. Recent studies (cf. Weber 2003) also underline a significant quality potential just using lexical measures. This was not done, however, as effects on the rest of the corpus could not be predicted, and it would have been an unfair tuning compared to the statistical package.

Also, recognition of named entities, proper names, product names etc. has been shown to improve the translation quality (Babych/Hartley 2003).

So there are significant tuning options just in the paradigm of rule-based systems; and there are customers which report error rates of only 3-4% for such systems.

# 2. Improvements

However, the question is not so much which approach is better; the more interesting question is what can be learned for the respective other approach, and how a hybrid system by which significant improvement in MT quality could be achieved should look like. To learn from the comparison, it is worthwhile to look at the translation results in more detail, and identify typical strengths and weaknesses of the respective approaches.

## 2.1 Statistical MT

This system basically works on chunks of input and assigns translations running a language model over the target words. Correlations of such chunks in source and target are learned, and used to translate the test corpus.

**Quality**

Translation quality is good if proper corresponding chunks can be identified in source and target language, like in (1)[3]; and fails if this is not the case, like in (29, 60). This counts for about 45% of the cases where translation quality is evaluated "wrong".

However, even if proper chunks are identified the translation fails in typical cases. Such failures can be described in linguistic terms, i.e. they can be generalised ("rule-based"). Typical failures are:

- German verb order and Satzklammer (split verbs) phenomena. Verbs in subordinate clauses must go from German last to English second position, and Satzklammer needs to be resolved.
  Here the system is not able to build a proper verb phrase (5, 27, 58), or drops one verb part altogether (31, 19).
- Constituent order: The system tends to keep the constituent order as in the source language (37, 68); cases where re-ordering is required (like in (63) where the German direct object is topicalised) tend to fail. Cf. also the wrong adverb placement in (57)
- Special constructions like German conditional clauses without subjunction.(47). The system translates plain indicative.
- Pronouns have several translations; the system tends to drop them altogether (22).

Such mis-handlings are systematic, they are responsible for about 55% of the 'wrong' evaluations, and it is hard to see how they could be overcome even if the training corpus could be extended significantly, because the "normal" material always outperforms the special cases.

Another systematic grammatical problem is to be mentioned, which is morphology. Statistical MT systems going from e.g. English into languages with richer morphology usually fail in assigning proper case information to their target output, in particular if the case indicates some functional relationship (like functional subject / object). This is less obvious in the current investigation as English does not use to many morphological markups.

On the lexical side, the statistical MT system performs quite well; so it is able to collect proper translation proposals from the training corpus. Sometimes wrong translations are given, however (4, 61, 64).

**Usability**

The crucial point is not that wrong lexical assignment can happen but that there is no possibility to control or influence the system behavior from a user's point of view. How can users add lexical items? How can they select a preferred translation in such a context? All this is crucial for a practical MT system.

[3] The numbers refer to the sentence numbers in the annex.

Another issue is domain-dependency. While statistical MT can be trained to a given domain with limited effort; this also means that it *has to be* trained to such domains every time anew. This is a never-ending task for a full-coverage MT system, and it is a severe problem in cases where no bilingual texts are available (which is nearly the majority of all cases). Even the best example-based systems (Richardson et al. 2001) have been tuned for one domain only (or one at a time).

From a practical and usability point of view, many questions remain to be solved before statistical MT systems can be considered to be operational.

### 2.2 Rule-based MT

These systems try to do a full parse on the input, and identify the basic syntactic functions in the sentence which are used for translation. Translation is done by looking up the words in the transfer dictionary and generating a proper word order and inflection.

**Quality**

The main sources of failure lie in the two main steps:

- **Parse failures** do not allow to identify the sentence parts; systems often use fall-back rules for those cases, but there will always be sentences which cannot be analysed properly. (cf. 25, 55)
- **Lexical failures** are the other main source of bad translations. This is not just that a word has no transfer entry in the dictionary; very often the problem is that there are *several* transfers in the dictionary and the system picks the wrong one.
  Examples are (10, 37, 57)
  In the tests mentioned above, two thirds of the "wrong" evaluation for the rule-based MT system are due to the problem of wrong lexical selection; so this seems to be more serious than the wrong-parse problem.
  A sub-section of this problem is translation of prepositions. They are notoriously difficult to translate, and there is much knowledge involved which is not rule-based but collocation-based; cf. (27, 56, 58).

In general, statistical MT performs better in these cases than rule-based MT. It is more robust than the fall-back strategies of rule-based systems, and it never picks translation readings which are outside of the domain (i.e. would simply not occur in a given corpus). Also, translation of prepositions contains less errors in statistical than in rule-based MT.

**Usability**

To select the right transfer from a set of options is a very difficult task, as current rule-based systems use systematic-linguistic features for disambiguation. They code in their transfer dictionaries under which conditions a term is transferred into a target term. Such conditions are mainly expressed in terms of features and values based on the conceptual model of underspecified morphosyntactic trees (good examples can be found in the OLIF (McCormick, 2001) and MILE (Calzolari et al, 2002) standardisation efforts for transfer entries). Examples are:

- Existence of certain **features** on the local node (e.g.: different transfers depending on gender),
- Existence of certain **syntactic functions** in a partial tree (e.g. different transfers of a verb depending on the presence of a direct object)
- Presence of certain surrounding **lexical material** (different transfer for adjective depending on the semantic type of the noun which it modifies; different transfer for nouns in compound specifier position vs. in head position)

and other such possibilities (more elaborate examples in (Thurmair 1990)).

Often however, either the text does not provide the required formal clues and neutralises readings, or the clues are more subtle to be detected by the current state of the art. Therefore it is not obvious how the selection process could be improved.

Of course, a rigid use of subject areas could prevent the system from picking out-of-area translations, but there are still sufficiently many cases of 1:n transfers left inside of such a subject area.

## 3. Conclusions

In the light of these discussions, the best way to proceed seems to be to create a hybrid system base a system on a rule-based architecture, and enrich it by features of statistical MT.

### 3.1 Rule-based backbone

The reasons to base it on a rule-based approach are the following:

1. It starts from a better quality baseline, and has already solved many of the usability and engineering problems which statistical MT still would have to overcome.

2. There are some ways how statistical MT can be improved:

- Preprocessing steps (better segmentation, morphological decomposition, name recognition etc.) definitely help to improve the MT quality by providing cleaner input to the statistical procedures.
- Replacing the (rather primitive) target language models by smarter linguistic-based generation components. Such components would use the lexical material produced by the statistical alignment, and try to 'make some sense' out of it, by putting them into the right constituent order and word formation. There have been related approaches in the paradigm of "shake and bake translation" in the early nineties (Whitelock 1992), however with limited success. But this approach would definitely improve results, and push some 'understandable' sentences into the 'grammatical' category.
- However, grammatical reference to the source sentence is still necessary, esp. in the area of grammatical functions (subject, object etc.). If this is not known, morphological case markings and/or word order cannot be stabilised. This kind of information requires significant linguistic analysis.

As a result, there are sources of knowledge which are indispensable for good MT, and it needs to be incorporated into a statistical backbone. A hybrid system based on such a statistical backbone is proposed in (Och et al. 2003), based on POS modeling, syntactic chunking probabilistic parsing and tree-tree alignment, with mixed quality results due to unreliable parses and the huge number of possible alternatives.

3. The main argument against rule-based MT is that it is costly to set up. There are three answers to this:

- There isn't such a thing as a free MT system. Building an MT system is always work.
- Cost is always relative, and is related to the savings which can be achieved, be it in productivity or in informativness. Examples show that investment into MT (esp. in the lexicon domain) pays off easily (Brundage 2001)
- Cost of a general-purpose MT system must not be compared to the cost of a special-purpose (one-domain) statistical system. Special purpose rule-based MT, with customised domain-specific dictionaries and grammars, can be set up in few months time. Cost for multi-domain general-purpose statistical MT is unknown as it does not exist.

For these reasons there is not really an alternative to a rule-based system backbone.

## 3.2 Statistical Enhancements

Assuming a decision in favor of a rule-based architecture, there are several ways how such systems could be improved by statistical means.

**Robust Parsing**

The idea is to improve rule-based parsing by statistical means. Instead of current approaches for probabilistic parsing only, the better strategy is to use probabilistic information to improve deep-linguistic analysis.

The side-effect of such a project would be to improve the analysis in robustness: In case of a parse failure, still the most probable analysis would be taken, just like in current statistical systems.

**Transfer Selection**

Instead of trying full statistical MT, the approach would be to find translation equivalents on word and phrase level for a given corpus / domain, and filter out all translation proposals which are not part of this corpus. After lexical transfer, standard target language generation components could be called.

This would reduce the hilarious results which MT is famous for, and leave only proposals which are valid for this domain.

Such an approach is promising also in cases of prepositions and other idiosyncratic translations, which make a good deal of the translation problems.

The challenge then would be to engineer such a solution: Create a special knowledge source for these cases, and have it interact with the current transfer components in a convincing way.

**Productivity Tools**

To increase productivity, statistical MT can be used as productivity tools in several respects:

- Pre-translation filter: Text analysis for the MT-translatability of a text. While most tools work on linguistic basis (Underwood/Jongejan 2001) (and repeat strengths and weaknesses of a rule-based MT system), a different technology may be better to detect such problems.
- Post-Translation filter: A statistical tool comparing MT output with 'standard' target text might help to locate problems which the MT system had: 'Strange' translations could be flagged, and postediting could focus on such segments first.

**Dictionary work**

Also in the preparation phase there are many options for statistical tools, mainly in the area to propose transfers from a given bilingual corpus. This is the intention of monolingual and bilingual terminology extraction tools (Thurmair 2003, Piperidis et al. 1997) which analyse corpus material to help to build linguistic resources.

Elaborate versions of such support users to create large bilingual linguistic dictionaries fast, and increase overall system productivity by shortening the coding phase. They assume, however, a rule-based type of MT system.

## References

Babych, B., Hartley, A. (2003). Improving Machine Translation Quality with Automatic named Entity Recognition. Proc. EACL-EAMT, Budapest.

Brundage, J. (2001). Machine Translation – Evolution not Revolution. Proc. MT Summit VIII Santiago

Calzolari, N., Bertagna, F., Lenci, A., Monachini, M., ed., (2002). Standards and best practice for multilingual computational Lexicons and MILE (the Multilingual ISLE Lexical Entry). ISLE-Report 2002

Charniak, E. (1997). Statistical parsing with a context-free grammar and word statistics. Proc. AAAI.

McCormick, S. (2001). The structure and content of the body of an OLIF v.2 File. www.olif.net

Nießen, S., Ney, H. (2000): Improving SMT Quality with Morpho-syntactic analysis. Proc. COLING 2000

(NIST, 2001) Automatic Evaluation of machine Translation Quality Using N-gram Co-Occurrence Statistics. www.nist.gov/speech/tests/mt

Och, F., Gildea, D, Khudanpur, S., et al. (2003): Syntax for Statistical Machine Translation. J. Hopkins Summer Workshop. www.clsp.jhu.edu/ws03/groups/translate

Piperidis, St., Boutsis, S., Demiros, J. (1997). Automatic Translation Lexicon Generation from Multilingual texts.. Proc. AAAI 1997.

Richardson, St., Dolan, W., Menezes, A., Pinkham, J. (2001): Achieving Commercial-quality Translation with Example-based Methods. Proc. MT Summit VIII, Santiago

Thurmair, G. (1990). Complex lexical transfer in METAL. Proc. TMI 3, Austin, Tx.

Thurmair, G. (2000): TQPro, Quality Tools for the translation process. proc. ASLIB, London

Thurmair, G. (2003). Making Term Extraction Tools Usable. Proc EAMT-CLAW Dublin.

Underwood, N., Jongejan, B. (2001). Translatability Checker: A Tool to Help Decide Whether to Use MT. Proc. MT Summit VIII, Santiago.

Vogel, S., Och, F, Ney, H. (2000). The Statistical tramslation Module in the Verbmobil System. Proc. KONVENS Ilmenau.

Vogel, S., Och. F., Tillmann, Chr., Nießen, S., Sawaf, H., Ney, H. (2000). Statistical Methods for Machine Translation. In. Wahlster, W. (ed.): Verbmobil: Foundations of Speech-to-Speech Translation. Springer.

Weber, N. (2003): MÜ-Lexikografie. Proc. GLDV, Köthen

Whitelock, P. (1992) Shake-and-bake Translation. Proc. COLING Nantes.

## Annex: Examples

The following list shows some examples of the output of the translation runs. Behind the sentence number there is

- the German source text (marked as "DE")
- the output of the statistical machine translation (marked as "SMT")
- the output of the rule-based translation ( "RMT").

Important aspects or missing parts are underlined.

01 DE Mit dieser Methode ermitteln Sie den ersten Kindknoten eines bestimmten Knotens.
SMT use this method to find out the first child node of a given node .
RMT With this method you find the first child node of a certain node out.

04 DE Die folgende Abbildung zeigt die Benutzersicht auf das R/3-System:
SMT the following diagram shows the View Point to the R /3 System :
RMT The following illustration shows the user view on the R/3 system:

05 DE Es gibt zwei unterschiedliche Möglichkeiten, aus ABAP-Programmen auf Datenbanken zuzugreifen:
SMT there are two different ways from ABAP programs access databases:
RMT There are two different possibilities of accessing databases from ABAP programs:

10 DE Der zweite Teilbaum beschreibt die Struktur des zu lesenden Datenbestandes .
SMT to read data with data with data with the second sub – tree describes the structure of the .
RMT The second partial tree describes the structure of the data stock to be read.

19 DE Von ihrem Arbeitsplatz aus sollen Ihre Mitarbeiter auch vereinfachte Transaktionen ausführen können .
SMT from their workplace should your employee also simplified transactions . [__]
RMT Your employees also shall be able to execute simplified transactions from their workstation.

22 DE tragen Sie sich bitte als Pilot ein , und ordnen Sie sich einem Flug zu .
SMT enter this as a pilot , and assign a flight .
RMT Please put down your name as a pilot and assign yourself to a flight.

25 DE liefert Liste von Schlüsselfeldern der Objekte , die bestimmten Auswahl - Kriterien genügen [*incomplete*]
SMT returns a list of the key fields of the objects satisfying the selection criteria
RMT List of key fields of the objects which are enough for certain choice criteria delivers

27 DE Die Reihenfolge der Ausgabe dieser Textelemente wird dann über das Druckprogramm gesteuert .
SMT the sequence of the text elements output this is then using the print program controlled .
RMT The order of the version of these tokens is then controled about the print program.

29 DE Die geschützten Komponenten sind eine Schnittstelle zu den Unterklassen von C 1 .
SMT the protected components are an interface and of C 1 .
RMT The protected components are an interface to the subclasses of C 1.

31 DE Beim Anlegen einer Tabelle wird ein Schlüssel nach programmtechnischen Gesichtspunkten vergeben.
SMT when you create a table is a key after a a technical assign . [__]
RMT When creating a table a key is allocated after technical points of view.

37 DE Für die Verknüpfung mit Organisationsobjekten müssen Sie kein HR im Einsatz haben.
SMT for the link with organizational objects must use in HR you do not have .
RMT You don't have to have any HR for the bonding with organization objects in use.

47 DE Ist diese Berechtigung nicht vorhanden, so wird das Empfangs-programm abgebrochen
SMT this authorization is not available , the receiving program terminated
RMT If this right isn't existing, then the receiving program is broken off

55 DE Über die Job - Übersicht ( Transaktion SM 37) erfahren Sie näheres .
SMT use the job overview ( transaction SM 37) tell you for further information .
RMT Learn you about the job overview (transaction sm 37) near.

56 DE Der Datenbankmonitor zeigt zum Redo-Log-Puffer die folgenden Informationen an:
SMT the Database Monitor shows the redo log buffer the following information:
RMT The database monitor displays the following information to the redo log buffer:

57 DE Ein Job kann jedoch erst ausgeführt werden, wenn er freigegeben wurde.
SMT a job but can be executed if it was released .
RMT A job can, however, be executed only if it was unlocked.

58 DE Sie können die Tabellenpuffer über die Parameter im Profil der Instanz neu einstellen .
SMT you can the table buffers using the parameters in the instance profile set a new .
RMT You can adjust the table buffers newly about the parameters in the profile of the instance.

60 DE Die effizientesten Schreiboperationen werden am Checkpoint durchgeführt ("Chunk Writes")
SMT the writes to the checkpoint to be performed ("chunk Writes")
RMT The most efficient write operations are executed at the checkpoint ("Chunk Writes")

61 DE Alle drei Workprozesse sind frei und die folgenden Jobs sind startfähig:
SMT all three work processes are empty and the following jobs are eligible to run:
RMT All three background processes are free and the following jobs are eligible to start:

63 DE Die Parameter der Datenbanksystemprüfung können Sie in der R/3- Tabelle DBCHECKORA konfigurieren.
SMT the parameters of the database system check [__] in R /3 configure table DBCHECKORA .
RMT The parameters of the database system test can configure you in the R/3 table DBCHECKORA.

64 DE Der Status der Aktionszeile im DBA – Einplanungskalender besitzt folgende Farbcodierung :
SMT the status of the Delete in the DBA Planning Calendar has the following color coding:
RMT The status of the action line in the DBA planning Calendar has the following color coding:

68 DE Um den Dialogmodus von saposcol zu verlassen, verwenden Sie den Befehl quit:
SMT dialog mode to the saposcol to leave , use the command quit:
RMT To exit the dialog mode of saposcol, you use the command quit:

# Ground Truth, Reference Truth & “Omniscient Truth” -- Parallel Phrases in Parallel Texts for MT Evaluation

**M. Vanni*** **C.R. Voss*** **C. Tate*** [§]

*Multilingual Computing Group
Army Research Lab
Adelphi, MD

[§]Dept. of Mathematics
University of Maryland
College Park, MD

{mvanni | voss }@arl.army.mil, ctate@math.umd.edu

**Abstract**

Recently introduced automated methods of evaluating machine translation (MT) systems require the construction of parallel corpora of source language (SL) texts with human reference translations in the target language (TL). We present a novel method of exploiting and augmenting these resources for task-based MT evaluation, assessing how accurately people can extract *Who, When,* and *Where* elements of information from TL output texts of different MT engines. This paper reports on the first phase of our research establishing a baseline MT evaluation process with (i) the construction and (ii) the annotation and inter-annotator rates of *an annotated extraction corpus,* and (iii) our results applying the corpus in the evaluation of three Arabic-to-English MT engines. In this corpus, the elements of interest are identified as parallel phrases across the parallel texts of the SL, the reference translations, and the MT engine outputs, where they are annotated and called, respectively the Ground Truth (GT), Reference Truth (RT), and Omniscient Truth (OT) items in the parallel texts. Our evaluation of three MT engines with the corpus yields precision and recall accuracy measures that, together with a loss measure, clearly rank the engines and, unlike other evaluation metrics, indicate diagnostically where output improvements will assist on extraction.

## 1 Introduction

Current methods of evaluating machine translation (MT) systems are costly: they require the construction of parallel corpora of source language (SL) texts with human reference translations in the target language (TL) prior to the run-time evaluations. We present a novel method of exploiting and augmenting these resources that we use for an experiment in task-based MT evaluation, assessing how accurately people can extract *Who, When,* and *Where* elements of information from TL output texts of different MT engines.

Our research approach is to divide into three stages, the analysis of which “end-to-end” MT engine-with-user combination produces the most complete and accurate information. First, we evaluate the MT output standalone (that will later be shown to users) for how adequately the engines preserve the content of the *Who, When,* and *Where* elements. Second, we conduct an experiment with users viewing the MT outputs of different engines and evaluate their responses (that they provide via our software tools) for how effectively they can extract the elements. Then, we use the results of these evaluations within a generalized linear model to test the relation of MT engine, document and subject variables in predicting the “end-to-end” MT engine-with-user accuracy in extracting the elements from MT output.

This paper reports on the first phase of the research approach with (i) the construction and (ii) the annotation, with inter-annotator rates, of ***an annotated extraction corpus***, and (iii) our results applying the corpus in the evaluation of three Arabic-to-English MT engines. In this corpus, the elements of interest are identified as parallel phrases across the parallel texts of the SL, the reference translations, and MT engines’ outputs, where the elements are annotated and called, respectively the ground truth (GT), reference truth (RT), and omniscient truth (OT) items in the texts. Our evaluation of the three MT engines with the corpus yields precision and recall accuracy measures that, together with a loss measure, clearly rank the engines and, unlike other evaluation metrics, indicate diagnostically where output improvements will assist on extraction.

## 2 Approach

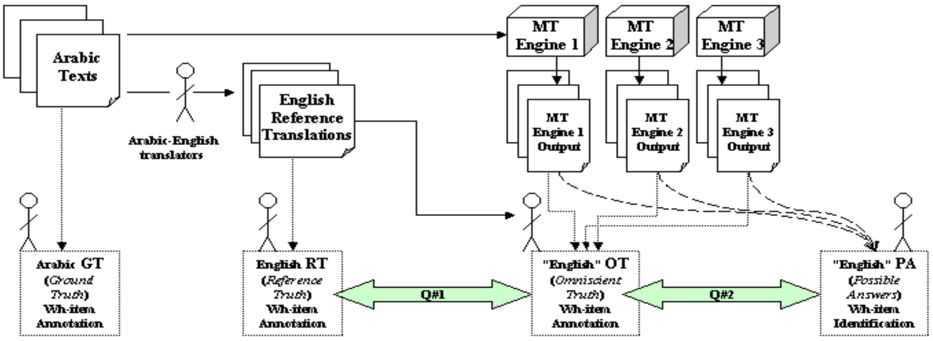
The construction of the annotated extraction corpus, illustrated in Figure 1, involves building the parallel texts, annotating them for the parallel phrases, and then augmenting the phrases in the MT output files with a higher-order, backoff categorization for evaluating the OT items in those files.

### 2.1 Parallel Texts

The corpus that we have created is effectively a three-way parallel corpus of the source language texts, reference translations, and MT outputs, aligned at the sentence level. We started with a collection of online Arabic language documents built by one native Arabic speaker with news article from ten different websites, where each article was selected for one of the who/when/where extraction tasks of the second stage of our research.

Four native Arabic speakers (including the one who built the collection), all bilingual in Arabic and English, then translated the documents into English to create the four reference translations for the corpus. We followed the guidelines established at the Linguistic Data Consortium for directing these individuals to create translations that preserve the full content of the documents as closely as possible and that do not add extra information which is not literally present in the text. They were instructed to translate the Arabic text on a sentence-by-sentence basis, creating English sentences that are fluent and do not contain Arabic constructions, such as sentences that start with the word “And” after the initial paragraph sentence.

To create the MT output files of the corpus, we ran the online Arabic documents through each of the three


Arabic Texts
Arabic-English translators
English Reference Translations
MT Engine 1
MT Engine 2
MT Engine 3
MT Engine 1 Output
MT Engine 2 Output
MT Engine 3 Output
Arabic GT (Ground Truth) Wh-item Annotation
English RT (Reference Truth) Wh-item Annotation
Q#1
"English" OT (Omniscient Truth) Wh-item Annotation
Q#2
"English" PA (Possible Answers) Wh-item Identification

Q#1: <u>how well do MT engines translate wh-items?</u>
**Answer: compare and code RT & OT in MT output**
> accurate if OT matches RT item
> broken if OT partially or reordered matches RT item
> scattered if noncontiguous OT items part match RT item
> lost if no OT item in MT output for RT item

Q#2: <u>how well do people identify wh-items?</u>
**Answer: compare PA & OT in MT output**
> hit if PA matches OT item
> partial hit if PA matches some of OT item
> miss if no PA item covers OT item
> false alarm if PA item does not cover OT item

**Figure 1.** Process of Constructing an Annotated Extraction Corpus

Arabic-to-English MT engines that we had available in their most recent release as of the end of October 2003. As needed, we converted the documents into the input format required by the MT engine. We also opted to run the MT engines with any settings left at their default value.

## 2.2 Parallel Phrases

Given our second-stage goal of evaluating how well people can extract *Who, When,* and *Where* elements of information from MT output for the purpose of ranking the "end-to-end" MT engine-with-user combinations, we experimented with defining these elements at different levels of granularity. The key was to determine the most straightforward, non-technical description of "chunks" of information in "noisy" MT output[1] that the people in our experiments, who were neither translators nor linguists, would be able to detect readily without extensive training.

We started out examining the category descriptions for PER (person), ORG (organization), LOC (location), and TIMEX (time expression) in the ACE program guidelines. Reading these guidelines and effectively learning the large and fine-grained distinctions among the categories that are extensively documented with examples requires several hours. Furthermore the categories are defined over the smallest atomic element of information, not the phrasal or chunk level that we needed in order to assess both the content of the MT output in the first phase of our work, and the feasibility of people extracting *Who, When,* and *Where* elements from the noisy MT output in the second phase of our work.

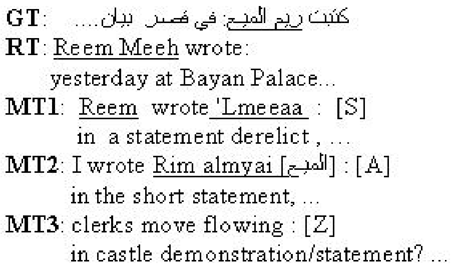
As a result, we established instead intuitive semantic descriptions, where the chunk could include attributes if that information was local within the syntactic phrase in the SL or reference translations. We pre-tested and refined the descriptions on members of our staff with no linguistic training, after giving them about twenty minutes training.

The identification of *Who, When,* and *Where* ground truth (GT) elements in the Arabic texts was set by one native Arabic translator and then vetted by a trained linguist in possession of the English human reference translations, who then marked up these documents for their parallel reference translation (RT) phrases.

The "who" category of our annotations consists of mentions of individual persons or groups of people, organizations, corporations, governments or other entities functioning as persons in the context of the SL passage. Here we include roles, names, objects with human identity and numbers referring to persons. The "where" category is comprised of names, proper/common nouns, and expressions such as prepositional phrases which refer to locations, regions, facilities, civil structures and other bounded geographic areas. The "when" category contains time and date expressions with standard proper noun month-day-year references, common nouns referring to time periods or instants, unique identifiers for temporally-defined events, or prepositional phrases referring to specific time periods.

After the GT-RT annotations were established, we developed the following procedure for identifying the corresponding "omniscient truth" (OT) elements in the MT outputs. Given a listing of the RT elements by document in order of appearance within each sentence of the document, the annotators searched within the same sentence of the MT output text for the OT that best approximated the RT element. The OT "chunks" were selected semantically by the annotators, so that even when they found incorrect English syntax or incomplete translations only roughly corresponding to the RT element, they could identify an OT item. The set of OTs for a document vary with the MT engine that generated the output text in the document. This can be seen in the example in Figure 2 where the underlined subject of the verb is translated by MT3, is transliterated by MT1 and MT2, and is separated across the verb in MT1.

[1] "Noisy" MT output refers to text output by MT engines that contains ungrammatical phrases with words out of order, incorrect or peculiar word selections, unrecognizable transliterated names, SL words left untranslated, and so on.



GT: كتبت ريم الميح: في قصر بيان....

RT: Reem Meeh wrote:
yesterday at Bayan Palace...

MT1: Reem wrote 'Lmeeaa : [S]
in a statement derelict , ...

MT2: I wrote Rim almyai [الميح] : [A]
in the short statement, ...

MT3: clerks move flowing : [Z]
in castle demonstration/statement? ...

**Figure 2.** Sample Parallel Phrases in Parallel Texts: *Who* ground truth (GT) phrase in Arabic source text, *Who* reference truth (RT) phrase in reference translation, *Who* omnisicient truth (OT) phrases with backoff codes in output texts of three Arabic-to-English MT engines

## 2.3 MT Output Backoff Classification

As annotators were reading the MT output texts and identifying the OT items to be marked up, they spontaneously started categorizing the patterns of errors in the MT output that directly affected their decision-making process of establishing the boundaries of an OT item. For example, in Figure 2, they designated the open class words such as "wrote" in the MT1 text that appear incorrectly inside of the translated phrases as "trapped words." As the markup process continued, the name for the OT items with such trapped words inside evolved into "split items."

When we observed that the annotators were regularly using their terms for error patterns to resolve differences in their OT markups, we realized this information was central to the OT identification process and decided to codified it by grouping their error analysis patterns into four classification categories (A, B, S, and Z) and then tested their consistency in assigning the classification labels to the OT items.

**Definitions of OT Item Classifications**

**A:** 1) Exact match, synonym, or paraphrase
2) Contiguous phrase
3) Words in grammatical word order

**B:** 1') Exact match, synonym, paraphrase
OR partial match with some content loss
2) Contiguous phrase
3') Words in grammatical word order
OR out of grammatical order

**S:** 1') Exact match, synonym, paraphrase
OR partial match with some content loss
2') Non-contiguous phrase
3') Words in grammatical word order
OR reordered OR out-of-order

**Z:** Lost OR not recognizable

The OT identification and backoff classification process worked as follows. First, the annotators would compare their respective OT items with the RT items for a match, within the relevant sentence, that preserved the RT meaning and that formed a grammatical element. These items were the best, or "A" cases. When there was no evidence for that form of an OT item in the MT output, they would do a backoff analysis and look for a chunk of contiguous words that *would* be a good OT item, *if* the words were re-arranged or had another word or two added in. Since these items were clearly not as easy to detect because they required spotting the relevant words in a partial or noisy pattern, these items became "B" cases. The split items, mentioned earlier, became the "S" cases. Finally, for those cases where the annotators could not identify any text in the relevant MT output sentence that conveyed the name in or the semantic content of the RT item (as occurs in the MT3 output for the subject's name "Reem Meeh" in Figure 2), the annotators designated that item "Z" to record that it was lost in translation.

# 3 Results

## 3.1 Backoff Classification

We evaluated the inter-annotator agreement rates on the ABSZ coding with the Kappa statistic (Cohen, 1960) for each MT engine, both across and within the who/when/where types after one round of annotation, but before the final resolution of the codes. The scores were all within the 0.6 to 0.8 "good agreement" range. Also, three of the nine Kappa scores for within who/where/when types were above 0.8 in the "very good" range. Most of the differences among annotators were at the A-B boundary.

The results of assigning each OT item to one of the four categories, A, B, S, or Z, are shown in Table 1. The total rows for each of the MT engines indicate, across who/when/where elements, how many are categorized as OTs (As, Bs, or Ss) and how many are lost in translation (Zs). The precision measure is the number of As divided by the number of OTs, and the recall measure is the number of As divided by the number of RTs. The loss measure is the number of Zs divided by the number of RTs. RT totals used in the Recall calculations for all MT engines are: 156 for all wh-items, 56 for Who items, 56 for Where items, and 44 When items.

| | **Backoff Classification** | | | | OT | **Accuracy Measures** | | |
|---|---|---|---|---|---|---|---|---|
| | A | B | S | Z | | **Prec** | **Rec** | **Loss** |
| **MT1Total** | 67 | 51 | **20** | **18** | **137** | **.49** | **.43** | **.12** |
| Who | 21 | 17 | 12 | 6 | 50 | .42 | .38 | .11 |
| Where | 34 | 15 | 2 | 5 | 51 | .67 | .61 | .09 |
| When | 12 | 19 | 6 | 7 | 36 | .33 | .27 | .16 |
| | | | | | | | | |
| **MT2Total** | **91** | 49 | 9 | 7 | **149** | **.61** | **.58** | **.05** |
| Who | 29 | 19 | 7 | 1 | 55 | .53 | .52 | .02 |
| Where | 41 | 12 | 1 | 2 | 54 | .76 | .73 | .04 |
| When | 21 | 18 | 1 | 4 | 40 | .53 | .48 | .09 |
| | | | | | | | | |
| **MT3Total** | 67 | **75** | 4 | 10 | **146** | **.46** | **.43** | **.06** |
| Who | 21 | 26 | 2 | 7 | 49 | .43 | .38 | .13 |
| Where | 33 | 22 | 0 | 1 | 55 | .60 | .59 | .02 |
| When | 13 | 27 | 2 | 2 | 42 | .31 | .30 | .05 |

**Table 1.** Counts of ABSZ Codes and Precision/Recall/Loss Percentages on MT Output of Annotation Extraction Corpus

### 3. 2 Interpretation

The precision, recall, and loss measures in Table 1 serve to tease apart the differences among the three Arabic-English MT systems that we tested. There are four results in this table. First, notice the substantially higher precision and recall scores of MT2 (.61 and .58), compared to those of MT1 (.49 and .43) and MT3 (.46 and .43), based on "A" scores. Second, while the precision and recall scores for MT1 and MT3 nearly identical, the loss scores based on "Z"s make is clear that MT1 is much weaker in preserving content. Third, MT1 is also weaker in preserving phrasal integrity, with more than twice the number of "S" split phrases in the output compared to the other two engines. Finally, Table 1 also makes clear that MT3 is the mostly likely engine to output less-than-correct "B" partial or broken-syntax translations.

To recap, the results of our work so far indicate we can rank the MT engines in our study on their accuracy and throughput in translating the wh-elements of interest for later extraction: MT2 provides the strongest overall results, MT1 has the weakest overall results because of its loss of content and phrasal integrity, and MT3 falls between the other two, with accuracy below that of MT2 but with better content throughput than MT1.

## 4 Related Work

We have developed a novel two-part approach to standalone MT engine evaluation that augments parallel text resources into an annotated extraction corpus and applies it in a focused *Who, When,* and *Where* backup classification of MT output text. This two-part approach is comparable to other current annotate-and-train/test approaches found in the processing of natural language texts for a wide range of applications, such as (i) tagged corpora for information extraction (Sundheim, 1991), (ii) bracketed corpora for parsing (Marcus, *et al*., 1993), and (iii) sense-tagged corpora for word sense disambiguation (Kilgariff and Palmer, 1999), to name but a few. These applications first require constructing corpora, developing well-documented annotation procedures for human annotators, determining the inter-annotator agreement rates, and resolving final annotations on the corpus. For many NLP applications, the annotated corpora then serve to train/test the algorithm for automating a particular task. In our work reported here, the annotated extraction corpus has served to develop the backoff classification algorithm for MT evaluation.[2]

While others have made *unannotated* parallel bilingual corpora central to their MT evaluation research[3], it is not yet clear what the results from these automated metrics signify. For example, Hovy and Ravichandran (2003) have shown that MT output that outperforms reference translations on these metrics may nevertheless be incomprehensible to human readers. Our approach with parallel corpora *annotated* for *Who, When, Where* extraction will allow us to test, in the second stage of our research, for a predictive model that can cross-validate our backoff evaluation performance measures with the effectiveness measures achieved by MT engine-with-user combinations carrying out extraction tasks.

## 5 Conclusions and On-Going Work

This paper reports on the first phase of our research establishing a standalone MT evaluation process with (i) the construction and (ii) the annotation and inter-annotator rates of *an annotated extraction corpus,* and (iii) our results applying the corpus in the evaluation of three Arabic-to-English MT engines. In this corpus, the elements of interest are identified as parallel phrases across the parallel texts of the SL, the reference translations, and the MT engine outputs, where they are annotated and called, respectively the Ground Truth (GT), Reference Truth (RT), and Omniscient Truth (OT) items in the parallel texts. Our evaluation of three MT engines with the corpus yields precision and recall accuracy measures that, together with a loss measure, clearly rank the engines and, unlike other evaluation metrics, indicate diagnostically where output improvements will assist on extraction.

We are currently conducting analyses, as part of the second stage of this research, on the results of task-based categorization, extraction, and template-completion experiments, where people read output text from the same three MT engines reported on in this paper. Given the results from the backoff classification found so far, we hypothesize that people will work most effectively with MT2 output. We also predict that there will be a range of individual differences in how well people are able to carry out these tasks on the output of MT1 and MT 3, as a function of how much experience they have with MT output.

**Acknowledgements**. This work was supported by the Center for Advanced Study of Language (CASL) at the University of Maryland, College Park, and the US DOD. We acknowledge the support of our lead translator, Jamal Laoudi, who supervised the translation and ground truthing of the document collection, and two other team members, Sooyon Lee and Joi Turner, who contributed to the OT and ABSZ annotation development and resolution steps.

## References

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*: 20 (pp. 37-46).

Doddington, G. (2002) "Automatic evaluation of machine translation quality using n-gram co-occurence statistics." In *Proceedings of HLT 2002,* Human Language Technology Conference, San Diego, CA.

Hovy, E. and D. Ravichandran (2003). Holy and Unholy Grails. Presentation at Panel, "Have we found the holy grail?" MT Summit IX, New Orleans, LA.

Kilgariff, A. and M. Palmer (1999). *Computers and the Humanities*: 34:1-2 (Special issue on Senseval1).

Marcus, M. B. Santorini, and M. Marcinkiewicz (1993). Building a large annotated corpus of English: the Penn Treebank. *Computational Linguistics*: 19.

Papineni, K., S. Roukos, T. Ward, and W. Zhu (2002). BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the ACL*, Philadelphia, PA

Sundheim, Beth, ed. (1991). In *Proceedings of the Third Message Understanding Conference* (MUC-3), San Diego, California. Morgan Kaufmann, San Mateo, CA.

[2] The corpus is also used in the second stage of our research on task-based extraction evaluation, not detailed in this paper.

[3] Papineni, *et al*. (2002), Doddington (2002).

# Gathering Empirical Data to Evaluate MT from English to Portuguese

**Diana Santos**

Linguateca, Oslo, SINTEF ICT
Pb 124 Blindern, 0314
Oslo, Norway
Diana.Santos@sintef.no

**Belinda Maia**

Fac. de Letras de Univ. do Porto
Via Panorâmica, s/n, 4150-564
Porto, Portugal
bmaia@mail.telepac.pt

**Luís Sarmento**

Linguateca, Porto, FLUP
Via Panorâmica, s/n, 4150-564
Porto, Portugal
las@letras.up.pt

## Abstract

In this paper we report on an experiment to gather quality analyses from several people, with a view to identifying problems and reaching consensus over (machine) translation from English to Portuguese. We start the paper by showing how this project is part of a larger framework of evaluation campaigns for Portuguese, and suggest the need for amassing consensual (or at least compatible) opinions. We describe the various tools (Metra, Boomerang, and TrAva) developed and explain the experiment, its results, shortcomings and lessons learned. We then present CorTA, a corpus of evaluated translations (English original, and several automatic translations into Portuguese) and make some remarks on how to use it for translation evaluation.

## Introduction

Let us begin by stating that the issue of evaluating translation is not new and is extremely complex (see e.g. Bar-Hillel, 1960). Machine translation (MT) evaluation has a long history, starting with the ALPAC (1966) report, which was extremely important for MT and NLP in general. However, we should also like to drw attention to two interesting facts: translation seems to remain one of the most popular NLP applications, and its output is judged by laymen in a way that no other complex intellectual activity is: while ordinary people would not think of criticizing a legal document written by a lawyer, an experiment designed by a physicist, or a diagnosis performed by a doctor, no one refrains from judging and criticizing the output of such a complex craft (or art) as translation.

In fact, translation is an interesting area because most people have strong opinions about the quality of particular (mis)translations (as opposed, for example to assessing the quality of IR results or abstracts). However, in most cases, it is remarkably difficult to elaborate objective criteria with which to classify, praise or reject specific translations.The work described in the present paper is an attempt to assess some of these analyses in a form that will later allow us to make generalizations.

Linguateca's efforts to start joint evaluation activities in the field of the processing of Portuguese, defined in the EPAV'2002 and Avalon'2003 workshops, selected three main areas[1]: morphosyntax, leading to the first *Morfolimpíadas* for Portuguese (Santos *et al.*, 2003, Santos & Barreiro, 2004); information retrieval, with resource compilation (Aires *et al.*, 2003) and participation of Portuguese in CLEF (Santos & Rocha, forthcoming); and machine translation (MT), reported here and in Sarmento *et al.* (forthcoming).

It should be noted that these are radically different areas with different challenges and different interested participants. For MT, despite projects initiated in Portugal and in Brazil, the Portuguese/Brazilian developing community has, on the whole, had very little impact on the outcome of current commercial systems, and specifically those available on the Web. However, and given that Portuguese is a major language in terms of the number of native speakers, there are plenty of international systems that feature translation into and from it, and there are many users of such systems worldwide. It was therefore thought that the best (initial) contribution that a Portuguese-speaking and Portuguese-processing community could offer was the identification of the specific problems (and challenges) posed by translation into Portuguese or from Portuguese. (We started with English as the other language.)

First, we thought about gathering test suites (of the translational kind of King & Falkedal, 1990), but in the initial process of discussing which phenomena should be extensively tested, there arose a more general concern with evaluating which kinds of problems were more obvious (and could also be consensually labeled) which led to the work described here.

The Porto node's concern with users in a language and translation teaching environment, and its close connection with the teaching activities at the Arts Faculty of the University of Porto, provided an excellent testbed a for testing the possibility of collecting (machine) translation evaluations during the study programme. The pedagogical objective was to increase future translators' awareness of MT tools and encourage their careful assessment of current MT performance.

## Gathering Judgements: TrAva

Our project had the double requirement of having trained translators with little formal knowledge of linguistics classifying the quality of the translation, and the need to create a classificatory framework that allowed comparison of examples, without making assumptions on the behaviour of specific MT systems.We have thus created a system for the empirical gathering of analyses called TrAva (Traduz e Avalia)[2], that has been publicized for general use by the community dealing with MT involving Portuguese, and whose continued use may supply new

[1] See http://www.linguateca.pt/AvalConjunta/, for information on the workshops and the three interest groups formed. ARTUR is the one for translation and other bilingual tasks.

[2] Available from http://www.linguateca.pt/TrAva/.

user requirements and functionalities, as well as larger amounts of classified data.

Due to the exploratory nature of this work, the use of the system by students and other researchers throughout the experiment has led to an almost continuous refinement of functionalities and several different versions. Although in the present paper we are restricted, for lack of space, to presenting only the current system, we must emphasize that the whole development proceeded bottom-up, and that the changes were motivated by the analysis of the input presented to the (previous versions of the) system.

TrAva is thus a system whose goal is to come to grips with some of the intuitively employed criteria of judging translation, by producing a relatively easy framework for cooperatively gathering hundreds of examples classified according to problems of (machine) translations.

From the analysis of the initial input to the system, it became clear that one should not rely on non-native competence to produce sentences to be translated, and thus we enforced the requirement that authentic English materials should be employed (and their origin documented, see Figure 1). Likewise, we required that only native speakers should classify translations, which means that so far we have only collected authentic English source language examples automatically translated into Portuguese and classified by Portuguese native speakers.

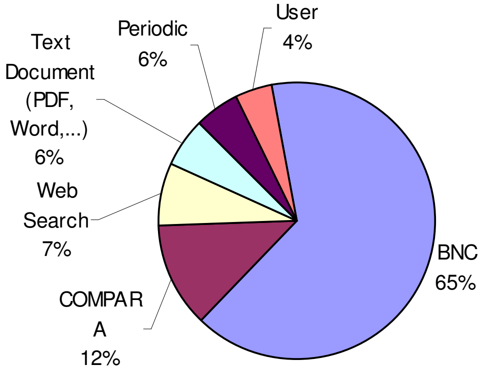


**Figure 1 - Distribution of the origins of English sentences in TrAva (1270 sentences)**

In order to be able to compare and gather large amounts of sentences with the same "classification", and also to reduce subjectivity (or error) in the classification of the English text, we used the British National Corpus (BNC), Aston & Burnard (1996), and its PoS-tagging, as a first organization criterion. (Note that the students were also being taught to use the BNC in their translation education, so no additional training was required for the MT evaluation exercise.) The user is requested to indicate a sequence of PoS tags and classify the problems in the translation of this particular sequence, and not anywhere else in the sentence. One may submit the same sentence with a different target sequence, when additional interesting problems are observed, but ideally one should be considering each problem or structure in turn.

Due to the availability of Web-based MT engines, compared to systems that require acquisition, installation and/or format conversion, and given that we did not want to restrict the evaluation work to in-house members, but instead to offer it as a joint activity to the community concerned with Portuguese language processing and even with MT, we chose to evaluate the performance of Web MT systems. As a preliminary step, two systems were developed: METRA (a meta-MT engine), http://poloclup.linguateca.pt/ferramentas/metra/ and Boomerang, a system that sequentially invokes MT in the two directions until the same output is produced, http://poloclup.linguateca.pt/ferramentas/boomerang/.

They helped us identify problems and solutions to the engineering of invoking remote systems,[3] and also gave us valuable insight into the relationships or dependencies among the seven MT engines involved. The final set used in TrAva contains the following four MT services: FreeTranslation, Systran, E-T Server and Amikai.[4]

## A four-fold Classification Activity

The user of TrAva has, first, to decide which part of the sentence s/he is going to evaluate, and PoS-classify it. The text is then submitted to the four MT engines referred to above and the results are presented to the user, who reports on how many translations display problems in translating the selected part. Only then can the user engage in the most time-consuming (and complex) classification activity, namely to identify, using TrAva's grid, the problems that appear in the translation(s).

Finally, and optionally, the user can also provide an alternative translation (this is encouraged), together with comments in free text. These comments have provided us with valuable input not only on several inadequacies of the current classification grids but also with feedback about the usability of the system. The alternative translation can also be considered a kind of classification (it may at least be used, in the future, as data for a re-classification, and for refining the grid).

A feature that may be difficult to understand is TrAva's requirement that the user classify more than one translation at once, and thus it requires some explanation on our part: Our main wish is to identify cases which are difficult enough not to have been (totally) solved by any system yet, rather than compare the systems. One would expect to have problems that originate in the differences between English and Portuguese and that are not covered by current state of the art systems, such as questions, the translation of reflexives, modal verbs, homographs, complex noun phrases, etc, to mention just a subset of the problems investigated. So, we were expecting many translations to display the same or similar errors.

However, when it comes to a fine-grained classification of the problem, it appears that different systems often make different errors, and we are aware that it may be confusing for a user to try to classify all of them in one fell swoop.

## Yet another Parallel Corpus: CorTA

One of the most relevant by-products of our experiment is CorTA (Corpus de Traduções automáticas Avaliadas), a corpus of annotated MT examples from English to

[3] One has to deal with timeout, or "system not available", with error messages, with excess length and consequent truncation, and – astonishingly – even sometimes with character codes and punctuation.

[4] URLS are: http://www.freetranslation.com, http://www.systransoft.com/, http://www.linguatec.de, http://standard.beta.amikai.com/amitext/

Portuguese with non-trivial search possibilities. This novel resource has currently around one thousand input sentences (about 65% coming from the BNC) and, in addition to the usual search in parallel corpora like DISPARA (Santos, 2002), it allows for selection by kind of error and by translation engine. IMS-CWB (Christ et al., 1999) is the underlying corpus processing system.

CorTA is available at www.linguateca.pt/CorTA/, and is meant to grow at the rate required by the cooperative compilation of evaluations through TrAva. It is "frozen" in the sense that we do not plan to continue its development before October 2004, but until then we wish to receive feedback and gather more data, in order to assess what could be done and in which direction(s) it should be further developed.

This corpus is different in several ways from the one described in Popescu-Belis *et a*l. (2002). Instead of a set of reference translations, it displays a set of (sometimes, correct, but usually incorrect) translations, which have not been hand-corrected, only hand-classified in relation to a subset of the problems they display.

Also, while the classification of Popescu-Belis *et al.*'s corpus is performed by a small group of experts (translation teachers), ours is cooperatively created by a set of people with little background, if any, in translation evaluation and is in principle open to any person who is a native speaker of one of the languages and knows the other well enough.

Although no numbers have been reported, we also expect the creation of such a corpus to be much more time-consuming than ours. On the other hand, their result will be a reference material, while ours, as it stands, can only be seen as a tool for empirical research in evaluation, translation, and human inter-agreement.

## Lessons Learned

Although the system was initially created to allow cooperation among MT researchers, we soon learned that one cannot expect people to gather enough material for reliable research, without having some financial or other reward (such as project funding). Thus, if one wants people to consistently use a system whose primary goal is to provide data for later research, one has to employ students and/or people who may directly benefit from using it (such as those writing assignments).

So – as is, in fact, also the case in other kinds of empirical data gathering, such as software engineering (Arisholm et al., 2002) – one has to use students and not experts or translation professionals. In the case of TrAva, however, given that, as pointed out in the initial section, every one seems to have intuitions about translation quality, we believe that students of translation are expert enough when compared to "real" laymen.

Another relevant lesson is that very often a problem can be classified according to source-language, transfer/contrastive, or target-language criteria, and that this is a source of confusion to users of TrAva and consequently also of CorTA. For example, suppose the user was interested in the complex noun phrase *the running text mode* and one system had provided **o modo correndo do texto* (!). One could classify this erroneous translation as (English) attachment ambiguity wrongly analysed; (contrastive) incorrect resolution of ambiguous *ing*-form (adj-> verb); (Portuguese) wrong article insertion/use, etc. All presuppose some model of how the system works – and may therefore be wrong – but by trying to guess the causes of the error, one may come to significant generalizations and, anyway, one cannot prevent people from thinking![5]

So, while TrAva may seem flawed because different users may use different strategies to classify the problems, we believe it is also a strength that allows higher-level cause classification instead of simple objective correction. One is then able to look for all cases in the corpus that come from wrong PoS assignment regardless of the actual words or even the English patterns employed.

As the project developed, various other things became clear. For instance, we recognized the desirability of asking people to provide a good human translation, and the need to classify the MT output as acceptable in both Brazilian and European Portuguese.

## Concluding Remarks

Obviously, the work we report here has never been thought of as an ultimate step in MT evaluation, but as a (maximally) unbiased pre-requisite for discovering a number of problems and for eventually producing a roadmap for MT into and from Portuguese.

We have not, at this stage, even tried to define metrics that could be employed to measure MT output, although we believe that CorTA could be a starting point for training automatic evaluators and for investigating the agreement with human intuitions about translation quality. There are a number of metrics and procedures used in MT evaluation (see Dabbadie *et al.*, 2002, for an overview), several of them making use of reference translations created by human translators, and specifying different translation goals (such as terminology coverage, NER handling, or syntactic correctness by counting the quantity of editing required). Because of their attempt at generality, they fail to consider the specific linguistic problems that the pairing of two particular languages poses. It is this language-dependent part that we want to address and to which we feel confident that the Portuguese-processing community can significantly contribute.

TrAva and CorTA are thus tools that allow everyone to look at specific problems of translation between the two languages and to suggest further ways to create representative samples (test suites) to test automatic translation per problem, instead of using "infamous" and irrelevant sentences from Shakespeare or the Bible or from the tester's (lack of) imagination: "this is a test", "hello, world", and the like.

One should not forget that significantly and surprisingly good machine translation(s) can already be found as output of MT systems on the Web, and it is important to consider this in any reliable assessment. Although the judgments currently stored in TrAva are by no means representative, it is interesting to report that, in the process of testing probable sources of problems, users were partially happy in up to 66% of the cases, i.e., they considered 66% of the translations faultless regarding the phenomena under investigation, see Figure 2. CorTA can thus also be used as repository of solved problems (or of

[5] On the contrary, instead of asking people to replicate machines, it would be more useful to ask them to think.

cases solved to a large extent), as well as of difficult cases to be used in future tests.

Finally, we must emphasize that, contrary to a test suite where the same lexical items are used many times in a controlled form, in CorTA, with TrAva, we can collect cases that display long-distance unforeseeable dependences and which would never even get addressed by more systematic means.[6] Real running text is always preferable for evaluation of real systems in the real world, especially if one's intentions are not limited to evaluating a few already known phenomena.

## Acknowledgments

The systems described in the present paper were implemented solely by the third author. The authors thank Fundação para a Ciência e Tecnologia for the grant POSI/PLP/43931/2001, co-financed by POSI. We thank all participants in the ARTUR interest group for their support and comments, as well as everyone who used TrAVa and provided us with their feedback. We also thank the members of the Policarpo project at NILC and of TRADAUT-PT at FLUNL for making their own approaches to MT testing available to ARTUR. Finally, Anabela Barreiro helped shape and improve several versions of TrAva with her feedback, and was responsible for devising several alternative input forms. The work reported here owes a lot to her being an enthusiastic power user of both TrAva and METRA.

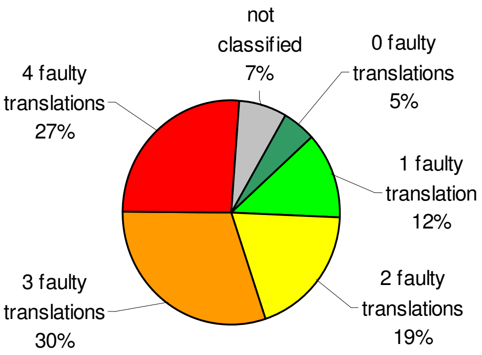


**Figure 2 - Distribution of user classifications in TrAva (1270 sentences)**

## References

Aires, R., Sandra Aluísio, P. Quaresma, Diana Santos & Mário Silva (2003). An initial proposal for cooperative evaluation on information retrieval in Portuguese. In Mamede et al. (eds.), Computational Processing of the Portuguese Language, 6th International Workshop, PROPOR 2003, Proceedings, Springer, pp. 227-234.

ALPAC: Automatic Language Processing Advisory Committee (1966). Language and machines: computers in translation and linguistics. Division of behavioral sciences, National Research Council, National Academy of Sciences, Washington.

Arisholm, Erik, Dag Sjøberg, Gunnar J. Carelius & Yngve Lindsjørn (2002). A Web-based Support Environment for Software Engineering Experiments. Nordic Journal of Computing 9 (4), 231-247, 2002.

Aston, Guy & Lou Burnard (1996). The BNC Handbook: Exploring the British National Corpus with SARA. Edinburgh: Edinburgh University Press.

Bar-Hillel (1960). Automatic Translation of Languages. In D. Booth & R.E. Meager (eds.), Advances in Computers. New York: Academic Press.

Christ, O., Schulze, B. M., Hofmann, A., & Koenig, E. (1999). The IMS Corpus Workbench: Corpus Query Processor (CQP): User's Manual. University of Stuttgart, March 8, 1999 (CQP V2.2).

Dabbabdie, M., A. Hartley, M. King, K.J. Miller, W.. M. El Hadi, A. Popescu-belis, F. Reeder & M. Vanni (2002). A Hands-On Study of the Reliability and Coherence of Evaluation Metrics. In M. King (ed.), Machine Translation Evaluation – Human Evaluators Meet Automated Metrics, Workshop Proceedings, LREC'2002, pp. 8--16.

King, M. & K. Falkedal. (1990). Using Test Suites in the Evaluation of Machine Translation Systems. In Proc. of the 13th International Conference on Computational Linguistics (COLING), Helsinki, pp.211--216.

Popescu-Belis, A., M. King & H. Bentanar (2002). Towards a corpus of corrected human translations. In M. King (ed.), Machine Translation Evaluation – Human Evaluators Meet Automated Metrics, Workshop Proceedings, LREC'2002, pp. 17--21.

Santos, Diana (2002). DISPARA, a system for distributing parallel corpora on the Web. In Ranchhod, E. & N.J. Mamede (eds.), Advances in Natural Language Processing (Third International Conference, PorTAL 2002,), Springer, pp. 209--218.

Santos, Diana & Anabela Barreiro (2004). On the problems of creating a consensual golden standard of inflected forms in Portuguese. In Proc. LREC'2004 (Lisbon, May 2004).

Santos, Diana, Luís Costa & Paulo Rocha (2003). Cooperatively evaluating Portuguese morphology. In Mamede et al. (eds.), Computational Processing of the Portuguese Language, 6th International Workshop, PROPOR 2003, Proceedings, Springer, pp.259--66.

Santos, Diana & Paulo Rocha (forthcoming). CHAVE: topics and questions on the Portuguese participation in CLEF. Proc. CLEF 2004 Working Notes (Bath, 16-17 September 2004).

Sarmento, Luís, Belinda Maia & Anabela Barreiro (forthcoming). O processo de criação do TrAva e do CorTA. In Santos, Diana (ed.), Avaliação conjunta: um novo paradigma no processamento computacional da língua portuguesa.

[6] A typical case is an NP erroneously torn between two different clauses in the source analysis, and therefore displaying lack of agreement in the translation. Independently of how the right error classification should be assigned in these cases, they clearly pinpoint errors that can only be observed and understood in a larger context.

# Compiling and Using a Shareable Parallel Corpus for Machine Translation Evaluation

**Debbie Elliott, Eric Atwell, Anthony Hartley**
School of Computing and Centre for Translation Studies, University of Leeds, Leeds LS2 9JT
debe@comp.leeds.ac.uk, eric@comp.leeds.ac.uk, a.hartley@leeds.ac.uk

### Abstract

TECMATE is a dynamic TEchnical Corpus for MAchine Translation Evaluation currently being compiled and used at the University of Leeds. A purpose-built corpus for machine translation (MT) evaluation differs in terms of size and content from corpora used for other kinds of linguistic analysis. For example, our research in automated MT evaluation requires source texts with human and machine translations as well as the scores for these translations given by human judges. These scores will allow us to test the reliability of experimental automated evaluation methods. Furthermore, a representative sample of machine translations annotated with fluency errors is also required to guide our research into automated error detection. In this paper, we summarise our rationale for corpus design and describe the different stages of corpus development. We provide an example of the content for one language pair and present findings from our recent evaluations of MT output using texts from the French-English sub-corpus. TECMATE will shortly be available online for research.

## Introduction

Shareable corpora for MT evaluation research are lacking. The largest known freely available resource is the DARPA corpus (White & O'Connell, 1994), which has been widely used for the testing of new automated evaluation methods (eg. Rajman & Hartley, 2002; White & Forner, 2001; Reeder et al., 2001; Vanni & Miller, 2002). The fluency, adequacy and informativeness scores associated with the translations from the corpus have been used to validate or reject experimental automated evaluation methods, enabling the investigation of correlations between human and automated scores. Although a valuable resource, the DARPA corpus has its limitations; all texts are newspaper articles, representing only a small part of MT use; the 300 source texts are in only three languages (French, Spanish and Japanese) and all human and machine translations are in American English. It is our intention, therefore, to provide a shareable resource that will complement the DARPA corpus.

## Rationale for Corpus Design

### Corpus Size

Before text collection began, informed decisions had to be taken with respect to corpus size. A large corpus would be impractical for human MT evaluation, as the greater the number of source texts, the more expensive and time-consuming it would be to evaluate the translations. Furthermore, our own research would require expert human translations of each text for comparison against MT output, and 'reference translations' (conveying the content of the source text without stylistic flourishes) to enable monolinguals to evaluate the fidelity of both the human and machine translations. These human translations are expensive to produce.

A large number of texts is not necessary for MT system comparison if reliable evaluation results can be obtained from a smaller corpus. We carried out a statistical analysis of the DARPA scores, for all three language pairs, to determine how many texts would be required to reliably compare MT systems. Results showed that for adequacy, fluency or informativeness evaluations, ten texts (approx. 3,500 words) would be sufficient to rank MT systems, and no more than forty texts (14,000 words) would be needed to offer a clear picture of system performance (Elliott et al., 2003).

### Text Types

In 2003, we conducted a worldwide survey of MT users to guide corpus design. The main purpose of the survey was to determine which text types were most frequently translated using MT systems and should, therefore, be represented in our corpus. Responses showed a great difference between the use of MT by companies/ organisations and by individuals who machine translate documents for personal use (Elliott et al., 2003). Individuals most often translated various kinds of web pages, followed by academic papers and newspaper texts. Companies, on the other hand, most frequently machine translated user manuals and technical documents on a large scale. As a result, the decision was taken to represent these texts in our corpus, along with a smaller number of legislative and medical documents. Corporate use of MT put newspaper texts in twelfth place.

### Language Pairs

Texts in a number of language pairs (translations into and out of English) will be required to test the portability of automated evaluation methods. To date, the French-English and English-French sub-corpora are complete, and we are currently working on the Spanish, German and Italian into English language pairs. We hope to add further language pairs, including typologically different languages at a later stage.

## Corpus Development

Text collection began with the French-English, followed by the English-French sub-corpus. Appropriate parallel texts in other language pairs were also discovered during this process. Our initial aim was to find French original texts with existing good quality human translations. Most freely available parallel corpora were unsuitable for our needs. However, extracts from technical reports were obtainable from the BAF Corpus[1]. The remaining documents were mined from the Web.

Finding good quality translations was a difficult task. Many were badly written, often by non-native speakers, and others, although of excellent quality, were localised to such an extent that they were unusable for MT evaluation. Obtaining copyright permissions was an arduous task, so methods were used to locate suitable documents that contained a permission notice to copy, distribute and modify the text and/or translations. Searches for "Guide de l'utilisateur" + "reproduction permitted" and "logiciel libre" + "copyleft" gave useful results, and many texts produced under the GNU Free (software) Documentation Licence and by the Free Software Foundation Europe were selected.

Although technical in nature, texts were chosen on the basis that they would be understandable to regular users of computer applications, enabling evaluators to confidently judge the quality of the translations.

All selected source texts and translations were checked for errors and translation correspondence. A number of corrections were made, as only perfect input and 'gold standard' translations would enable us to reliably evaluate the quality of the MT output. An English reference translation was then produced for each text. Machine translations of all source texts were generated from three commercial systems (Systran, Reverso Promt and Comprendium) and one online system (SDL's FreeTranslation).

## Corpus Content

Each language pair comprises forty source texts of approximately 400 words (equal to the longer texts in the DARPA corpus), and the same categories of text types:

- 10 software user manuals (extracts)
- 10 technical press releases
- 5 technical FAQs (Frequently Asked Questions)
- 5 technical reports (extracts)
- 5 legislative documents (extracts)
- 5 medical documents (extracts)

(The press releases were included at a later stage to represent a greater variety of verb tenses, as the documents initially collected were found to contain mostly imperative and present tense verbs.)

Each source text has an expert human translation, a reference translation, and currently four machine translations. The size of each sub-corpus is approximately 110,000 words. Expert human translations and machine translations will have three human evaluation scores per segment (usually a sentence or heading) for both fluency and adequacy; due to the subjective nature of translation evaluation, one score per segment is insufficient. In addition to these scores, the machine translations of twelve of the source texts (around 20,000 words in total) have been annotated with errors using the Systemic Coder[2] and our new fluency error categorisation scheme.

## Evaluation of MT Output

### Texts and Evaluators

In our first evaluation, the five translations of a sample of twelve source texts from the French-English sub-corpus were evaluated by thirty monolingual native speakers of English (mostly postgraduate students at the University of Leeds) who had little or no knowledge of French. The intention was to prevent untranslated words in the machine translations from being understood, therefore influencing evaluator judgements.

### Design of the Experiment

To provide detailed scores for comparison with results from our new automated evaluation methods, we required translations to be judged at segment level. Each evaluator rated one translation of each source text; judging six translations for fluency and six for adequacy. Both evaluations were based on the DARPA methods. To avoid the "training effect" no evaluator saw more than one translation of the same text.

Thirty evaluator packs were compiled, each comprising translations from different systems in different orders. As every translation would be judged for each attribute by three different evaluators, the same translation would appear in a different position in each pack, preventing the text order from affecting judgements. In half of the packs, the six fluency evaluations appeared first; the other half began with the adequacy evaluations. Judges were not told that the texts were translations. Scores were entered electronically to facilitate their collation and avoid transcription errors.

### Fluency

With access only to the translation, evaluators rated each "candidate segment" (most often a sentence or heading) using the Fluency Metric (Figure 1). To simplify the metric, judges were not provided with definitions for scores 2, 3 and 4. For both evaluations, they were asked

[1] http://www-rali.iro.umontreal.ca/arc-a2/BAF/Description.html

[2] http://www.wagsoft.com/Coder/index.html

not to go back to a segment once a judgement had been made.

**Fluency**
Look carefully at each segment and give each one a score according to how much you think the text reads like fluent English written by a native speaker. Give each segment of text a score of 1, 2, 3, 4, or 5 where:
**5 = All** of the segment reads like fluent English written by a native speaker
**1 = None** of the segment reads like fluent English written by a native speaker

Figure 1: Fluency Metric

## Adequacy

Judges compared the “candidate text” segments with the aligned "reference text” (reference translations) and used the Adequacy metric (Figure 2) to score each segment.

**Adequacy**
For each segment, read carefully the reference text on the left. Then judge how much of the same *content* you can find in the candidate text, *regardless of grammatical errors, spelling errors, inelegant style or the use of* synonyms. Give each segment of text a score of 1, 2, 3, 4, or 5 where:
5 = All of the content in the reference text is present in the candidate text
1 = None of the content is present (OR the text completely contradicts the information given on the left hand side).

Figure 2: Adequacy Metric

## Results

Three scores were obtained for each segment for each of the two evaluations. A mean score was the calculated per segment of each translation. These scores were used to generate a mean score per text and subsequently per system. Figure 3 and Figure 4 summarise the human evaluation results for both fluency and adequacy.

| System | Fluency Score | Adequacy Score |
|---|---|---|
| FreeTranslation | 2.827 | 3.644 |
| Comprendium | 3.221 | 4.013 |
| Reverso | 3.466 | 4.142 |
| Systran | 3.519 | 4.136 |
| Human | 4.893 | 4.826 |

Figure 3: Mean Segment Scores by System

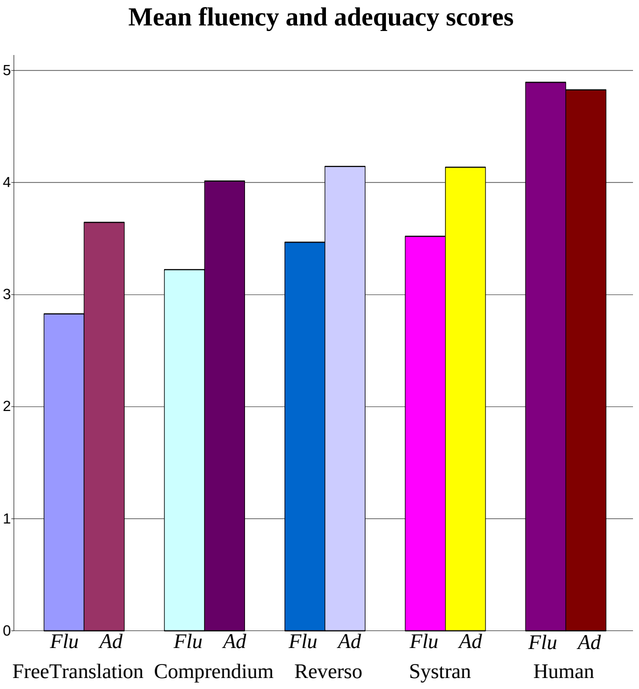


Figure 4: Comparison between MT and Human Translation scores

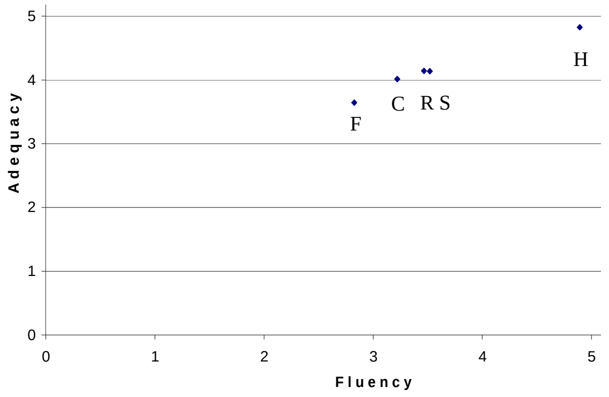


Figure 5: Association between fluency and adequacy values for each system

Systran was the highest scoring MT system for fluency and Reverso for adequacy, by a very small margin. FreeTranslation was the lowest scoring system for both attributes. The machine translations scored consistently more highly for adequacy, indicating that despite a lower level of fluency, the content of raw MT output can be useful. Conversely, there was little difference between the fluency and adequacy scores for the human translations. For all five ‘systems’, a high degree of association was found between values for the two attributes, as shown in Figure 5. Pearson’s correlation coefficient was used to test this hypothesis: using the mean system scores for fluency and adequacy in Figure 3, the value of *r* = 0.98803, showing a very strong correlation between the two variables. This correlation indicates that evaluating either fluency or adequacy would be sufficient to predict values for the other attribute. This supports earlier findings (eg. White, 2001).

## Evaluation Time Required

Each evaluator judged 327 segments, rating approximately half for adequacy and half for fluency. The average time taken to complete the fluency evaluation was 33 minutes. The adequacy evaluation contained more reading material and took 48 minutes on average to complete. Without including an introduction to the task, time needed to read instructions, and at least one break, 30 evaluators each required 81 minutes to complete the evaluations. Therefore, the total time needed to evaluate five translations of twelve texts amounted to 40.5 hours.

## Conclusions and Further Work

As our experiment shows, machine translation evaluation by humans is expensive and time-consuming. Not only does it involve the careful selection of source texts, often accompanied by good quality human translations, it also requires the preparation of materials (here, segmented aligned texts and metrics) and a sufficient number of human judges. However, these evaluations are necessary to create shareable corpora, with the added value of human scores, to allow for the testing of results from experimental automated evaluation methods.

In terms of corpus development, our next stage will involve the completion of existing language pairs and obtaining human judgements for a greater number of texts. We also plan to investigate correlations between human scores from our recent evaluation and the ranking of the same translations at text level (a cheaper way to evaluate).

We are currently fine-tuning our fluency error classification scheme for French-English machine translations. The annotated texts will be available as a component of the corpus at a later stage. Furthermore, we intend to extend the scheme to additional language pairs, to compare translation errors in English output from different source languages. Statistics resulting from the annotated texts will guide our selection of errors for automated detection. Finally, we will seek to validate our automated methods by using our corpus to find a correlation between human judgements on fluency and adequacy and automated scores.

Each sub-corpus of TECMATE will be made available online when completed. It is hoped that the texts will be of use for research in MT evaluation and other areas of translation studies.

## References

Elliott, D., Hartley, A & Atwell, E. (2003). Rationale for a multilingual corpus for machine translation evaluation. In Proceedings of CL2003: International Conference on Corpus Linguistics (pp. 191-200). Lancaster University, UK.

Rajman, M. & Hartley, A. (2002). Automatic Ranking of MT Systems. In Proceedings of the 3rd International Conference on Language Resources and Evaluation. Las Palmas de Gran Canaria, Spain.

Reeder, F., Miller, K., Doyon, K. & White, J. (2001). The Naming of Things and the Confusion of Tongues. In Proceedings of the 4th ISLE Evaluation Workshop, MT Summit VIII. Santiago de Compostela, Spain.

Vanni, M. & Miller, K. (2002). Scaling the ISLE Framework: Use of Existing Corpus Resources for Validation of MT Evaluation Metrics across Languages. In Proceedings of the 3rd International Conference on Language Resources and Evaluation. Las Palmas de Gran Canaria, Spain.

White, J. (2001). Predicting Intelligibility from Fidelity in MT Evaluation. In Proceedings of the 4th ISLE Evaluation Workshop, MT Summit VIII. Santiago de Compostela, Spain.

White, J. & Forner, M. (2001). Predicting MT fidelity from noun-compound handling. In Proceedings of the 4th ISLE Evaluation Workshop, MT Summit VIII. Santiago de Compostela, Spain.

White, J. & O'Connell, T. (1994). The ARPA MT evaluation methodologies: evolution, lessons, and future approaches. In Proceedings of the 1994 Conference, Association for Machine Translation in the Americas. Columbia, MD.

# Improving Word Alignment in an English – Malay Parallel Corpus for Machine Translation

**Suhaimi Ab. Rahman** **Normaziah Abdul Aziz, Ph.D**

Language Engineering Research Lab
MIMOS
Malaysia Technology Park
57000 Kuala Lumpur, Malaysia
smie@mimos.my, naa@mimos.my

## Abstract

A bilingual parallel corpora is an important resource in constructing an English – Malay Bilingual Knowledge Base that is heavily referred to in our English to Malay machine translation system. We present an approach that we applied at word level alignment from a bilingual parallel corpora to improve the translation quality of our English to Malay Example-based machine translation. Initially, *one-to-one* word alignment was applied against the source and target languages. We revised this method to a *many-to-one* word alignment. The comparison of translation results for both method shows that our *many-to-one* word alignment is capable to improve the translation quality.

## 1. Introduction

We have compiled an English-Malay bilingual parallel corpus consist of 250,000 words in the domain of agriculture and health. These two domains were put in placed as an initial deployment of the English-Malay machine translation service to the rural community users. This research and development (R&D) project's goal is to address the language barrier issue as part of narrowing the Digital Divide problems in Malaysia[1]. In country such as Malaysia where English is a second language for majority of its people, language is one of the factors that ought to be addressed in the digital divide issues. Hence, the need of tools such as online machine translation systems is important to ensure that the non-English speakers community could broaden their knowledge resources unlimited to their native language as discussed in (Aziz, N., et al, 2002).

We started our applied research work with University of Science Malaysia's (USM) prototype machine translation. USM's works surrounding this research have been described in various technical platforms such as (al Adhaileh, Tang, 2001) and (al Adhaileh, Tang, Zahrin, 2002), among others. We continued the work by upgrading USM's *proof-of-concept* version to a *real usage* version for deployment to the digital divide communities.

Work on alignment of parallel corpus for machine translation, sense disambiguation, information retrieval for multilingual environment and other language related researches have been actively discussed at various perspective and levels of discussions such as in (Chen, 1993), (Dagan, Church and Gale, 1993), (Gaussier, 1998), and (Ahrenberg, et.al., 2000) among others. However, our discussion in this paper is based on an experience that we encountered while developing and testing in upgrading a prototype machine translation and not out of a theoretical research exercise. Referring to our English-Malay parallel corpora, this paper discusses on how we revised the word level alignment from the bilingual parallel corpora to improve the translation quality of the English to Malay Example-based machine translation (EBMT).

## 2. Parallel Corpora for EBMT

A bilingual of English and Malay parallel corpora is a significant resource in constructing an English – Malay Bilingual Knowledge Base (BKB). This BKB is heavily referred to in our English to Malay example-based machine translation system.

As for the process of alignment in our English-Malay parallel corpora, initially, an auto sentence alignment process together with an English-Malay dictionary mapping are applied to align our English-Malay parallel text. An alignment algorithm that uses English-Malay dictionary mapping offers a potential for higher accuracy of word alignment that leads to better translation quality. After these two processes, we manually review the result of English-Malay bi-texts through post-editing to improve the English – Malay word alignment. The bilingual parallel text will be used in constructing a bilingual knowledge bank (semi-) automatically through

[1] This work is funded by MIMOS' Bridging Digital Divide Programme, under the 8th Malaysia Plan.

available parsers and alignment tools. A representation schema named *Synchronous Structured String-Tree Correspondence* (S-SSTC) is used to annotate the translation example pairs, describing the correspondence relation between the source and target sentences (Al-Adhaileh, 2002).

Referring to the alignment process for content of our parallel corpora, here we are addressing the problem of "many English words to be represented in one Malay word", which then improves the linguistic quality of translation by our English to Malay EBMT. The following are a few cases of English words and its translation of Malay word(s). Note that in these cases, the number of the translated word(s) of the target language is one or lesser than the number of words from its source language. Using such replacement will produce a better translation.

| | |
|---|---|
| as well as --- juga | in order to --- supaya |
| as long as --- selagi | such as --- seperti |

The following Figure 1 shows some examples with the above words in the bilingual parallel text that is used for our EBMT.

| **English (E)** | **Malay (M)** |
|---|---|
| E1 : Wild flowers ***such as*** orchids and primroses are becoming rare. | M1 :Bunga-bunga hutan ***seperti*** orkid dan primros semakin jarang ditemui. |
| E2 : ***As long as*** you maintain your diet, you don't need to worry about your health. | M2 ***: Selagi*** anda menjaga pemakanan anda, anda tidak perlu risau mengenai kesihatan anda. |

Figure 1: Example of the English-Malay Bilingual Corpus.

## 3. Word Level Alignment

It is necessary to align the two texts of the target and source language to extract information from the parallel corpora. The alignment process is meant to associate chunks of text in the source language document with the ones of the translated version in the target language as discussed in (Somers, H.) In our work, the alignment is done at sentence and word level.

The initial auto alignment algorithm at word level splits each word in a sentence, one by one. We refer to this approach as *one-to-one* word alignment method. Due to the nature of Malay and English at linguistic level, there are instances where several words in English are best represented or translated to one Malay word. This is also true at instances where one English word needs to be represented in a few Malay words, when translated. However, here we are addressing "English phrases that is to be represented in one Malay word" alignment.

### 3.1 Many-to-One Approach

The processes that are involved in our many-to-one word alignment are as follows:

1. Get the aligned source sentence.
2. Generate the list of word-form from the source sentences from lexicon parser process.
3. In order to get many-to-one word, the process will refer to the lexicon parser, which contains phrases based on English grammar.
4. The logical dictionary mapping is used to retrieve the meaning for each word-form.
5. Improve the word-level alignment output, when necessary by manual post-editing.

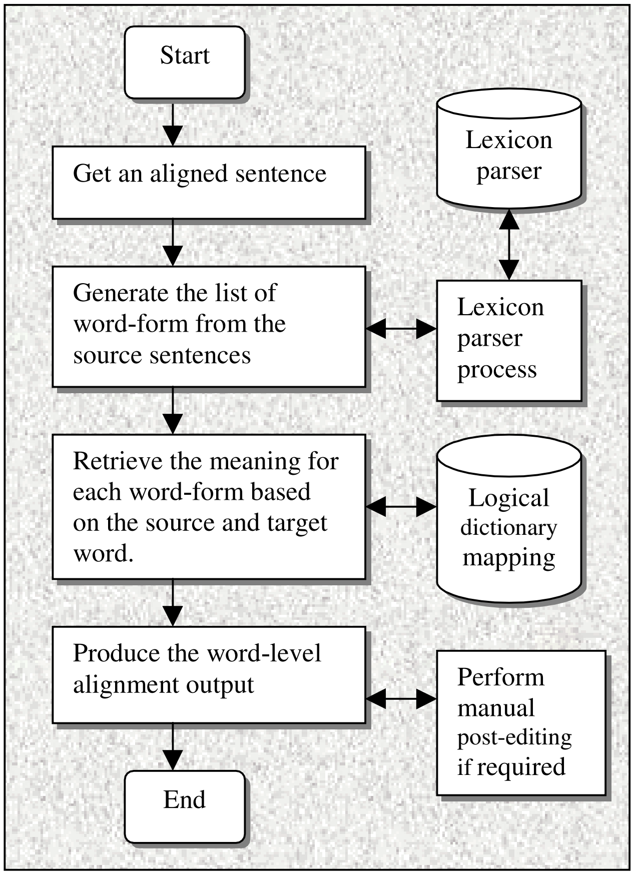

Start
Get an aligned sentence
Lexicon parser
Generate the list of word-form from the source sentences
Lexicon parser process
Retrieve the meaning for each word-form based on the source and target word.
Logical dictionary mapping
Produce the word-level alignment output
Perform manual post-editing if required
End

Figure 2: The process involved in the many-to-one approach word alignment

In this process, the lexicon parser is of great use where we can add the phrases related to this *many-to-one* word alignment issue. It contains the identified phrase together with its lexicon tag.

Besides that, we further improve the algorithm which then eliminates redundancy data and thus, reduce its run-time. In other words, i) after compiling the *many-to-one* word-level alignment, we manage to reduce the number of alignment between source word and target word; ii) reduce the occurrence number of *null words* returned after the *many-to-one* alignment being made; where, null word is the word-level alignment that carries no meaning. With these steps taken, we have a better version of parallel corpus to work on.

### 3.2 The Parallel Corpus Alignment

Below is an example of the different word-level alignment using *one-to-one* and *many-to-one* for the following English-Malay corpus.

**Source (English) :**
$_{0}$The $_{1}$doctor $_{2}$advises $_{3}$her $_{4}$to $_{5}$rest $_{6}$***as*** $_{7}$***long*** $_{8}$***as*** $_{9}$she $_{10}$needs$_{11}$.$_{12}$

**Target (Malay) :**
$_{0}$Doktor $_{1}$menasihati $_{2}$perempuan $_{3}$itu $_{4}$untuk $_{5}$berehat $_{6}$selagi $_{7}$dia $_{8}$perlu$_{9}$.$_{10}$

The alignment of both source and target sentences are described in Figure 4 a) in a one-to-one alignment approach and Figure 4 b) in a many-to-one alignment method. The dependency tree for each approaches are also shown respectively.

## 4. Test Results

We assign to the English corpus *E* translating to the Malay corpus *M* with a particular alignment. For example, sentence $E_1$ corresponds to the target sentence $M_1$. From the parallel corpus ($E_1$,$M_1$) in Figure 4, it shows the difference alignment output generated by using *one-to-one* and *many-to-one* methods. The phrase word ***$_{0}$such as$_{1}$*** from one-to-one method is separated into two words: ***$_{0}$such$_{1}$as$_{2}$***. Meanwhile, many-to-one method combined the phrase word ***$_{0}$such$_{1}$as$_{2}$*** into one word ***$_{0}$such as$_{1}$***. The combination of this phrase word ***$_{0}$such as$_{1}$*** is produced in the lexicon parser process. Other sentences which contain identified phrase that are in the lexicon parser will go through the same process as described.

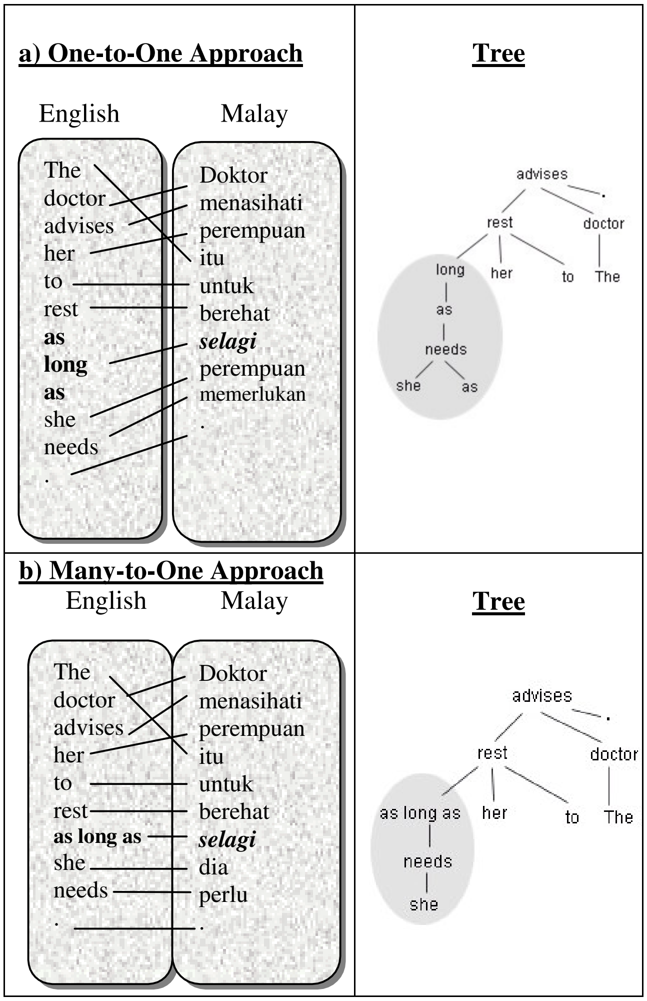


Figure 4: Word-level alignment for the translation pair (a) one-to-one approach. (b) many-to-one method.

The following Figure 5 shows the different results of word level alignment using one-to-one and many-to-one word alignment method.

| Example : Bilingual Corpus (E,M):<br>$E_1$ : Wild flowers ***such as*** orchids and primroses are becoming rare.<br>$M_1$ : Bunga-bunga hutan seperti orkid dan primros semakin jarang ditemui. | |
|---|---|
| **One-to-one word alignment method** | **Many-to-one word alignment method** |
| Wild -- > hutan | Wild -- > hutan |
| flowers -- > Bunga-bunga | flowers -- > Bunga-bunga |
| ***such -- > seperti*** | ***such as -- > seperti*** |
| ***as -- > null*** | orchids -- > orkid |

<table>
<tr><td>orchids  -- &gt; orkid</td><td>and  -- &gt;  dan</td></tr>
<tr><td>and  -- &gt;  dan</td><td>priroses  -- &gt; primros</td></tr>
<tr><td>primroses  -- &gt; primros</td><td>are -- &gt; null</td></tr>
<tr><td>are -- &gt; null</td><td>becoming -- &gt; semakin</td></tr>
<tr><td>becoming -- &gt; semakin</td><td>rare -- &gt; jarang ditemui</td></tr>
<tr><td>rare -- &gt; jarang ditemui</td><td>. -- &gt; .</td></tr>
<tr><td></td><td></td></tr>
<tr><td colspan="2"><b>Test:</b></td></tr>
<tr><td colspan="2"><b><i><u>New input sentence</u></i></b><br>E1: You have to eat more vegetables <b><i>such as</i></b> salad, spinach and mustard.</td></tr>
<tr><td colspan="2"><b>Results:</b></td></tr>
<tr><td colspan="2"><b><i><u>Translation results using one-to-one word alignment</u></i></b><br>M1a:   Anda ada untuk makan lebihan banyak sayuran sebagai seperti salad, bayam  dan sawi.</td></tr>
<tr><td colspan="2"><b><i><u>Translation results using many-to-one word alignment</u></i></b><br>M1b:   Anda perlu makan banyak sayuran seperti salad,bayam dan sawi.</td></tr>
</table>

Figure 5:  Word level alignment, testing and result using *one-to-one* and *many-to-one* approach.

Referring to the test above, there are two different results generated form the EBMT system. The example of the input sentence focusing to the phrase word $_0$***such as***$_1$. By referring *one-to-one* method, it shows that the translation is more on word-to-word translation. This is because i) the dependency tree or sub-tree for the phrase word $_0$***such as***$_1$ is not found in our Bilingual Knowledge Base (BKB), in the context of the input sentence; and ii) the phrase word $_0$***such as***$_1$ is not in the lexicon parser. Meanwhile, for *many-to-one* method, the translation is more accurate because i) the dependency tree or sub-tree of the phrase word $_0$***such as***$_1$ found in the BKB; and ii) the lexicon parser process found the phrase word $_0$***such as***$_1$ in the lexicon parser.

We revised the *one-to-one* auto alignment to a *many-to-one* word alignment for relevant cases. After running several test data of 100 English sentences with such words (e.g. as long as, such as, years old, in order to), we discovered that this *many-to-one* word alignment manage to ensure the construction of a more accurate of our Bilingual Knowledge Base, thus better quality of translation result i.e. from the Malay linguistic perspective.

## 5. Conclusion

The discussion above shows that the translation improvement could be made via a *many-to-one* word alignment of a bilingual parallel corpus, in the context of an English and Malay parallel text. The improvement is significant to us when we are refining the translation quality (from the perspective of Malay language). At the same token, we managed to reduce the processing time at a factor of 4 for searching the proper word alignment between the source and target word in the bilingual parallel corpora.

## References

Al-Adhaileh, M.H. (2002). Synchronous Structured String Tree Correspondence (S-SSTC) and its Application for  Machine Translation, PhD Thesis, University of Science Malaysia.

Al-Adhaileh Mosleh H. & Tang Enya Kong. (2001). Converting a Bilingual Dictionary into a Bilingual Knowledge Bank based on the Synchronous SSTC. Proceedings of MT Summit VIII, Santiago de Compostela, Spain, 18 Sept 2001.

Al-Adhaileh, Mosleh H., Tang Enya Kong and Zaharin Yusoff. (2002).  A Synchronization Structure of SSTC and its  Applications in Machine Translation. The COLING 2002 Post-Conference Workshop on Machine Translation in Asia, Taipei, Taiwan.

Aziz, N., et al. (2002).  Is Machine Translation Still Relevant?, in MIMOS 2002 Tech-Symposium Proceedings.

Chen, S. F. (1993) . Aligning sentences in bilingual corpora using lexical information. In Proceedings of ACL-93, Columbus OH.

Dagan, I., Church, K. W., and Gale, W. A. (1993). Robust Bilingual Word Alignment for Machine Aided Translation.  In Proceedings of the Workshop on Very Large Corpora: Acad. & Industrial Perspectives, Columbus OH.

Gaussier, E. (1998). Flow Network Models for Word Alignment and Terminology Extraction from Bilingual Corpora. In Proceedings of COLING-ACL-98, Montreal, (pp. 444-450).

Lars Ahrenberg, Magnus Merkel, Anna Sågvall Hein & Jörg Tiedemann (2000).  Evaluating Word Alignment Systems. Proceedings of the Second International Conference on Linguistic Resources and Evaluation (LREC-2000), Athens, Greece, 31 May - 2 June, 2000, Volume III: 1255-1261.

Somers, H., Bilingual Parallel Corpora and Language Engineering, available at http://www.emille.lancs.ac.uk/lesal/somers.pdf

# Alignment of Parallel Corpora Exploiting Asymmetrically Aligned Phrases

**Patrik Lambert and Núria Castell**

TALP Research Center
Universitat Politècnica de Catalunya
Jordi Girona Salgado, 1-3, 08034 Barcelona, Spain
{lambert,castell}@talp.upc.es

**Abstract**

This paper presents a simple way of producing symmetric, phrase-based alignments, combining two single-word based alignments. Our algorithm exploits the asymmetries in the superposition of the two word alignments to detect the phrases that must be aligned as a whole. It was run with baseline word alignments produced by the Giza++ software and improved these alignments. The ability to treat some groups of words as a whole is essential in applications like machine translation. The paper also addresses the difficulty of the alignment evaluation task.

## 1. Introduction

A parallel corpus aligned at a word level is a resource directly usable for the building of bilingual lexica and terminology. It is also a valuable resource for several natural language processing applications such as machine translation and word sense disambiguation. A publicly available, widely used software to produce baseline single word based alignments is Giza++ (Och, 2000; Och and Ney, 2003). It implements various translation models: the so-called IBM models 1 to 5, introduced by (Brown et al., 1993), and the HMM model, introduced by (Vogel et al., 1996). The models are trained on a bilingual corpus with the EM algorithm (Baum, 1972), "bootstrapping" from a simpler model to a more complex model. The final alignment (Viterbi alignment) is the best one according to a Viterbi search.

The models implemented by the Giza++ software have limitations. The first one is a consequence of the mapping used, which only allows to link one source word to each target word. The second one is inherent to single-word based alignments: alignment of multiple word or phrases which do not decompose easily in word-into-word translations are not possible.

As pointed out in (Och and Ney, 2003), the first problem can be solved if the Viterbi alignment is calculated in both source-target and target-source directions. If the alignment in one direction is not complete, the alignment in the other direction completes it. The combination of source-target and target-source alignments is also a useful resource to detect the second problem. This is because the phrases that cannot be aligned word-to-word (like idiomatic expressions) are not well aligned by Giza++, so that the source-target and target-source alignments are typically not symmetric.

In section 2., we present an algorithm that detects these asymmetries in the superposition of source-target and target-source alignments, and replaces them by appropriate symmetric alignments. Section 3. discusses the alignment evaluation task. Section 4. describes the experiments. Some conclusions are given in section 5..

## 2. Symmetrisation Algorithm

The central idea is that if the asymmetry is caused by a language feature such as an idiomatic expression, it will be repeated various times in the corpus, otherwise it will occur only once. Our symmetrisation process has the following two stages:

**Building of asymmetries memory.** Detect all the asymmetries present in the corpus and store them with their number of occurrences. A word does not belong to an asymmetry if it is linked to exactly one word, which in turn has exactly one link to it.

**Alignment correction.** Detect again asymmetric zones and for each asymmetry, try to correct the alignment:

1. Look if the limitation associated to the mapping can be solved: if the asymmetry contains various words linked to a word x, itself aligned to only one of them, links are added so that x be aligned to the other words.
2. Look if the asymmetry contains phrases qualified to be aligned as a group: it should include at least one source and one target word. Two parts of a non-contiguous phrase can't be more than three words away from each other. If the asymmetry is suitable for group alignment, follow steps 3 and 4. Otherwise, the asymmetry has generally no linguistic basis and it is advisable to take the intersection of source-target and target-source alignments.
3. Split the source and target strings in fragments, combine each source fragment with each target fragment and see how many times the combination has occurred in an asymmetry. Select the combination that has occurred more times in the corpus. If it is above a predefined threshold, add links so that both fragments be aligned as a group (many-to-many alignment). Continue with the other fragments until all words have been grouped or until no remaining combination has more than the threshold number of occurrences in the corpus.
4. If no combination had occurred more than the threshold, apply a combination of source-target and target-source alignments, like their intersection or union.

## 3. Alignment Evaluation

A consensus on word alignment evaluation methods has started to appear. These methods are described in (Mihalcea and Pedersen, 2003). Submitted alignments are compared to a manually aligned reference corpus (gold standard) and scored with respect to precision, recall, F-measure and Alignment Error Rate (AER). An inherent problem of the evaluation is the ambiguity of the manual alignment task. The annotation criteria depend on each annotator. Therefore, (Och and Ney, 2003) introduced a reference corpus with explicit ambiguous (called P or Possible) links and unambiguous (called S or Sure) links. Given an alignment $\mathcal{A}$, and a gold standard alignment $\mathcal{G}$, we can define sets $\mathcal{A_S}$, $\mathcal{A_P}$ and $\mathcal{G_S}$, $\mathcal{G_P}$, corresponding to the sets of Sure and Possible links of each alignment. The set of Possible links is also the union of S and P links, or equivalently $\mathcal{A_S} \subseteq \mathcal{A_P}$ and $\mathcal{G_S} \subseteq \mathcal{G_P}$. The following measures are defined (where $T$ is the alignment type, and can be set to either S or P):

$$P_T = \frac{|A_T \cap G_T|}{|A_T|}, \;\; R_T = \frac{|A_T \cap G_T|}{|G_T|}, \;\; F_T = \frac{2P_T R_T}{P_T + R_T}$$

$$AER = 1 - \frac{|A_P \cap G_S| + |A_P \cap G_P|}{|A_P| + |G_S|}$$

Note that $|G_P|$ is not taken into account in the AER. Therefore, including more P links in the reference alignment can only lower the error rate. The definition also implies that if $G_S \subseteq A_P \subseteq G_P$, the AER is equal to zero.

The next step in the evaluation is to be able to compare the values obtained. However, it is a delicate task because they are very dependent on the exact method used as well as on the reference corpus.

### 3.1. Influence of the Evaluation Method

The scores are greatly affected by the representation of NULL links (between a word and no other word: whether they are assigned an explicit link to NULL or removed from the alignments). Explicit NULL links contribute to a higher error rate because in this case the errors are penalised twice: for the incorrect link to NULL and for the missing link to the correct word.

Another influent factor is the way of weighting each link: $n$ words linked as a group represent $n^2$ links instead of $n$ links. To correct this effect, (Melamed, 1998) proposed to attach a weight to each link. The weight $w(\mathrm{x}, \mathrm{y})$ of a link between two words x and y would be inversely proportional to the number of links in which x and y are involved.

In conclusion, experiments are not comparable unless they are evaluated with exactly the same method.

### 3.2. Influence of the Reference Corpus

Apart from their dependence in the annotator's criteria (the decision of what is translation of what), the results vary in function of the proportion of ambiguous and unambiguous links. If the reference corpus contains a small number of very sure S links and many P links, adding more links to the submitted alignment will only slightly modify the value of $|A_P \cap G_S|$ and $|A_P \cap G_P|$ since they tend easily to $|G_S|$ and $|A_P|$, respectively. However the increase of $|A_P|$ will lower the AER. So this reference corpus will favour high precision alignments. On the contrary, if the reference corpus only contains S links, more submitted links will be needed to increase $|A_P \cap G_S|$ and high recall alignments will be more rewarded than in the previous case.

A related issue is that a reference corpus with many ambiguous links allows many different submitted alignments to have the same AER, while some of them are obviously poorer. Consider for instance the sentence pair 76 of the reference corpus of (Och and Ney, 2000), displayed in figure 1.

nous souhaitons parvenir à une décision cette semaine .
it is our hope to make a decision this week .

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| . | . | . | . | . | . | . | . | . | . | . | . | S |
| semaine | . | . | . | . | . | . | . | . | . | . | S | . |
| cette | . | . | . | . | . | . | . | . | . | S | . | . |
| décision | . | . | . | . | . | P | P | P | S | . | . | . |
| une | . | . | . | . | . | P | P | S | P | . | . | . |
| à | . | . | . | . | . | P | P | P | P | . | . | . |
| parvenir | . | . | . | . | . | P | P | P | P | . | . | . |
| souhaitons | . | P | P | P | P | P | . | . | . | . | . | . |
| nous | . | P | P | P | P | P | . | . | . | . | . | . |
| NULL | . | . | . | . | . | . | . | . | . | . | . | . |
| | NULL | it | is | our | hope | to | make | a | decision | this | week | . |

Figure 1: Example alignment with few Sure links and many ambiguous links

With such a reference, both alignments of figure 2 would get the same score of zero error rate (as well as all the alignments for which $G_S \subseteq A_P \subseteq G_P$), although the lower one is much poorer.

Therefore, if the gold standard contains ambiguous links, they should only allow alignment combinations that are considered equally correct.

## 4. Alignment Symmetrisation Experiments

We present results on two corpora. First we give their characteristics. Next, we detail the evaluation of the Giza++ alignments and their symmetrisation.

In all the experiments the NULL links were removed. Here we only show results in which each link has the same weight. The first 200 sentence pairs of each test corpus were used to optimise some parameters of the symmetrisation application (this doesn't require training). The whole test corpus, including these 200 sentence pairs, was used for the evaluation.

### 4.1. Training and Test Data

#### 4.1.1. Verbmobil Corpus

These data come from a selection of spontaneous speech databases available from the Verbmobil project[1].

[1] http://verbmobil.dfki.de/verbmobil

|  | NULL | it | is | our | hope | to | make | a | decision | this | week | . |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| . | . | . | . | . | . | . | . | . | . | . | . | S |
| semaine | . | . | . | . | . | . | . | . | . | . | S | . |
| cette | . | . | . | . | . | . | . | . | . | S | . | . |
| décision | . | . | . | . | . | . | . | . | S | . | . | . |
| une | . | . | . | . | . | . | . | S | . | . | . | . |
| à | . | . | . | . | . | . | . | . | . | . | . | . |
| parvenir | . | . | . | . | . | . | S | . | . | . | . | . |
| souhaitons | . | . | S | S | S | . | . | . | . | . | . | . |
| nous | . | . | S | S | S | . | . | . | . | . | . | . |
| NULL | . | . | . | . | . | . | . | . | . | . | . | . |

|  | NULL | it | is | our | hope | to | make | a | decision | this | week | . |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| . | . | . | . | . | . | . | . | . | . | . | . | S |
| semaine | . | . | . | . | . | . | . | . | . | . | S | . |
| cette | . | . | . | . | . | . | . | . | . | S | . | . |
| décision | . | . | . | . | . | . | . | . | S | . | . | . |
| une | . | . | . | . | . | . | . | S | . | . | . | . |
| à | . | . | . | . | . | . | . | . | S | . | . | . |
| parvenir | . | . | . | . | . | . | . | S | . | . | . | . |
| souhaitons | . | . | . | S | . | . | . | . | . | . | . | . |
| nous | . | . | . | . | . | S | . | . | . | . | . | . |
| NULL | . | . | . | . | . | . | . | . | . | . | . | . |

Figure 2: Two possible submission alignments with AER=0. Only the upper one is acceptable.

The databases have been selected to contain only recordings in US-English and to focus on the appointment scheduling domain. Then their counterparts in Catalan and Spanish have been generated by means of human translation (Arranz et al., 2003)[2]. Dates and times were categorised automatically (and revised manually). The test corpus consists of four hundred sentence pairs manually aligned by a single annotator. See the characteristics of the data in table 1.

|  |  | Spanish | English |
|---|---|---|---|
| Training | Sentences | 28000 ≈ 28K |  |
|  | Words | 201893 | 209653 |
|  | Vocabulary | 4894 | 3167 |
|  | Singletons | 2139 | 1251 |
| Test | Sentences | 400 |  |
|  | Words | 3124 | 3188 |

Table 1: Characteristics of Verbmobil corpus

#### 4.1.2. Hansards Corpus

The corpus consists of the debates in the 36th Canadian parliament. We used a version of the Hansards aligned by Ullrich Germann at the level of sentences or smaller fragments (Germann, 2001). From the over 1.3 million of parallel text chunks, we selected those of 40 words or less. The size of this corpus is much larger than that of Verbmobil and the domain much more open so that the vocabulary is very large (see table 2). The test data were created by Franz Och and Hermann Ney (Och and Ney, 2000). They contain a restricted set of sure links and a large set of possible links.

|  |  | French | English |
|---|---|---|---|
| Training | Sentences | 1008K |  |
|  | Words | 16,95M | 14,60M |
|  | Vocabulary | 76130 | 59534 |
|  | Singletons | 32644 | 24370 |
| Test | Sentences | 484 |  |
|  | Words | 8482 | 7681 |

Table 2: Characteristics of Hansards corpus

### 4.2. Giza++ Baseline

The first decision to take in the symmetrisation process is the default starting point, which is systematically selected when our algorithm can't find an adequate group (step 4 of the algorithm). Combining the source-target and target-source information of the Giza++ alignments, we can obtain a high precision with low recall alignment (taking the intersection), a low precision with high recall alignment (taking the union), or intermediate combinations. The evaluation of different possible sets are presented in table 3.

As outlined in section 3.2., the best combination depends on the reference corpus. Both reference corpora contain more links than the Giza++ alignments because they have many-to-many alignments whereas Giza++ only produces one-to-one alignments. For Verbmobil, the reference corpus contains only S links. The recall plays an important role and the union is the best combination. The reference corpus for the Hansards task contains few S links and many P links. The intersection is the best combination because it keeps fewer, more precise links.

Results with weighted links, as described in section 3.1., are presented in a research report (Lambert and Castell, 2004). In most cases the effect of the weighting of the links is simply to move up the scores. However for the Hansards corpus it produces a qualitative change: the intersection gets a score worse than the union.

### 4.3. Symmetrisation Evaluation

From the results of the previous section and further experiments, the default starting point of the symmetrisation was set to be the union (of source-target and target-source Giza++ alignments) for the Verbmobil corpus and their intersection for the Hansards corpus.

Table 4 presents the evaluation of the symmetrisation process in these two cases. The symmetrisation increases the recall but introduces also some noise, so the precision is lower. However the outcome is a decrease of the error rate from 18.6 to 17.7 in the case of Verbmobil, and from 9.1 to 7.4 in the case of Hansards. The larger effect in the case of the Hansards could be due to the much greater size of the asymmetries repository. This allows a higher coverage but also permits to increase the threshold number of occurrences of an asymmetry, which implies a gain in precision. This threshold number was 3 for the Hansards, and 2 for Verbmobil.

[2] It is referred to as “subset-1” in the paper

Verbmobil corpus

| Experiment | $P_S$ (%) | $R_S$ (%) | $F_S$ (%) | $P_P$ (%) | $R_P$ (%) | $F_P$ (%) | AER (%) |
|---|---|---|---|---|---|---|---|
| English to Spanish | 92.82 | 64.18 | 75.89 | 92.82 | 64.18 | 75.89 | 24.11 |
| Spanish to English | 93.95 | 67.51 | 78.57 | 93.95 | 67.51 | 78.57 | 21.43 |
| Intersection | **97.62** | 57.59 | 72.44 | **97.62** | 57.59 | 72.44 | 27.56 |
| Union | 90.37 | **74.11** | **81.43** | 90.37 | **74.11** | **81.43** | **18.57** |

Hansards corpus

| Experiment | $P_S$ (%) | $R_S$ (%) | $F_S$ (%) | $P_P$ (%) | $R_P$ (%) | $F_P$ (%) | AER (%) |
|---|---|---|---|---|---|---|---|
| English to French | 60.89 | 91.04 | 72.97 | 90.29 | 30.74 | 45.86 | 9.41 |
| French to English | 62.08 | 85.81 | 72.04 | 90.58 | 28.50 | 43.36 | 11.42 |
| Intersection | **74.06** | 82.79 | **78.18** | **98.10** | 24.97 | 39.80 | **9.13** |
| Union | 53.45 | **94.06** | 68.17 | 85.56 | **34.27** | **48.94** | 11.36 |

Table 3: Giza++ evaluation

Verbmobil corpus

| Experiment | $P_S$ (%) | $R_S$ (%) | $F_S$ (%) | $P_P$ (%) | $R_P$ (%) | $F_P$ (%) | AER (%) |
|---|---|---|---|---|---|---|---|
| Giza++ Union | 90.37 | 74.11 | 81.43 | 90.37 | 74.11 | 81.43 | 18.57 |
| Symmetrisation | 88.68 | 76.75 | 82.28 | 88.68 | 76.75 | 82.28 | 17.72 |

Hansards corpus

| Experiment | $P_S$ (%) | $R_S$ (%) | $F_S$ (%) | $P_P$ (%) | $R_P$ (%) | $F_P$ (%) | AER (%) |
|---|---|---|---|---|---|---|---|
| Giza++ Intersection | 74.06 | 82.79 | 78.18 | 98.10 | 24.97 | 39.80 | 9.13 |
| Symmetrisation | 65.05 | 89.49 | 75.34 | 94.92 | 29.73 | 45.27 | 7.37 |

Table 4: Evaluation of the symmetrisation process

## 5. Conclusions

We used the Giza++ application to produce symmetric, phrase-based alignments with lower alignment error rate. In fact, our symmetrisation process could be applied to any two alignments of the same sentence pairs. The resulting alignments can in turn improve those applications where aligned corpora are a valuable resource. For instance, the obtained alignments could be used as phrase tuples in transducer machine translation. Thus our algorithm may be a simple way of improving machine translation results.

In this paper we also pointed out some critical issues concerning the evaluation methods. All of them stress the care with which evaluation results must be compared.

## 6. Acknowledgements

This work has been granted by the Spanish Government under grant TIC2002-04447-C02.

## 7. References

Arranz, Victoria, Núria Castell, and Jesús Giménez, 2003. Development of language resources for speech-to-speech translation. In *Proc. of the International Conference on Recent Advances in Natural Language Processing (RANLP)*. Borovets, Bulgary.

Baum, L. E., 1972. An inequality and associated maximization technique in statistical estimation for probabilistic functions of markov processes. *Inequalities*, 3:1–8.

Brown, Peter F., Stephen A. Della Pietra, Vincent J. Della Pietra, and R. L. Mercer, 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

Germann, Ullrich, 2001. Aligned hansards of the 36th parliament of canada. release 2001-1a. http://www.isi.edu/natural-language/download/hansard/index.html.

Lambert, Patrik and Núria Castell, 2004. Evaluation and symmetrisation of alignments obtained with the giza++ software. Technical Report LSI–04–15–R, Technical University of Catalonia. http://www.lsi.upc.es/dept/techreps/techreps.html.

Melamed, I. Dan, 1998. Manual annotation of translational equivalence. Technical Report 98-07, IRCS.

Mihalcea, Rada and Ted Pedersen, 2003. An evaluation exercise for word alignment. In Rada Mihalcea and Ted Pedersen (eds.), *HLT-NAACL 2003 Workshop: Building and Using Parallel Texts: Data Driven Machine Translation and Beyond*. Edmonton, Alberta, Canada: Association for Computational Linguistics.

Och, Franz Josef, 2000. Giza++: Training of statistical translation models. http://www.isi.edu/˜och/GIZA++.html.

Och, Franz Josef and Hermann Ney, 2000. Improved statistical alignment models. In *Proc. of the Annual Meeting of the Association for Computational Linguistics*. Hongkong, China.

Och, Franz Josef and Hermann Ney, 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Vogel, Stephan, Hermann Ney, and Christoph Tillmann, 1996. HMM-based word alignment in statistical translation. In *COLING'96: The 16thInt. Conf. on Computational Linguistics*. Copenhagen, Denmark.

# Sentence Alignment in Parallel, Comparable, and Quasi-comparable Corpora

**Percy Cheung and Pascale Fung**
Human Language Technology Center
Department of Electrical & Electronic Engineering
HKUST, Clear Water Bay
{eepercy,pascale}@ee.ust.hk

## Abstract

We explore the usability of different bilingual corpora for the purpose of multilingual and cross-lingual natural language processing. The usability of bilingual corpus is evaluated by the lexical alignment score calculated for the bi-lexicon pair distributed in the aligned bilingual sentence pairs. We compare and contrast a number of bilingual corpora, ranging from parallel, to comparable, and to non-parallel corpora.

We compare different methods of mining parallel sentences and bilingual lexicon from bilingual corpora. These methods make several sentence-level assumptions on the bilingual corpora. We have found that some of them are applicable to bilingual parallel documents but non-applicable to non-parallel, comparable documents. None of the sentence-level assumptions can be made about non-parallel and quasi-comparable corpora. The latter contain bilingual documents that may or may not be on the same topic.

By postulating additional assumptions on comparable documents, we propose a completely unsupervised method to extract useful material, such as parallel sentences and bilexicons, from quasi-comparable corpora. The lexical alignment score for the comparable sentences extracted with our unsupervised method is found to be very close to that of the parallel corpus. This shows that our extraction method is effective.

## Introduction

There is an explosively increasing amount of new content being loaded to the Internet every day. These online resources constitute practically an unlimited amount of raw material of corpora for natural language processing, such as multilingual information extraction, question answering, machine translation, and so on (Resnik & Smith, 2003)

One of the most challenging tasks in multilingual information extraction is to identify the comparable documents that are more or less within the same topic. This requires the comparison of documents in different languages that are *not* translations of each other.

What is a comparable document? EAGLES Guidelines1 gives a definition of "comparable corpora".

> "*A comparable corpus is one which selects similar texts in more than one language or variety. There is as yet no agreement on the nature of the similarity, because there are very few examples of comparable corpora.*"

The degree of comparability of different documents varies, but we believe that the more comparable the corpora are, it is more useful for various NLP research task.

We can view both parallel and non-parallel corpora as "extreme" cases of comparable corpora. Our objective is to extract parallel sentences from non-parallel, and quasi-comparable corpora.

In this paper, we describe a method for quantifying the comparability of a bilingual corpus. Then we compare different methods for mining parallel sentences and bilingual lexicon, from bilingual corpora with different degrees of comparability. These methods are based on different assumptions about the characteristics of bilingual corpora. We have found that some assumptions for bilingual parallel documents are non-applicable to non-parallel documents. Finally, by postulating additional assumptions on comparable documents, we propose a completely unsupervised method to extract useful material, such as parallel sentences and bilexicons, from a quasi-comparable corpus

## Bilingual Corpora

We compare and contrast bilingual corpora, ranging from the parallel, non-parallel but comparable, and to non-parallel and not very comparable corpora—quasi-comparable corpora.

The Hong Kong Laws Corpus is a parallel corpus with sentence level alignment; it is used as parallel sentence source for statistical machine translation systems. There are 313,659 sentence pairs in Chinese and English. Alignment of parallel sentences from this type of database has been the focus of research for the last decade and can be achieved with many off-the-shelf, publicly available alignment tools.

Previous works have extracted bilingual word senses, lexicon and parallel sentence pairs from noisy parallel corpora. This type of corpora is often called comparable corpora. Corpora like the Hong Kong News Corpus, and the Xinhua News Corpus are in fact rough translations of each other, focused on the same thematic topics, with some insertions and deletions of paragraphs. Sentence and bilingual extraction methods from such corpora can be found in (Fung & McKeown, 1995; Fung & Lo, 1998; Zhao & Vogel, 2002).

On the other hand, TDT3 Corpus is a truly non-parallel and quasi-comparable corpus. It contains transcriptions of various news stories from radio broadcasting or TV news report from 1998-2000 in English and Chinese. In this corpus, there are about 7,500 Chinese and 12,400 English documents, covering 60 different topics. 1,200 Chinese and 4,500 English documents are manually labeled as relevant to a topic and are *in-topic*. The remaining documents are labeled as *off-topic* since they are only weakly relevant to a topic or irrelevant to all topics. The high percentage of off-topic gives rise to more variety of sentences in term of content and structure. From the *in-*

[1] http://www.ilc.cnr.it/EAGLES96/home.html

*topic* documents, most are found to be comparable. A few of the Chinese and English document are almost parallel document and contain some parallel sentences. Nevertheless, the existence of considerable amount of *off-topic* document makes the whole corpus quasi-comparable. The TDT 3 corpus also contains 110,000 Chinese, 290,000 English sentences, giving more than 30 billion possible sentence pairs. A very small portion of the sentence pairs will turn out to be parallel, but many are sentence pairs describing comparable content, with some addition or deletion of minor information or details. The objective of our proposed method is to automatically identify documents that are on the same topic, and then extract parallel sentence pairs from these documents.

## Comparing Bilingual Corpora

We argue that the usability of bilingual corpus is determined by how well the sentences are aligned. We postulate that if the sentence pairs in the corpus are indeed translations of each other, then bilingual word pairs identified in the dictionary will co-occur frequently in this corpus.

Lexical alignment score is defined as the sum of the mutual information score of the bilingual lexicon (bilexicon):

$$S(W_c, W_e) = \frac{f(W_c, W_e)}{f(W_c) f(W_e)}$$

$$S = \sum_{all(W_c, W_e)} S(W_c, W_e)$$

where $f(W_c,W_e)$ is the co-occurrence frequency of bilexicon pair $(W_c,W_e)$ in the aligned sentence pairs. $f(W_c)$, $f(W_e)$ is the occurrence frequency of Chinese word $W_c$ and English word $W_e$, in the respective language sentences set.

We use different alignment methods to extract bilingual parallel sentence pairs from the parallel corpus (Hong Kong Law), a comparable noisy parallel corpus (Hong Kong News), and a non-parallel, quasi-comparable corpus (TDT 3). The lexical alignment scores are computed from the extracted sentence pairs and shown in the following table. We can see that the scores are in direct proportion to the parallel-ness or comparability of the corpus.

| Corpus | Parallel | Comparable | Quasi-Comparable |
|---|---|---|---|
| Bilexicon score | 359.1 | 253.8 | 160.3 |

Table 1. Corpus comparability

In the following section, we describe the different methods we use for extracting bilingual sentence pairs from parallel, comparable, and not-so-comparable corpora.

## Comparing Alignment Methods

All previous work on sentence alignment from parallel corpus makes use of one or multiple of the following assumptions:

1. There are no missing translations in the target document;
2. Sentence lengths: a bilingual sentence pair are similarly long in the two languages;
3. Sentence position: Sentences are assumed to correspond to those roughly at the same position in the other language.
4. Bi-lexical context: A pair of bilingual sentences which contain more words that are translations of each other tend to be translations themselves.

For noisy parallel corpora without sentence delimiters, assumptions for bilingual word pairs are made as follows:

5. Occurrence frequencies of bilingual word pairs are similar
6. The positions of bilingual word pairs are similar
7. Words have one sense per corpus
8. Following 7, words have a single translation per corpus
9. Following 4, the contexts in two languages of a bilingual word pair are similar.

Different sentence alignment algorithms based on both sentence and lexical information can be found in Manning and Schűtze (1999), Wu (2000), and Veronis (2002). These methods have also been applied recently in a sentence alignment shared task at NAACL 2003[2]. We have learned that as bilingual corpora become less parallel, it is better to rely on information about word translations rather than sentence length and position.

For comparable corpora, previous bilingual sentence or word pair extraction work are based soly on bilexical context assumption (Fung & McKeown, 1995; Rapp, 1995; Grefenstette, 1998; Fung & Lo, 1998; Kikui, 1999; Barzilay & Elhadad, 2003; Masao & Hitoshi, 2003; Kenji & Hideki, 2002). Similarly, for quasi-comparable corpora, we cannot rely on any other sentence level or word level statistics but the bi-lexical context assumption.

More recent works on mining parallel sentences from non-parallel comparable corpus are (Munteanu & Marcu, 2002; Zhao & Vogel, 2002). Both work use a translation-model based alignment model trained from parallel corpus and adaptively extract more parallel sentences and bilingual lexicon in the comparable corpus. There are several differences between the two methods. Zhao and Vogel (2002) used a generative statistical machine translation alignment model, while Munteanu and Marcu (2002) used suffix trees. In Zhao and Vogel (2002), the comparable corpus consists of Chinese and English versions of new stories from the Xinhua News agency, while Munteanu and Marcu (2002) used unaligned segments from the French-English Hansard corpus and finds parallel sentences among them.

Existing algorithms (Barzilay & Elhadad, 2003; Masao & Hitoshi, 2003; Kenji & Hideki, 2002), for extracting parallel sentences from comparable documents follow similar steps: firstly extract comparable documents and then extract parallel corpus from comparable documents. They differ in the training and computation of document similarity scores and sentence similarity scores. Examples of document similarity computation include counting word overlap and cosine similarity. Examples of sentence

[2] http://www.cs.unt.edu/~rada/wpt/

similarity computation include word overlap count, cosine similarity, and classification scores of a binary classifier trained from parallel corpora, generative alignment classifier.

We propose a method to find parallel sentences and new word translations from unequal number of sentences in news stories in Chinese and English. In our work, we use simple cosine similarity measures and we dispense with using parallel corpora to train an alignment classifier.

## An Alignment Method for Quasi-comparable Corpora

In addition to the bi-lexical context assumption described in the previous section, we postulate an additional assumption about non-parallel, quasi-comparable corpus:

- Bi-lexicon translation probability: Bilingual lexicon with better translation probabilities can improve bilingual document (sentence) matching.
- Topic: Documents and passages that are on the same topic tend to contain parallel or comparable sentences;
- Seed parallel sentences: Documents and passages that are found to contain *at least* one pair of parallel sentences are likely to contain more parallel sentences.

Based on these assumptions, we propose a first method in extracting useful material from quasi-comparable corpora.

Similar to the iterative process in statistical word alignment methods, we propose that while better document matching leads to better parallel sentence extraction, better sentence matching leads to improved bilingual lexical extraction, the latter in turn improves the document and sentence matches. We propose a multi-level bootstrapping algorithm that iteratively improves the quality of the parallel sentences extracted.

### Multi-level Bootstrapping

**Step 1: Extract Comparable Documents**

The aim of this step is to extract the Chinese-English document pairs that are similar in term distributions.

The documents are word segmented with the Language Data Consortium (LDC) Chinese-English dictionary 2.0. Then the Chinese documents are glossed with the same dictionary. When a Chinese word has multiple possible translations, it is disambiguated with a cohesion scores based method (Gao et al., 2001). Both the glossed Chinese document and English are represented in vector forms, in which the inverse document (where a "document" is a single sentence) frequency is used as the term weight.

Pair-wise similarities are calculated for all possible Chinese-English document pairs, and bilingual documents with similarities above a certain threshold are considered to be comparable. For quasi-comparable corpora, this document alignment step also serves as topic alignment.

**Step 2: Extract Parallel Sentences**

In this step, we extract parallel sentences from the matched English and Chinese documents in the previous section. Each sentence is again represented as word vectors. For each extracted document pair, the pair-wise cosine similarities are calculated for all possible Chinese-English sentence pairs. Sentence pairs above a set threshold are considered parallel and extracted from the documents.

**Step 3: Update the Bilingual Lexicon**

The occurrence of unknown words can adversely affect parallel sentence extraction by introducing erroneous word segmentations. Hence, we need to refine the bi-lexicon by learning new word translations from the intermediate output of parallel sentences extraction. In this work, we focus on learning translations for name entities since these are the words most likely missing in our baseline lexicon. The Chinese name entities are extracted first (Zhai et al., 2004). Translations of these terms are learned from the extracted sentence pairs based on (Fung & Lo, 98) as follows:

**Step 4: Refine Comparable Documents**

This step replaces the original corpus by the set of documents that are found to contain at least one pair of parallel sentences. Other documents that are comparable to this set are also included since we believe that even though they were judged to be not similar at the document level, they might still contain one or two parallel sentences. The algorithm then iterates to refine document extraction and parallel sentence extraction. An alignment score is computed in each iteration, which counts, on average, how many known bilingual word pairs actually co-occur in the extracted "parallel" sentences. The alignment score is high when these sentence pairs are really translations of each other.

### Evaluation

We have evaluated our algorithm on a comparable corpus of TDT3 data. We use our method and a baseline method to extract parallel sentences from this corpus and manually examine the precision of these parallel sentences.

The baseline method shares the same preprocessing, document matching and sentence matching with our proposed method. However, it does not iterate to update the comparable document set, the parallel sentence set, or the bilingual lexicon. . The precision of parallel sentence extract is 43% for the top 2,500 ranked pair. For our approach, the precision of extracted parallel sentences is 67% for the top 2,500 ranked pair, which is 24% higher. In addition, we also found that the precision of parallel sentence pair extraction increases steadily over each iteration in our method, until convergence.

The main contribution of the unsupervised multi-level bootstrapping is in steps 3 and 4 and in the iterative process. The iterative lexicon-sentence alignment process has been previously applied to alignment tasks from parallel corpus. By using the correct alignment assumptions, we have demonstrated that a bootstrapping iterative process is also possible for finding parallel sentences and new word translations from comparable corpus.

## Conclusion

We explore the usability of different bilingual corpora for the purpose of multilingual natural language processing. We compare and contrast a number of bilingual corpora, ranging from the parallel, to comparable, and to non-parallel corpora. A lexical alignment score calculated for the bi-lexicon pair distributed in the aligned bilingual sentence pairs then evaluates the usability of each type of corpus.

We compared different alignment assumptions for mining parallel sentences from these different types of bilingual corpora and proposed new assumptions for quasi-comparable corpora.

By postulating additional assumptions on seed parallel sentences of comparable documents, we propose a multi-level bootstrapping algorithm to extract useful material, such as parallel sentences and bilexicons, from *quasi-comparable corpora*. This is a completely unsupervised method. Evaluation results show that our approach achieves 67% accuracy and a 23% improvement from baseline. This shows that the proposed assumptions and algorithm are promising for our objective. The lexical alignment score for the comparable sentences extracted with our unsupervised method is found to be very close to that of the parallel corpus. This shows that our extraction method is effective.

## References

Regina Barzilay and Noemie Elhadad, (2003). "Sentence Alignment for Monolingual Comparable Corpora", Proc. of EMNLP, 2003, Sapporo, Japan.

Pascale Fung and Kathleen Mckeown. (1997). Finding terminology translations from non-parallel corpora. In The 5th Annual Workshop on Very Large Corpora. Pages 192--202, Hong Kong, Aug. 1997."

Pascale Fung and Lo Yuen Yee. (1998). "An IR Approach for Translating New Words from Nonparallel, Comparable Texts". In Coling 1998

Jianfeng Gao, Jian-Yun Nie, Endong Xun, Jian Zhang, Ming Zhou, Changning Huang. (2001). "Improving Query Translation for Cross-Language Information Retrieval using Statistical Models. In SIGIR'01 September 9-12,2001, New Orleans, Louisiana, USA.

Gregory Grefenstette, editor. (1998). "Cross-Language Information Retrieval". Kluwer Academic Publishers, 1998.

Hiroyuki Kaji. (2003). Word sense acquisition from bilingual comparable corpora, in Proceedings of the NAACL, 2003, Edmonton, Canada, pp 111-118.

Genichiro Kikui. (1999). Resolving translation ambiguity using non-parallel bilingual corpora. In Proceedings of ACL99 Workshop on Unsupervised Learning in Natural Language

Christopher D. Manning and Hinrich Schűtze. (1999). Foundations of Statistical Natural Language Processing. The MIT Press.

Kenji Matsumoto and Hideki Tanaka. (2002) Automatic alignment of Japanese and English Newspaper articles using an MT system and a bilingual Company name dictionary. In LREC-2002, pages 480-484

Dragos Stefan Munteanu, Daniel Marcu. (2002). Processing Comparable Corpora With Bilingual Suffix Trees. In Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002).

Reinhard Rapp. (1995). Identifying word translations in non-parallel texts. Proceedings of the 33rd Meeting of the Association for Computational Linguistics. Cambridge, MA, 1995. 320-322

Philip Resnik and Noah A. Smith. (2003) " The Web as a Parallel Corpus", Computational Linguistics 29(3), pp. 349-380, September 2003.

Masao Utiyama and Hitoshi Isahara. (2003). Reliable measures for aligning Japanese-English news articles and sentences. In Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics. Sapporo, Japan.

Jean Veronis (editor). (2000). Parallel Text Processing: Alignment and Use of Translation Corpora. Dordrecht: Kluwer. ISBN 0-7923-6546-1. Aug 2000.

Dekai Wu. (2000). Alignment. In Robert Dale, Hermann Moisl, and Harold Somers (editors), Handbook of Natural Language Processing. 415-458. New York: Marcel Dekker. ISBN 0-8247-9000-6. Jul 2000.

Bing Zhao, Stephan Vogel. (2002). Processing Comparable Corpora With Bilingual Suffix Trees, In Proceedings of the ICSLP 2002.

Zhai, Lufeng, Pascale Fung, Richard Schwartz, Marine Carpuat and Dekai Wu. (2004). Using N-best list for Named Entity Recognition from Chinese Speec. In the Proceedings of the NAACL 2004 , to appear

# Propbanking in Parallel

**Paul Kingsbury, Nianwen Xue, Martha Palmer**
Department of Computer and Information Science
University of Pennsylvania
{kingsbur, xueniwen, mpalmer}@unagi.cis.upenn.edu

**Abstract**

This paper describes an effort to provide semantic role annotation for parallel Chinese/English corpora that we believe has the potential of benefiting statistical machine translation. This level of annotation, called a Parallel Proposition Bank, abstracts away from divergences in word order and syntactic categories to facilitate a mapping from a clausal structure in one language to the corresponding clausal structure in the other language. It collects together split arguments, making it easier to find their foreign language counterparts. It also provides for a level of coarse-grained word sense disambiguation based primarily on differences in subcategorization frames that could simplify the task of lexical choice. Although there are still many language specific characteristics of the semantic annotation, it moves us one step closer to a general semantic representation that is language independent.

## Introduction

Concurrent with the completion of the PropBank project at Penn (Palmer et al (submitted), Kingsbury and Palmer 2002), the decision was made to extend the annotation methodology both to independent corpora in other languages and to multilingual parallel corpora. The intention is first to gain resources for shallow semantic analysis in languages other than English; thus, monolingual PropBanking efforts have begun for Chinese (Xue and Palmer 2003), Korean and one is planned for Arabic. A second, more salient goal is the facilitation of machine translation systems. Just as there is evidence that syntactic parses improve the accuracy of MT systems (Yamada and Knight 2001, Charniak etal. 2003), it is expected that semantic parses will also improve accuracy, by showing explicit dependency relationships between elements of a sentence. PropBanking further includes a degree of coarse-grained sense-tagging which could also facilitate accurate translations.

PropBanking in parallel requires a number of resources. A first obvious step is the collection or creation of a parallel corpus annotated with syntactic structures. The Penn Chinese Treebank comprises almost 250K words of Xinhua news and 250K words of Sinorama (a Taiwanese multilingual news magazine) (Xue et al. 2004). There is on-going effort at the University of Pennsylvania to treebank the English translation of the first 100 thousand words of the treebanked Xinhua news, as well as the corresponding 250K word English Sinorama corpus. More important is the pre-existence of argument-structure lexicons for each of the languages in question. More than 3300 lexical items of English already have entries in the Propbank frame lexicon, and there are more than 4500 Chinese PropBank entries. A third component of the parallel propbanking endeavor is to explore the transferability between the languages at the level of frameset. It is hoped that this transferability can be exploited in future Machine Translation systems.

## The Propbank

### Generalities and the English Propbank

PropBank is a shallow semantic parse of running text, marking the argument structure of the verbs and deverbal adjectives. It comprises two separate but interdependent parts. The first is an annotated corpus wherein every verb and its arguments are explicitly marked. The corpus in question for English is the Wall Street Journal portions of the Penn TreeBank II (Marcus et al, 1994), while for Chinese the corpus is the Chinese TreeBank (Xue et al, 2004). Of more interest is the second part of the PropBank resource, the so-called 'frames files.' These are collectively a lexicon detailing the specific arguments expected to appear with any given verb. Arguments are assigned a (relatively) theory-neutral numbered label and are assigned a verb-specific mnemonic label. Different senses of a verb are assigned to different 'framesets' containing independent definitions of arguments. Senses are defined on both semantic and syntactic grounds. For example, the English verb 'afford' is seen in contexts such as the following:

1. These days Nissan can afford that strategy, even
   though profits aren't exactly robust. (wsj_0286)
2. Last year the public was afforded a preview of Ms.
   Bartlett's creation in a tablemodel version, at a BPC
   exhibition. (wsj_0984)

Although each example shows two arguments, the passive morphology on the second sentence shows that a third argument must be possible, providing a syntactic motivation for the framing of 'afford' as follows:

afford.01 'be able to sustain the cost of something'
arg0: entity sustaining cost
arg1: costly thing

afford.02 'provide, make available'
arg0: provider
arg1: thing provided
arg2: recipient

Framesets are also distinguished when the meanings of the usages are sufficiently different, even if the number of roles is the same. For example, the verb 'stem' also takes two framesets[1], each with two roles, on the basis of sentences such as the following:

3. Travelers Corp.'s third-quarter net income rose 11%, even though claims stemming from Hurricane Hugo reduced results $40 million. (wsj_0144)
4. If the company can start to ship during this quarter, it could stem some, if not all of the red ink, he said. (wsj_1973)

Under most circumstances a relatively proficient speaker of English will be able to distinguish between these senses, motivating their classification into separate framesets.

stem.01 'arise'
arg1: entity arising, coming about
arg2: arising from what?

stem.02 'stanch, cause to stop flowing'
arg0: causer of non-flowing
arg1: thing no longer flowing

Verb senses are thus defined at a level considerably more coarse-grained than the senses used in WordNet (Palmer, et. al., 2004), but the disambiguation still results in an explosion of related verbs. The English TreeBank contains approximately 3300 separate lexical items identified as verbs, but even the coarse-grained distinctions produce more than 4600 framesets.

### Special Issues for the Chinese Propbank

The same annotation philosophy has been extended to the Penn Chinese Proposition Bank (Xue and Palmer, 2003). In Chinese, the same syntactic alternations that form the basis for the English PropBank annotation also exist in robust quantities, even though it may not be the case that the same exact verbs (meaning verbs that are close translations of one anther) have the exact same range of syntactic realization for Chinese and English. For example, in (5), "xin-nian/New Year zhao-dai-hui/reception" plays the same role in (a) and (b), even though it occurs in different syntactic positions. This regularity is captured by assigning the same argument label ARG1 to both instances. It is worth noting that the predicate "ju-xing/hold" does not have passive morphology in (5a), despite what its English translation suggests. Like the English Propbank, the adjunct-like elements receive more general labels like TMP or LOC. The tag set for Chinese and English PropBanks are to a large extent similar and more details can be found in (Xue and Palmer, 2003).

5. a. [$_{\text{ARG1}}$ *xin-nian*/New Year *zhao-dai-hui*/reception] [$_{\text{ARGM-TMP}}$ *jin-tian*/today][$_{\text{ARGM-LOC}}$ *zai*/at *diao-yu-tai*/Diaoyutai *guo-bin-guan*/state guest house] *ju-xing*/hold
"A New Year reception was held in Diaoyutai State Guest House today."

b. [$_{\text{ARG0}}$ *tang-jia-xuan*/Tang Jiaxuan] [$_{\text{ARGM-TMP}}$ *jin-tian*/today] [$_{\text{ARGM-LOC}}$ *zai*/at *diao-yu-tai*/Diaoyutai *guo-bin-guan*/state guest house] *ju-xing*/hold [$_{\text{ARG1}}$ *xin-nian*/New Year *zhao-dai-hui*/reception]
"Tang Jiaxuan was holding the New Year Reception in Diaoyutai State Guest House today."

For polysemous verbs we also distinguish different framesets. (6) and (7) illustrate the different framesets of "tong-guo/pass", which correspond with major senses of the verb, loosely defined. The frameset in (6) roughly means "pass by voting" while the frameset illustrated by (7) means "pass through".

6. a. [$_{\text{ARG0}}$ *mei-guo*/the U.S. *guo-hui*/Congress] *zui-jin*/recently *tong-guo*/pass *le*/ASP [$_{\text{ARG1}}$ *zhou-ji*/interstate *yin-hang-fa*/banking law]
"The U.S. Congress recently passed the inter-state banking law."
b. [$_{\text{ARG1}}$ *zhou-ji*/interstate *yin-hang-fa*/banking law] *zui-jin*/recently *tong-guo*/pass *le*/ASP
"The inter-state banking law passed recently."

7. a. [$_{\text{ARG0}}$ *huo-che*/train] *zhen-zai*/now *tong-guo*/pass [$_{\text{ARG1}}$ *sui-dao*/tunnel]
"The train is passing through the tunnel."
b. [$_{\text{ARG0}}$ *huo-che*/train] *zheng-zai*/now *gong-guo*/pass.
"The train is passing."

Despite these similarities between the languages, there are also some Chinese-specific issues that have to be dealt with in the process of creating frame files. One issue is the disambiguation of preverbal prepositional phrases. As illustrated in (8), these preverbal PPs can be dependent on the verb, as in (8a), or the postverbal NP as in (8b). In English, since all such PPs are postverbal, this disambiguation can be done straightforwardly in syntax by attaching them at different levels. Such a simple solution does not exist for Chinese. Instead, this is handled as part of the PropBanking effort by marking verb-dependent PPs, such as that of (8a), as a semantic argument of the verb. The noun-dependent PP in (8b) will be related to the post-verbal NP and will have no predicate-argument label relative to the verb.

8. a. *zai*/at *jiu-hui*/banquet *shang*/on *cai*/Cai *da-shi*/ambassador **[**$_{\text{PP}}$ *dui*/to *yi-xiang*/always *guan-xin*/support *zu-guo*/motherland *jian-she*/development *de*/DE *hai-wai*/overseas *tong-bao*/compatriot**]** [$_{\text{V}}$ *fa-biao*/deliver] *le*/ASP [$_{\text{NP}}$ *re-qing*/enthusiam *yang-yi*/overflow *de*/DE *jiang-hua*/speech].
"At the banquet, Ambassador Cai made an enthusiastic speech to the overseas compatriots."

b. *zeng-yin-quan*/Zeng Yinquan **[**$_{\text{PP}}$ *jiy*/on *jian-li*/establish *guo-ji*/international *jin-rong*/financial *xin*/new *zhi-xu*/order**]** [$_{\text{V}}$ *fa-biao*/express] [$_{\text{NP}}$ *jian-jie*/view] .

[1] This ignores two other possible senses of 'stem' which do not happen to occur in the corpus, namely 'reduce to just a stem' as in a morphological stemmer and 'remove the stems of something which inherently has a stem' as in stemmed cherries.

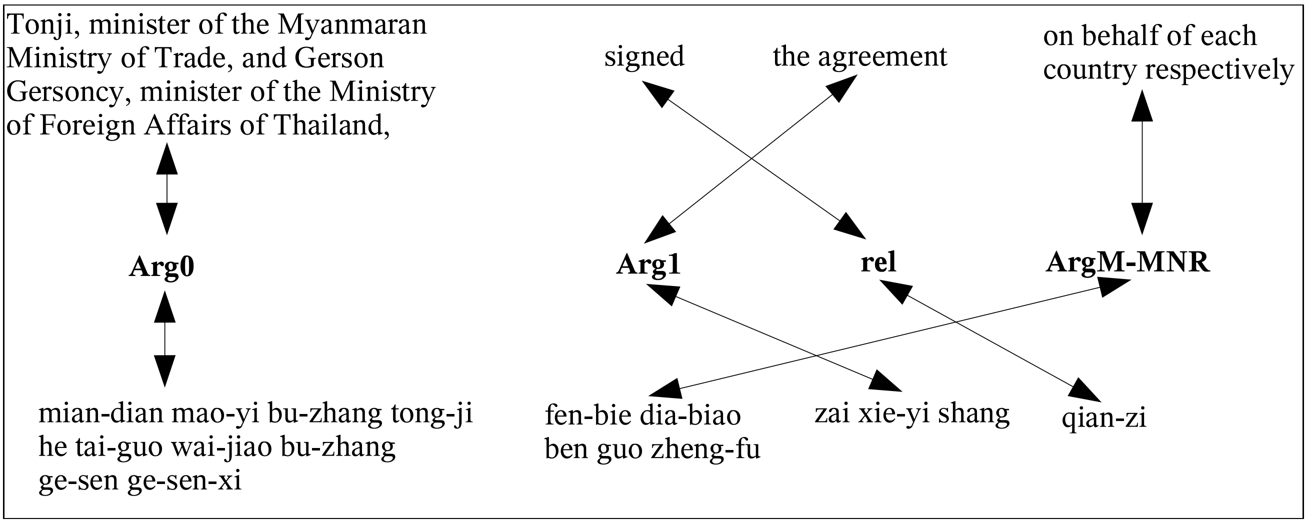

Tonji, minister of the Myanmaran Ministry of Trade, and Gerson Gersoncy, minister of the Ministry of Foreign Affairs of Thailand,
signed
the agreement
on behalf of each country respectively
Arg0
Arg1
rel
ArgM-MNR
mian-dian mao-yi bu-zhang tong-ji he tai-guo wai-jiao bu-zhang ge-sen ge-sen-xi
fen-bie dia-biao ben guo zheng-fu
zai xie-yi shang
qian-zi

Figure 1: Mapping between Chinese and English arguments

“Zeng Yinquan expressed his own view on the establishment of a new international financial order."

Another phenomenon which is much more common in Chinese than in English is split arguments. One such split is between the possessor (PSR) and the possessee (PSE). Here the possessor and possessee are abstract notions and do not necessarily indicate a strict possession relation. This is illustrated in (9).

9. [ARG1-psr *zhong-guo*/China *jing-ji*/economy *zeng-zhang*/growth] *ye*/also *jiang*/will [V *fang-man*/slow down] [ARG1-pse *su-du*/speed]
“The speed of Chinese economic growth will also slow down.”

## Transferability of Framesets

The value of the PropBanking effort lies in the fact that the semantic representations implemented in the frame files of the two languages abstract away from the syntactic idiosyncrasies of the individual languages and create a platform where the predicate-argument structure mapping can take place. If these mappings can be recovered automatically, then it will have a profound impact on machine translation. Although the extent to which such mapping can be performed in a straightforward manner is yet to be determined, a preliminary examination shows that the PropBank annotations would facilitate such a mapping in a number of ways. First, the PropBank representation abstracts away from divergences in the word order and the syntactic category of the two languages and allows for a straightforward mapping at the predicate-argument structure level. This is illustrated in (10) and graphically in Figure 1.

10. [ARG0 Tonji, minister of the Myanmaran Ministry of Trade, and Gerson Gersoncy, minister of the Ministry of Foreign Affairs of Thailand], [FRAMESET.01 signed] [ARG1 the agreement] on behalf of each country respectively.
[ARG0 *mian-dian*/Myanmar *mao-yi*/trade *bu-zhang*/minister *tong-ji*/Tonji *he*/and *tai-guo*/Thailand *wai-jiao* *bu-zhang*/foreign minister *ge-sen ge-sen-xi*/Gerson Gersoncy] *fen-bie*/respectively *dai-biao*/represent *ben*/own *guo*/country *zheng-fu*/government [ARG1 *zai*/at *xie-yi*/agreement *shang*/above] [FRAMESET.01 *qian-zi*/sign].

Second, the PropBank annotation also abstracts away from the split argument phenomenon in the two languages. Split arguments may occur in different places and with different predicates in the two languages, but the PropBank annotation addresses this by marking the pieces as belonging to the same argument. This is illustrated in (11), adapted from (9):

11. [ARG1-psr *zhong-guo*/China *jing-ji*/economy *zeng-zhang*/growth] *ye*/also *jiang*/will [V *fang-man*/slow down] [ARG1-pse *su-du*/speed]
[ARG1-pse The speed] [ARG1-pse of Chinese economic growth] will also slow down.

Having the frameset information also enables us to map framesets that have compatible argument structures across languages. In many cases the framesets of a verb in one language map to different lexical items in another. For example, “leave” has two framesets and each takes a different set of arguments. They are mapped to different lexical items in Chinese:

leave.01: *li-kai*
Arg0: entity leaving
Arg1: place left
Arg2: attribute of Arg1

12. This flight leaves Shanghai at midnight.
*hang-ban*/flight *wu-ye*/midnight *li-kai*/leave *shang-hai*/Shanghai

leave.02: *liu-gei*
Arg0: giver
Arg1: thing given
Arg2: benefactor

13. John left Mary a big fortune.
*yue-han*/John *liu-gei*/leave *ma-li*/Mary *yi*/one *da-bi*/big sum *cai-chan*/fortune

## Conclusion

This paper has described the basis of the PropBank annotation that is being applied to parallel Chinese/English corpora. The English and Chinese PropBanks provide a level of annotation that highlights the dependency structure of a clause and the semantic roles played by the dependents. It abstracts away from surface idiosyncrasies such as word order, syntactic category and split constituents. The expectation is that this level of annotation, in addition to aiding the development of increasingly sophisticated monolingual information processing tools, will also prove useful to various kinds of machine translation systems. Transfer-based machine translation approaches could benefit from corpus-based transfer lexicons extracted from PropBanked parallel corpora. Statistical machine translation systems could re-rank potential target language outputs based on the similarity between their semantic role labels and those of the source language sentence. Although still preserving many language-specific characteristics, this level of annotation is one step closer to a general-purpose semantic representation.

## Acknowledgments

Work on the English PropBank has been funded by DARPA grant N66001-00-1-8915 and DOD grant MDA904-00-C-2136. The Chinese PropBank has been funded by DOD grant MDA904-02C0412. The Parallel PropBanks are being funded by NSF ITR Supplement 0343005.

## References

Charniak, E., Knight, K. & Yamada K. (2003) Syntax-based Language Models for Machine Translation. *Proceedings of MT Summit IX 2003*. New Orleans.

Kingsbury, P. & Palmer, M. (2002) From Treebank to Propbank. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation* (LREC-2002), Las Palmas, Spain.

Marcus, M., Kim, G., Marcinkiewicz, M.A., MacIntyre, R., Bies, A., Ferguson, M., Katz, K., & Schasberger, B. (1994) The Penn Treebank: Annotating predicate argument structure. In *ARPA Human Language Technology Workshop*, pp 114-119, Plainsboro NJ.

Ng, H.T., Wang, B., & Chan, Y.S. (2003). Exploiting Parallel Texts for Word Sense Disambiguation: An Empirical Study. In the *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics* (ACL-03). Sapporo, Japan.

Palmer, M., Gildea, D. & Kingsbury, P. (submitted) The Proposition Bank: An Annotated Corpus of Semantic Roles, submitted to *Computational Linguistics.*

Palmer, M., Babko-Malaya, M., Dang, H., Different Sense Granularities for Different Applications, *$2^{nd}$ Workshop on Scalable Natural Language Understanding Systems, at HLT/NAACL-04,* Boston, Mass, May 6, 2004.

Xue, N. & Palmer, M. (2003) Annotating Propositions in the Penn Chinese Treebank. In *Proceedings of the Second Sighan Workshop,* in conjunction with ACL'03, Sapporo, Japan.

Xue, N., Xia, F., Chiou, F. & Palmer, M. 2004. The Penn Chinese Treebank: Phrase Structure Annotation of a Large Corpus, *Natural Language Engineering,* 10(4):1-30.

Yamada, K. & Knight, K. 2001. A Syntax-based Statistical Translation Model. *Proceedings of the Conference of the Association for Computational Linguistics,* (ACL-2001).

# Browsing Multilingual Information with the MultiSemCor Web Interface

**Marcello Ranieri, Emanuele Pianta, Luisa Bentivogli**

ITC-irst
Via Sommarive 18, 38050 Povo (Trento) - Italy
{ranieri,pianta,bentivo}@itc.it

**Abstract**

Parallel and comparable corpora represent a crucial resource for different Natural Language Processing tasks like machine translation, lexical acquisition, and knowledge structuring but are also suitable to be consulted by humans for different purposes, such as linguistic teaching, corpus linguistics, translation studies, lexicography, multilingual information browsing. To enhance their exploitation by human users, specially designed interfaces need to be developed. In this paper we present the design and implementation of the MultiSemCor Web Interface. MultiSemCor is a parallel English/Italian corpus, which is being developed at ITC-irst starting from the English corpus SemCor. In MultiSemCor the texts are aligned at word level and semantically annotated with WordNet senses. The MultiSemCor Web Interface allows the users to exploit at best the potentiality of the corpus. We will describe the main functions of the interface, which provides two distinct browsing modalities: a bi-text-oriented modality and a word-oriented modality, which amounts to a bilingual semantic concordancer. Moreover, the MultiSemCor Web Interface is integrated with the on-line MultiWordNet browser, which gives access to the reference lexicon for MultSemCor.

## 1 Introduction

In the last years, the importance of parallel and comparable corpora has become more and more evident within the human language technology field, where these resources are used for the extraction of multilingual information in many tasks such as machine and machine-aided translation, linguistic teaching, lexicography, and knowledge structuring.

To enhance the exploitation of parallel corpora by humans, suitable interfaces need to be developed. Such interfaces should give access to all the information available in the corpus in an easy and intuitive way, and should possibly be integrated with other linguistic resources such as on-line dictionaries.

In this paper we will focus on the design and implementation of the MultiSemCor Web interface. MultiSemCor (Bentivogli & Pianta, 2002) is a parallel English-Italian corpus, aligned at word level and annotated with PoS, lemma and word sense. It has been obtained starting from SemCor, an English corpus semantically tagged with WordNet senses.

The rest of the paper is organized as follows. In Section 2 we summarise the methodology developed for the creation of the MultiSemCor corpus and its composition up to now. In Section 3 we describe in detail the MultiSemCor Web interface, its main browsing functionalities and novel characteristics. In Section 4 we outline some existing related work before concluding in Section 5.

## 2 The MultiSemCor Corpus

MultiSemCor is a parallel English-Italian corpus, which is being developed at ITC-irst starting from SemCor, a subset of the English Brown corpus containing almost 700,000 running words. In SemCor all the words are tagged by PoS, and more than 200,000 content words are also lemmatized and sense-tagged according to WordNet (Fellbaum, 1998). The strategy for creating MultiSemCor consists in having SemCor texts translated into Italian by professional translators; aligning Italian and English texts at word level; and then transferring the word sense annotations from English to the aligned Italian words. Both the word alignment and the annotation transfer are carried out automatically.

The main hypothesis underlying this methodology is that, given a text and its translation into another language, the translation preserves to a large extent the meaning of the source language text. A pilot study estimated that this methodology can be applied with a precision of 95% and a recall of 75%. The automatic projection of annotations from one language to another has been adopted as a strategy aiming at reducing the effort needed for obtaining annotated corpora (Pianta & Bentivogli, 2003): the result is an Italian corpus annotated with PoS, lemma and word sense, but also an aligned parallel corpus lexically annotated with a shared inventory of word senses. More specifically, the sense inventory used is MultiWordNet (Pianta et al., 2002), a multilingual lexical database in which the Italian component is strictly aligned with the English Princeton Wordnet.

At present MultiSemCor is composed of 116 English texts aligned at sentence level with their corresponding 116 Italian translations. The total amount of running words is 230,738 for English and 233,178 for Italian. The word alignment and transfer methodology has been applied to 29 texts out of the 116 texts available. These 29 texts are aligned at word level and annotated with PoS, lemma, and word sense. As regards English, we have 55,935 running words and 29,655 words semantically annotated (from SemCor). As for Italian, the corpus is composed of 59,726 running words among which 23,095 words are annotated with word senses that have been automatically transferred from English.

MultiSemCor will be useful for a variety of tasks. From a computational point of view we are planning to use it to automatically enrich the Italian component of MultiWordNet. As a matter of fact, out of the 23,095 Italian words automatically sense-tagged, 5,292 are not yet present in MultiWordNet and will be added to it. Moreover, MultiSemCor is also suitable to be consulted by humans for different purposes, such as language teaching and learning, translation studies, lexicography, multilingual information browsing.

## 3 The Interface

To help human users exploiting at best the potentiality of MultiSemCor, a Web-based browser has been realized. In its design we faced a number of interesting issues, such as making available to the users information about corpus annotation, bilingual text alignment, bilingual semantic concordancing, integration between corpora and lexical resources. To meet all these requirements, two distinct browsing modalities have been implemented. The first is *text-oriented* and the second is *word-oriented*. Each of these two modalities is embodied in a dynamic Web page.

### 3.1 Bi-text Browsing

In the text-oriented browsing modality, for each bi-text the user can access the following information:

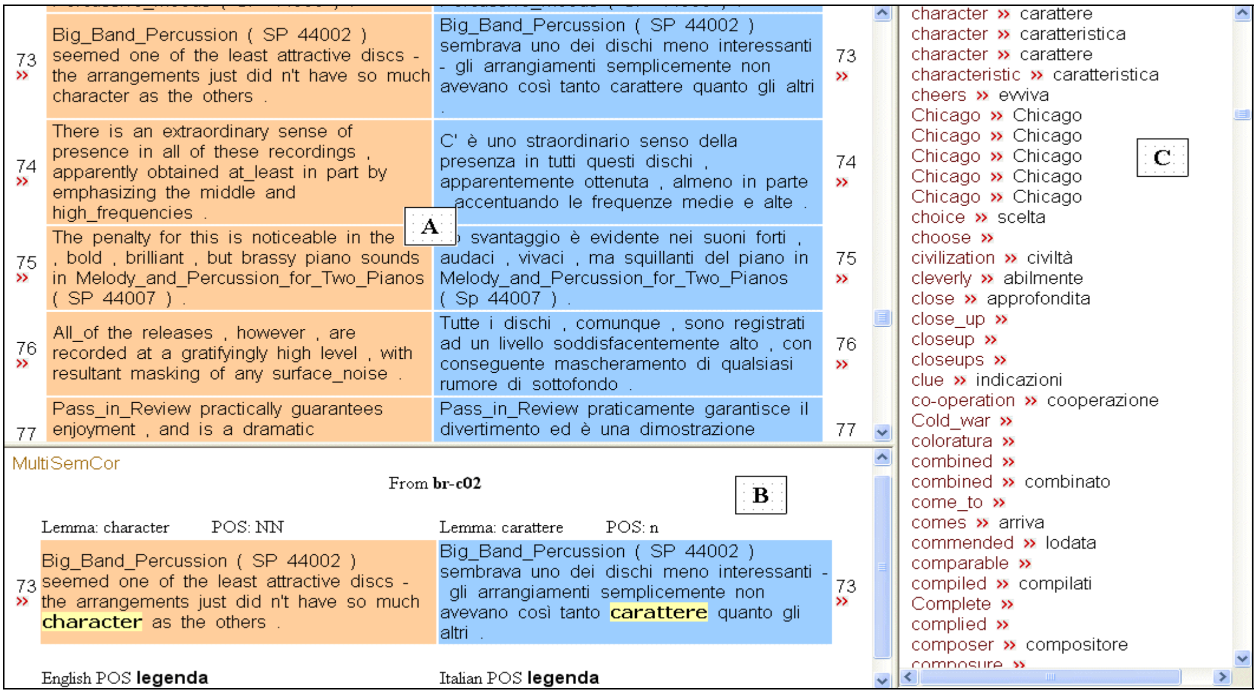
A. alignment at sentence level
B. alignment at word level
C. dictionary of all the tokens of the text, with links to the sentences in which they occur

These functionalities have been implemented through a web-page organized in three sections corresponding to the three kinds of information above, see Figure 1. Section A contains the whole bi-text and shows the alignment at sentence level. This has been realized through a simple two column table, where each column contains the text in one of the two languages, and each row shows the alignment between a sentence and its translation. This solution shows the alignment between sentences, while keeping the possibility for the user to read the entire two texts in a natural way.

Section B allows the user to focus on a specific sentence and shows the available alignments at word level for that sentence. Showing word level alignments through a Web interface, while keeping the readability of the sentence in which the words occur is not as straightforward as showing sentence level alignments. Alignments could be shown for instance by marking aligned words with various colours, a colour for each alignment, or by putting the two sentences in a two column table, where each row contains a word alignment. However, we think that the former solution may be visually awkward, and for long sentences it makes the correspondence between words hard to trace. The latter solution makes the correspondence between words easier to read, but makes the entire sentence difficult or impossible to read, because of the vertical layout of words, and because the order of words in the target sentence needs to be completely changed. To solve the problem we choose to show only one word alignment per time, by highlighting the aligned words in the source and target sentence. Note that along with the word alignment, Section B also provides the available morphosyntactic information about the aligned words.

Section C of the interface contains a list of all the tokens in the current text in alphabetic order, with the translation in the other language. In fact there are two such lists, one for English-to-Italian, and one for Italian-to-English correspondences. Each token is hyperlinked with the sentence in which the token occurs.

In the example in Figure 1, the user is browsing the text br-c02 in Section A of the interface. By clicking on the word *character* contained in sentence nr. 73, he/she gets two results. Section B highlights the alignment between the word *character* in the English sentence, and *carattere* in the Italian translation. On the other hand, the top of Section C shows all the translations of the word *character* in the current text. Note that the user can now ask for the interface to show the passages in which the other translations of *character* are to be found.



| | | | |
|---|---|---|---|
| 73 » | Big_Band_Percussion ( SP 44002 ) seemed one of the least attractive discs - the arrangements just did n't have so much character as the others . | Big_Band_Percussion ( SP 44002 ) sembrava uno dei dischi meno interessanti - gli arrangiamenti semplicemente non avevano così tanto carattere quanto gli altri . | 73 » |
| 74 » | There is an extraordinary sense of presence in all of these recordings , apparently obtained at_least in part by emphasizing the middle and high_frequencies . | C' è uno straordinario senso della presenza in tutti questi dischi , apparentemente ottenuta , almeno in parte accentuando le frequenze medie e alte . | 74 » |
| 75 » | The penalty for this is noticeable in the , bold , brilliant , but brassy piano sounds in Melody_and_Percussion_for_Two_Pianos ( SP 44007 ) . | svantaggio è evidente nei suoni forti , audaci , vivaci , ma squillanti del piano in Melody_and_Percussion_for_Two_Pianos ( Sp 44007 ) . | 75 » |
| 76 » | All_of the releases , however , are recorded at a gratifyingly high level , with resultant masking of any surface_noise . | Tutte i dischi , comunque , sono registrati ad un livello soddisfacentemente alto , con conseguente mascheramento di qualsiasi rumore di sottofondo . | 76 » |
| 77 | Pass_in_Review practically guarantees enjoyment , and is a dramatic | Pass_in_Review praticamente garantisce il divertimento ed è una dimostrazione | 77 |

MultiSemCor

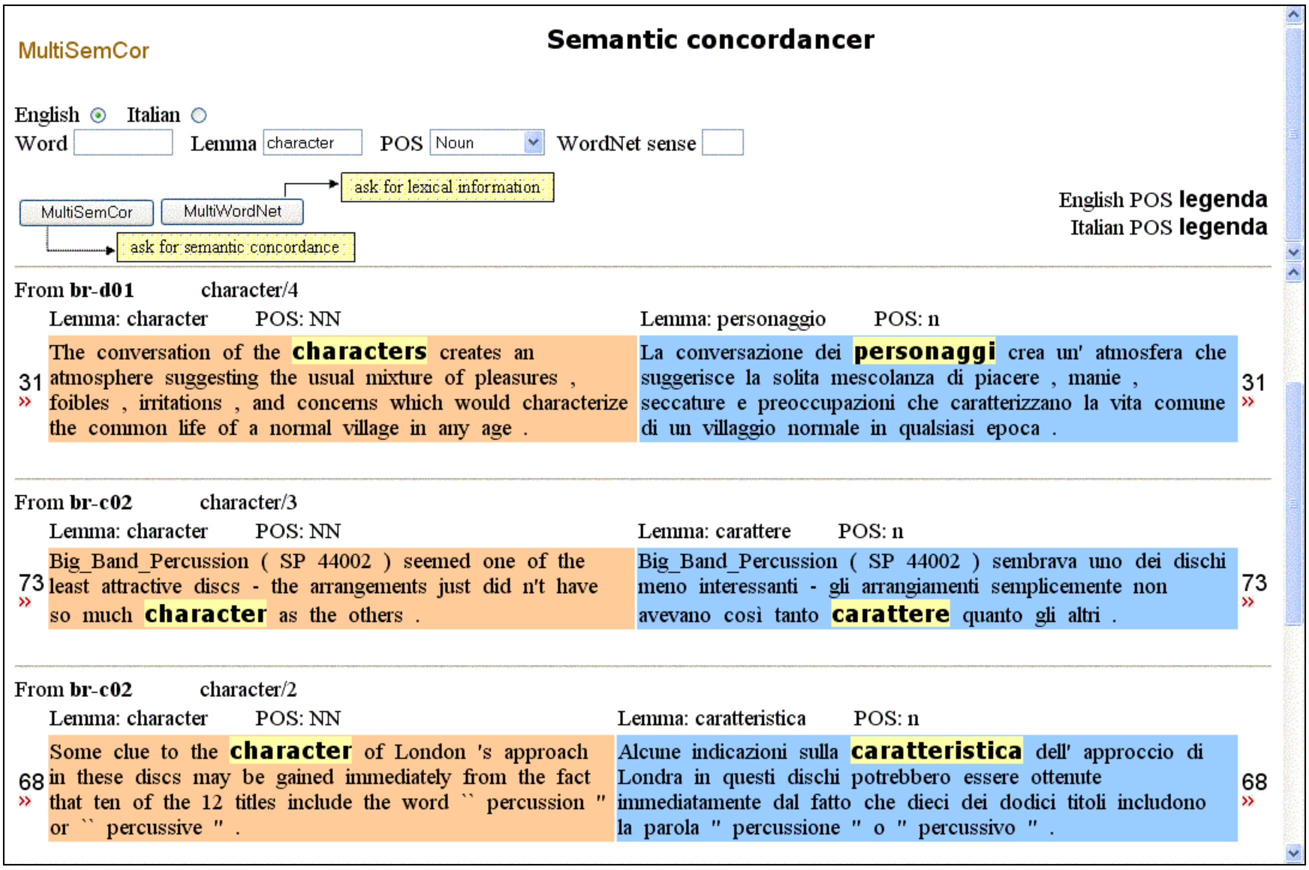
From **br-c02**

| | | | |
|---|---|---|---|
| | Lemma: character POS: NN | Lemma: carattere POS: n | |
| 73 » | Big_Band_Percussion ( SP 44002 ) seemed one of the least attractive discs - the arrangements just did n't have so much **character** as the others . | Big_Band_Percussion ( SP 44002 ) sembrava uno dei dischi meno interessanti - gli arrangiamenti semplicemente non avevano così tanto **carattere** quanto gli altri . | 73 » |

English POS **legenda** Italian POS **legenda**

| | | |
|---|---|---|
| character | » | carattere |
| character | » | caratteristica |
| character | » | carattere |
| characteristic | » | caratteristica |
| cheers | » | evviva |
| Chicago | » | Chicago |
| Chicago | » | Chicago |
| Chicago | » | Chicago |
| Chicago | » | Chicago |
| Chicago | » | Chicago |
| choice | » | scelta |
| choose | » | |
| civilization | » | civiltà |
| cleverly | » | abilmente |
| close | » | approfondita |
| close_up | » | |
| closeup | » | |
| closeups | » | |
| clue | » | indicazioni |
| co-operation | » | cooperazione |
| Cold_war | » | |
| coloratura | » | |
| combined | » | |
| combined | » | combinato |
| come_to | » | |
| comes | » | arriva |
| commended | » | lodata |
| comparable | » | |
| compiled | » | compilati |
| Complete | » | |
| complied | » | |
| composer | » | compositore |
| composure | » | |

Figure 1: the browser in the *text-oriented* modality



English ● Italian ○
Word [ ] Lemma [character] POS [Noun] WordNet sense [ ]

From **br-d01** character/4

Lemma: character POS: NN

31 » The conversation of the **characters** creates an atmosphere suggesting the usual mixture of pleasures , foibles , irritations , and concerns which would characterize the common life of a normal village in any age .

Lemma: personaggio POS: n

La conversazione dei **personaggi** crea un' atmosfera che suggerisce la solita mescolanza di piacere , manie , seccature e preoccupazioni che caratterizzano la vita comune di un villaggio normale in qualsiasi epoca . 31 »

From **br-c02** character/3

Lemma: character POS: NN

73 » Big_Band_Percussion ( SP 44002 ) seemed one of the least attractive discs - the arrangements just did n't have so much **character** as the others .

Lemma: carattere POS: n

Big_Band_Percussion ( SP 44002 ) sembrava uno dei dischi meno interessanti - gli arrangiamenti semplicemente non avevano così tanto **carattere** quanto gli altri . 73 »

From **br-c02** character/2

Lemma: character POS: NN

68 » Some clue to the **character** of London 's approach in these discs may be gained immediately from the fact that ten of the 12 titles include the word `` percussion '' or `` percussive '' .

Lemma: caratteristica POS: n

Alcune indicazioni sulla **caratteristica** dell' approccio di Londra in questi dischi potrebbero essere ottenute immediatamente dal fatto che dieci dei dodici titoli includono la parola '' percussione '' o '' percussivo '' . 68 »

Figure 2: the result of a query in the semantic concordancer

## 3.2 Semantic Concordancer

The second modality for browsing the corpus is word-oriented, and amounts to a bilingual semantic concordancer, that is a tool able to provide all the occurrences of a certain word sense in a corpus. More precisely, in the MultiWordNet concordancer the user can alternatively search for all the occurrences of a *word form*, *lemma*, or *word sense* (according to MultiWordNet). The user can also constrain the search to a certain PoS. Free combinations between all these constraints (language, word form, lemma, word sense, PoS) are allowed. For instance the user can search for all the occurrences of: the word form *characters*; or the word form *character* as verb; or the lemma *character* in all of its senses; or the lemma *character* in its third sense (according to MultiWordNet).

The system will return a KWIC-like concordance of all the tokens in the corpus that match the request, within the sentence in which they occur; each sentence is presented along with its translation. Morphosyntactic information and the WordNet sense are also reported, as shown in Figure 2. An hyperlink connects each semantic concordance to the text-oriented browser, so that the user can easily get the bi-text in which a certain sentence occurs.

In Figure 2, the user has asked for the semantic concordance of the lemma *character* as a noun. Three aligned sentence in which the lemma occurs can be seen in the picture. Note that both singular and plural forms of the lemma have been selected, and the various senses of the word *character* (nr. 4, 3, and 2 with reference to MultiWordNet) are all translated with different Italian words.

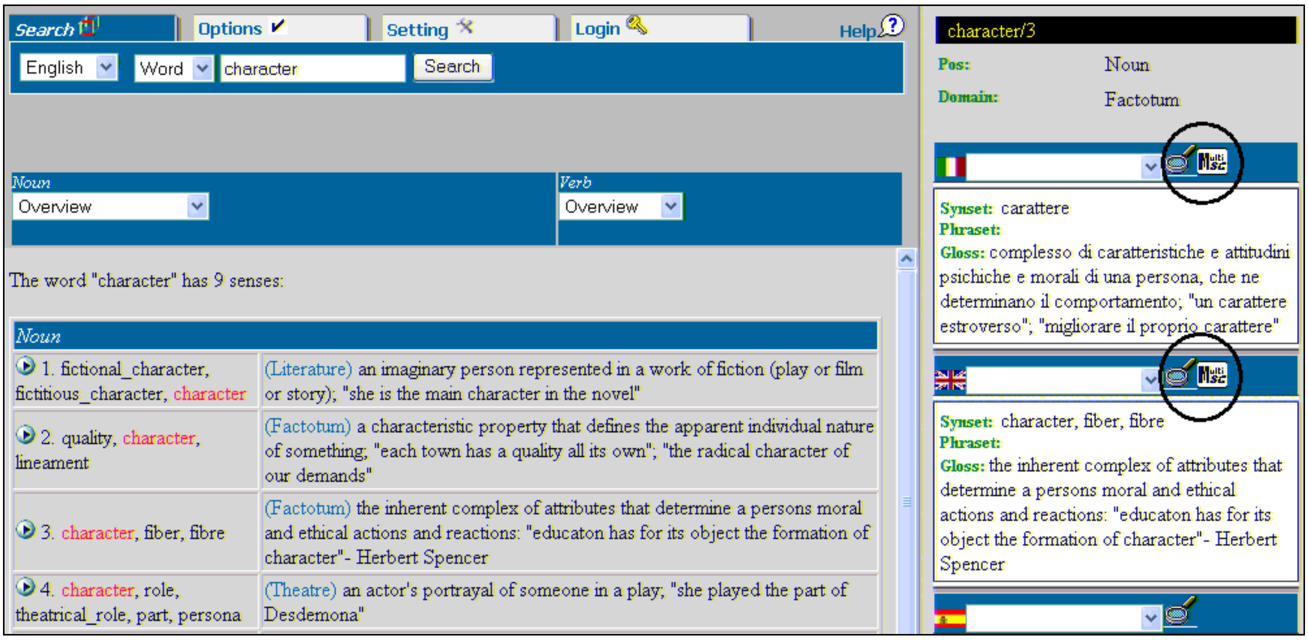
## 3.3 Integration with MultiWordNet

Another important characteristic of the MultiSemCor Web interface is that it allows for the integration between the semantically annotated corpus and its reference lexicon, i.e. MultiWordNet.

This integration has a twofold effect. On the one side, while browsing the MultiSemCor word senses the user can consult MultiWordNet for a better understanding of the semantic annotation. On the other side, while browsing MultiWordNet the user can get examples of usage of a certain word sense from MultiSemCor. To our knowledge, MultiSemCor is the first interface to a multilingual corpus integrated with an on-line lexical resource.

The same form used in Figure 2 to ask for a semantic concordance, can be exploited to access the MultiWordNet lexical information related to a word form or lemma. See the “MultiWordNet” button next to the “MultiSemCor” button in the picture above. Figure 3 shows the result of searching lexical information about the lemma *character* in the standard MultWordNet interface. The two circles in the picture highlight two special icons. Clicking on one of them amounts to activating the MultiSemCor semantic concordancer on the specific sense which is in the focus of the interface.

From an implementation point of view, the MultiSemCor browser has been developed in PHP. The MultiSemCor corpus is encoded according to the XCES guidelines and it is stored in a MySQL database.



English | Word | character | Search

The word "character" has 9 senses:

| Noun | |
|---|---|
| 1. fictional_character, fictitious_character, character | (Literature) an imaginary person represented in a work of fiction (play or film or story); "she is the main character in the novel" |
| 2. quality, character, lineament | (Factotum) a characteristic property that defines the apparent individual nature of something; "each town has a quality all its own"; "the radical character of our demands" |
| 3. character, fiber, fibre | (Factotum) the inherent complex of attributes that determine a persons moral and ethical actions and reactions: "educaton has for its object the formation of character"- Herbert Spencer |
| 4. character, role, theatrical_role, part, persona | (Theatre) an actor's portrayal of someone in a play; "she played the part of Desdemona" |

character/3

| | |
|---|---|
| Pos: | Noun |
| Domain: | Factotum |

Synset: carattere
Phraset:
Gloss: complesso di caratteristiche e attitudini psichiche e morali di una persona, che ne determinano il comportamento; "un carattere estroverso"; "migliorare il proprio carattere"

Synset: character, fiber, fibre
Phraset:
Gloss: the inherent complex of attributes that determine a persons moral and ethical actions and reactions: "educaton has for its object the formation of character"- Herbert Spencer

Figure 3: The MultiWordNet browser

## 4 Related Work

A number of institutions are active to collect, promote, and make available mono- and multilingual language resources and tools. The most important institutions, such as the LDC, ELRA/ELDA, Tractor, UCREL, and RALI, all distribute parallel or multilingual corpora. Also some parallel concordancers have been made available to the community. The most well known are: MultiConcord, ParaConc, WordSmith Tools, Web Concordancer, and TransSearch. Also, a number of projects built parallel corpora, and made them available through a Web interface. The project that is most similar to MultiSemCor is the multilingual English-Catalan-Castillan parallel corpus, developed at Universitat Pompeu Fabre of Barcelona. See http://terminotica.upf.es/academic/. This is the only available interface giving access to word-level alignment. Other on-line interfaces allow for the browsing of sentence-level alignment, and for a token-based search in the text:

- the bilingual English-Chinese parallel corpus by the Hong Kong Virtual Language Center: http://www.edict.com.hk/concordance
- the bilingual English-Portuguese parallel corpus Compara, by the Linguateca group: http://www.linguateca.pt/COMPARA
- the bilingual English-Slovene parallel corpus by the University of Ljubljiana-Slovenia: http://nl2.ijs.si/index-bi.html

Other projects made available only an on-line sample. These are the Web TCE interface to the bilingual English-Norwegian parallel corpus at the University of Oslo, and the TransSearch interface to the Canadian Hansard Corpus.

## 5 Conclusion

In this paper we presented an on-line, freely accessible Web interface to MultiSemCor, a parallel English/Italian corpus, annotated at lexical level. The interface gives access to a large amount of bilingual information through two main modalities, addressing the needs of users with different background. Moreover, it allows for the integrated access to the MultiWordNet on-line lexical database. A first version of the on-line MultiSemCor browser is available at the following address: http://tcc.itc.it/projects/multisemcor.

## References

Bentivogli, L., & Pianta, E. (2002). Opportunistic Semantic Tagging. In Proceedings of the Third International Conference on Language Resources and Evaluation (pp. 1401--1406). Las Palmas, Canary Islands – Spain, May 29-31, 2002.

Fellbaum, C. (ed.) (1998). Wordnet: An Electronic Lexical Database. Cambridge (Mass): The MIT Press.

Pianta, E., & Bentivogli, L. (2003). Translation as Annotation. In Proceedings of the AI*IA 2003 Workshop “Topics and Perspectives of Natural Language Processing in Italy”(pp. 40--48). Pisa, Italy, September 2003.

Pianta, E., Bentivogli, L., & Girardi, C. (2002). MultiWordNet: Developing an aligned multilingual database. In Proceedings of the 1st International Global WordNet Conference (pp. 293--302), Mysore, India, January 21-25, 2002.

MultiWordNet, http://tcc.itc.it/projects/multiwordnet/multiwordnet.php

# Application of Translation Corresponding Tree (TCT) Annotation Schema in Example-Based Machine Translation

**WONG Fai[†], HU Dong Cheng[†], MAO Yu Hang[†], TANG Chi Wai[‡], DONG Ming Chui[‡]**

[†]Speech and Language Processing Research Center, Tsinghua University, 100084 Beijing, China
[‡]Faculty of Science and Technology of University of Macao, PO Box 3001, Macao SAR
derekfw@umac.mo, hudc@mail.tsinghua.edu.cn, myh-dau@mail.tsinghua.edu.cn,
sekevin@inesc-macau.org.mo, dmc@inesc-macau.org.mo

## Abstract

In this paper, we present an Example-Based Machine Translation (EBMT) system for Portuguese to Chinese translation. In our approach, the examples used for translation are annotated under the representation schema of Translation Corresponding Tree (TCT). Each Translation Corresponding Tree describes a translation example (a pair of bilingual sentences). It represents the syntactic structure of source language sentence (i.e. Portuguese in our system), as well as denotes the translation correspondences (i.e. Chinese translation) for each node in the representation tree. In addition, syntax transformation rules are also encapsulated at each node in the TCT representation that captures the differentiation of grammatical structure between the source and target languages. With this annotation schema, translation examples are effectively represented and organized in the bilingual knowledge database. In the translation process, the source sentence is parsed. The output, syntactic tree, is then used for finding the similar TCTs or constituency parts of TCTs from the knowledge DB. By referring to the translation information coded in the TCTs, target language translation is synthesized.

## Introduction

The construction of bilingual knowledge base, in the development of example-based machine translation systems (Sato and Nagao, 1990), is vitally critical. In the translation process, the application of bilingual examples concerns with how examples are used to facilitate translation, which involves the factorization of an input sentence into the format of stored examples and the conversion of source texts into target texts in terms of the existing translations by referencing to the bilingual knowledge base. Theoretically speaking, examples can be achieved from bilingual corpus where the texts are aligned in sentential level, and technically, we need an example base for convenient storage and retrieval of examples. The way of how the translation examples themselves are actually stored is closely related to the problem of searching for matches. In structural example-based machine translation systems (Grishman, 1994; Meyers et al., 1998; Watanabe et al., 2000), examples in the knowledge base are normally annotated with their constituency (Kaji et al., 1992) or dependency structures (Matsumoto et al., 1993), which allows the corresponding relations between source and target sentences to be established at the structural level. All of these approaches annotate examples by mean of a pair of analyzed structures, one for each language sentence, where the correspondences between inter levels of source and target structures are explicitly linked. However, we found that these approaches require the bilingual examples that have '*parallel*' translations or '*close*' syntactic structures (Grishman, 1994), where the source sentence and target sentences have explicit correspondences in the sentences-pair. For example, in (Wu, 1995), the translation examples used for building the translation alignments are strictly selected based on constraints. As a result, these approaches indirectly limit their application in using the translation examples that are '*free translation*' for the development of example-based machine translation system. In this paper, we overcome the problem by designing a flexible representation schema, called Translation Corresponding Tree (TCT). We use the Translation Corresponding Tree (TCT) as the basic structure to annotate the examples in our bilingual knowledge base for the Portuguese to Chinese example-based machine translation system.

## Translation Corresponding Tree Representation

Translation Corresponding Tree structure, as an extension of structure string-tree correspondence representation (Boitet and Zaharin, 1988), is a general structure that can flexibly associate not only the string of a sentence to its syntactic structure in source language, but also allow the language annotator to explicitly associate the string from its translation in target language for the purpose to describe the correspondences between different languages.

### The TCT Structure

The TCT representation uses a triple sequence intervals [SNODE(*n*)/STREE(*n*)/STC(*n*)] encoded for each node in the tree to represent the corresponding relations between the structure of source sentence and the substrings from both the source and target sentences. In TCT structure, the correspondence is made up of three interrelated correspondences: 1) one between the node and the substring of source sentence encoded by the interval SNODE(*n*), which denotes the interval containing the substring corresponding to the node, 2) one between the subtree and the substring of source sentence represented by the interval STREE(*n*), which indicates the interval of substring that is dominated by the subtree with the node as root, and 3) the other between the subtree of source sentence and the substring of target sentence expressed by the interval STC(*n*), which indicates the interval containing the substring in target sentence corresponding to the subtree of source sentence. The associated substrings may be discontinuous in all cases. This annotation schema is quite suitable for representing translation example, where it preserves the strength in describing non-standard and non-projective linguistic phenomena for a language (Boitet and Zaharin, 1988; Al-Adhaileh et al., 2002), on the other hand, it allows the annotator to flexibly define the corresponding translation substring from the target sentence to the representation tree of source sentence when it is necessary. This is actually the idea behind the formalism of Translation Corresponding Tree.

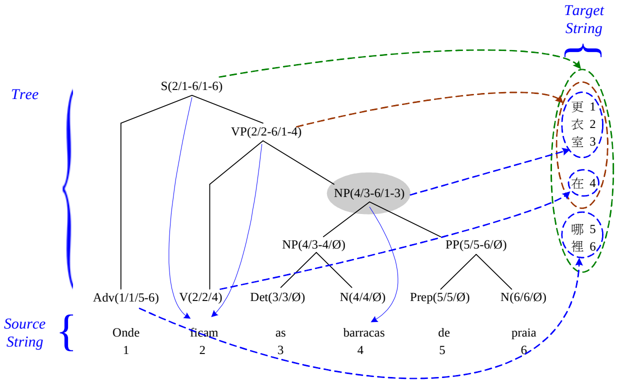


Figure 1: An TCT representation for annotating the translation example "*Onde ficam as barracas de praia? (Where are the bathhouses?)/更衣室在哪裡?*" and its phrase structure together with the correspondences between the substrings (of both the source and target sentences) and the subtrees of sentence in source language.

As illustrated in Figure 1, the translation example “*Onde ficam as barracas de praia?/ 更衣室在哪裡?*” is annotated  in a TCT structure. Based on the interpretation structure of the source sentence “*Onde ficam as barracas de praia?*”, the correspondences between the substrings (of source and target sentences) and the grammatical units at different inter levels of the syntactic tree of the source sentence are expressed in terms of sequence intervals. The words of the sentences pair are assigned with their positions respectively, i.e. “*Onde* (1)”, “*ficam* (2)”, “*as* (3)”, “*barracas* (4)”, “*de* (5)” and “*praia* (6)” for the source sentence, as well as for the target sentence. But considering that Chinese uses ideograms in writing without any explicit word delimiters, the process to identify the boundaries of words is considered to be the task of word segmentation, instead of assigning indices in word level with the help of word segmentation utility, a position interval is assigned to each character for the target (Chinese) sentence, i.e. “*更* (1)”, “*衣* (2)”, “*室* (3)”, “*在* (4)”, “*哪* (5)” and “*裡* (6)”. Hence, a substring in source sentence that corresponds to the node of its representation is denoted by the intervals encoded in SNODE(*n*) for the node, e.g. the shaded node, *NP*, with interval, SNODE(*NP*)=4, corresponds to the substring “*barracas*” in source sentence that has the same interval. A substring of source sentence that corresponds to a subtree of its syntactic tree is denoted by the interval recorded in STREE(*n*) attached to the root of the subtree, e.g. the subtree of the shaded node, *NP*, encoded with the interval, STREE(*NP*)=3-6, corresponds to the substring “*as barracas de praia*” in source sentence. While the translation correspondence between the subtree of source sentence and substring in the target sentence is denoted by the interval assigned to the STC(*n*) of each node, e.g. the subtree rooted at shaded node, *NP*, with interval, STC(*NP*)=1-3, corresponds to the translation fragment (substring) “*更衣室*” in target sentence.

## Expressiveness of Linguistic Information

Another inherited characteristic of TCT structure is that it can be flexibly extended to keep various kinds of linguistic information, if they are considered useful for specific purpose, in particularly the linguistic information that differentiating the characteristics of two languages which are structural divergences (Wong et al., 2001). Basically, each node representing a grammatical constituent in the TCT annotation is tagged with grammatical category (part of speech). Such feature is quite suitable for the describing specific linguistic phenomena due to the characteristic of a language. For instance, in our case, the crossing dependencies (syntax transformation rules) for the sentence constituents between Portuguese and Chinese are captured and attached to each node in the TCT structure for a constituent that indicates the order in forming the corresponding translation for the node from the subtrees it dominated. In many phrasal matching approaches, such as constituency-oriented (Kaji et al., 1992; Grishman, 1994) and dependency-oriented (Matsumoto et al., 1993; Watanabe et al., 2000), crossing constraints are deployed implicitly in finding the structural correspondences between pair of representation trees of a source sentence and its translation in target. Here, in our TCT representation, we adopted the use of constraint (Wu, 1995) for a constituent unit, where the immediate subtrees are only allowed to cross in the inverted order. Such constraints, during the phase of target language generation, can help in determining the order in producing the translation for an intermediate constituency unit from its subtrees when the corresponding translation of the unit is not associated in the TCT representation.

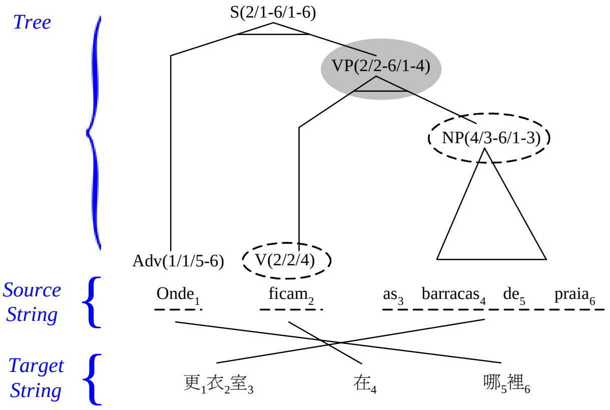


Figure 2: The transfer relationships between the sentence-constituents of source language and its translation in target language are recorded in TCT structure.

Figure 2 demonstrates the crossing relations between the source and target constituents in an TCT representation structure. In graphical structure annotation, a horizontal line is used to represent the inversion of translation fragments of its immediate subtrees.

## Construction of Example Base

In the construction of bilingual knowledge base (example base) in example-based machine translation system (Sato and Nagao, 1990; Watanabe et al., 2000), translation examples are usually annotated by mean of a pair analyzed structures, where the corresponding relations between the source and target sentences are established at the structural level through the explicit links. Here, to facilitate such examples representation, we use the Translation Corresponding Tree as the basic annotation structure.

### TCT Generation

In our example base, each translation pairs is stored in terms of an TCT structure. Conceptually speaking, the construction of the example base can be viewed as the process in building the TCT structures for the example cases. To a translation example, the system will automatically process and generate a preliminary TCT representation structure for it. The resultant annotation tree is then further edited by human through the use of an TCT editing program if any amendment to the representation structure is necessary.

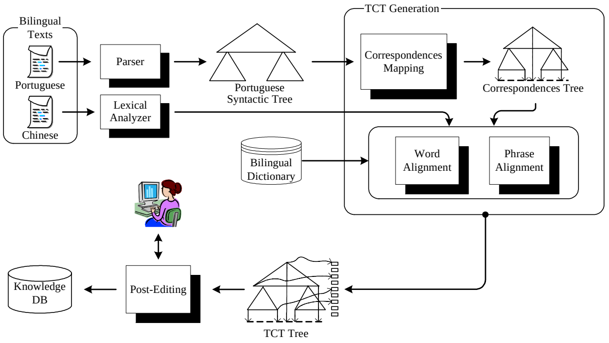


Figure 3: The construction of bilingual knowledge base based on the representation structure of TCT.

In the generation process, it starts by analyzing the grammatical structure of Portuguese sentence with the aid of a Portuguese parser, and a shallow analysis to the Chinese sentence is carried out by using the Chinese Lexical Analysis System (ICTCLAS) (Zhang, 2002) to segment and tag the words with a part of speech. The grammatical structure produced by the parser for Portuguese sentence is then used for establishing the correspondences between the surface substrings and the inter levels of its structure, which includes the correspondences between nodes and its substrings, as well as the correspondences between subtrees and substrings in the sentence. Next, in order to identify and establish the translation correspondences for structural constituents of Portuguese sentence, it relies on the grammatical information of the analyzed structure of Portuguese and a given bilingual dictionary to search the corresponding translation substrings from the Chinese sentence. Finally, the consequent TCT structure will be verified and edited manually to obtain the final representation, which is the basic element of the knowledge base. The overall process in constructing the bilingual knowledge base is depicted in Figure 3, and Figure 4 illustrates the example "*Actos anteriores à publicidade da acção (Publicity of action prior to acts) / 在訴訟公開前所作之行爲*" with its corresponding TCT structure.

### Translation Equivalents

Through the notation of translation corresponding structure for representing translation examples in the bilingual knowledge base, the translation units between the Portuguese sentence and its target translation in Chinese are explicitly expressed by the sequence intervals STREE(*n*) and STC(*n*) encoded in the intermediate nodes of an TCT structure, that may represent the phrasal and lexical correspondences. For instance, from the translation example being annotated under the TCT representation schema as shown in Figure 4, the Chinese translation "*訴訟* " of Portuguese word "*acção*" is denoted by [STREE(*n*)=6/STC(n)=2-3] in the terminal node. For phrasal translation, we may visit the higher level constituents in the representing structure of TCT and apply the similar coding information to retrieve the corresponding translation for the unit that representing a phrasal constituent in a sentence. In order that the representation examples can be effectively consulted, each TCT structure is being indexed by its nodes in the bilingual knowledge base. Thus, all the possible sub-TCTs (translation units) or the constituency structures of an TCT can be easily retrieved for reference.

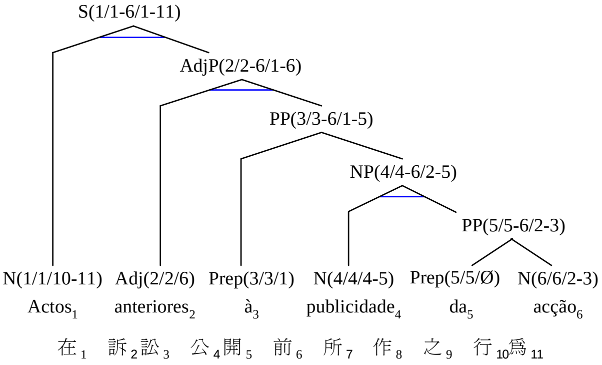


Figure 4: A TCT structure constructed for the translation example "*Actos anteriores à publicidade da acção (Publicity of action prior to acts) / 在訴訟公開前所作之行爲*".

## Example-Based Translation Based on TCT

In example-based machine translation systems, a corpus of translation examples used to facilitate the translation rather than linguistic rules is the significant component (Sato and Nagao, 1990). In our approach, translation examples are annotated under the representation structure of TCT. Each TCT structure consists of a sentence in source language, e.g. Portuguese in our case, an associated constituency structure that describing the source sentence, the mapping between the inter levels of abstracted structure and its surface string of the sentence, as well as the corresponding relations against its translation in target language, e.g. Chinese, including the translation fragments and the constraints of crossing dependencies between the source and target phrasal units. During the translation process, a new input sentence is first analyzed into the form of representation structure, followed by retrieving the related examples that contain the same words or comprise the same constituency structures as the input sentence from the example base, and use them to synthesize the final translation for the input sentence guided by the syntactic information of sentential constituents and the translation correspondences of the referenced examples. The overall picture of the translation processes is depicted in Figure 5.

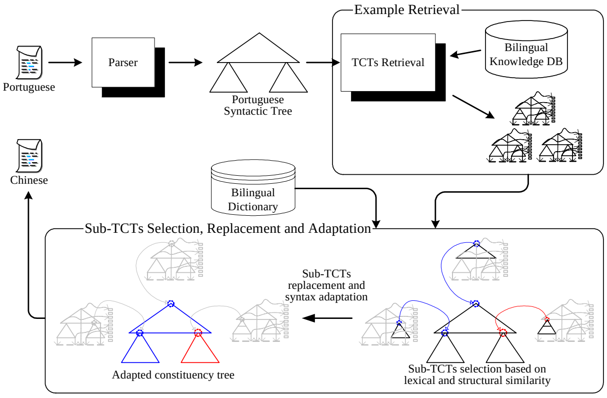

Portuguese
Parser
Portuguese Syntactic Tree
Example Retrieval
TCTs Retrieval
Bilingual Knowledge DB
Chinese
Bilingual Dictionary
Sub-TCTs Selection, Replacement and Adaptation
Sub-TCTs replacement and syntax adaptation
Adapted constituency tree
Sub-TCTs selection based on lexical and structural similarity

Figure 5: The overall translation processes by using the TCT representation examples as the bilingual knowledge base (example base).

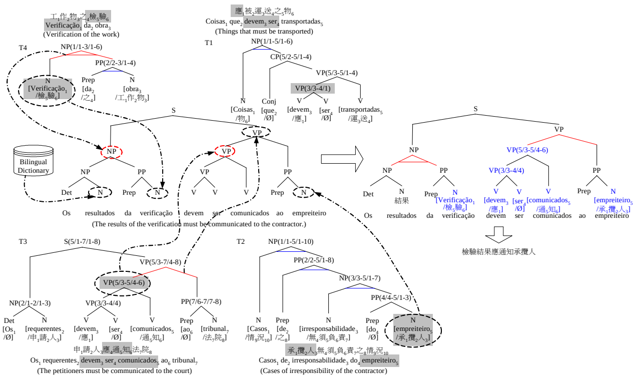


Figure 6: Translation by matching and replacing.

To translate a Portuguese sentence, in our system, can be viewed as the process to construct an TCT structure for describing the input sentence guided by the collection of annotated TCT representations of examples from the example base, follow by traversing the resultant representation structure according to the order being controlled by the crossing constraints encoded in each node (grammatical unit) to produce the target translation for the source sentence in Chinese. During the process, the internal structure of the source sentence is first analyzed with the help of a parser and a syntactic representation tree of the sentence is produced as the parsing result. Then for each subgraph (constituency unit) of the constructed tree, the system retrieves a list of close related TCTs or sub-TCTs from the example base based on the constraint that the constituency units (TCTs or sub-TCTs) that have similar grammatical structure (as well as the grammatical categories labeled for the root nodes and the dominated nodes) as that of the source sentence are recalled. In addition, the content words of the root node of the constituency unit will also be considered for determining the examples that are completely matched to the source sentence. After the related examples are identified and obtained from the example base, the next step is to select the set of TCTs or sub-TCTs to form a complete TCT structure that can best describe the source sentence by replacing the subtrees of source sentence with the chosen sub-TCTs. For those of unmatched terminal nodes, the corresponding Chinese translation can be consulted from a given bilingual dictionary and filled to complete the construction of TCT structure for the sentence. In the case if more than one example is found, the system will evaluate the distance between the chosen examples and the source sentence based on the edit distance function. The replacement process to construct the target TCT for the source sentence is demonstrated in Figure 6. Finally, the corresponding translations appeared in the resultant TCT structure are combined to form the target translation in Chinese.

## Acknowledgement

The research work reported in this paper was supported by the Research Committee of University of Macao under grant CATIVO:3678.

## References

Al-Adhaileh, M.H., Tang, E.K. & Zaharin, Y. (2002). *A Synchronization Structure of SSTC and Its Applications in Machine Translation*. The COLING 2002 Post-Conference Workshop on Machine Translation in Asia, Taipei, Taiwan.

Boitet, C. & Zaharin, Y. (1988). *Representation trees and string-tree correspondences*. In Proceedings of COLING-88, Budapest, pp.59-64.

Grishman, R. (1994). *Iterative Alignment of Syntactic Structures for a Bilingual Corpus*. In Proceedings of Second Annual Workshop on Very Large Corpora (WVLC2), Kyoto, Japan, pp.57-68.

Kaji, H., Kida, Y. & Morimoto, Y. (1992). *Learning Translation Templates from Bilingual Text*. In Proceedings of CoLING-92, Nantes, pp.672-678.

Matsumoto, Y., Isimoto, H. & Utsuro, T. (1993). *Structural Matching of Parallel Texts*. 31st Annual Meeting of the Association for Computational Linguistics, Columbus, Ohio, pp.23-30.

Meyers, A., Yangarber, R. & Ralf, B. (1998). *Deriving Transfer Rules from Dominance-Preserving Alignments*. In Proceedings of Coling-ACL (1998), pp.843-847.

Sato, S. & Nagao, M. (1990). *Toward Memory-Based Translation*. In Proceeding of Coling (1990), Vol.3, pp.247-252.

Watanabe, H., Kurohashi, S. & Aramaki, E. (2000). *Finding Structural Correspondences from Bilingual Parsed Corpus for Corpus-based Translation*. In Proceedings of COLING-2000.

Wong, F., Mao, Y.H., Dong, Q.F. & Qi, Y.H. (2001). *Automatic Translation: Overcome the Barriers between European and Chinese Languages*. In Proceedings (CD Version) of First International UNL Open Conference 2001, SuZhou China.

Wu, D. (1995). *Grammarless extraction of phrasal translation examples from parallel texts*. In Proceedings of TMI-95, Sixth International Conference on Theoretical and Methodological Issues in Machine Translation, v2, Leuven Belgium, pp.354-372.

Zhang, H.P. (2002). *ICTCLAS*. Institute of Computing Technology, Chinese Academy of Sciences: http://www.ict.ac.cn/freeware/003_ictclas.asp.

# Improving Word Alignment Quality Using Linguistic Knowledge

**Bettina Schrader**

Cognitive Science Doctorate Programme: Rules and Patterns
Institute for Cognitive science, University of Osnabrück
Kolpingstraße 7, D-49069 Osnabrück, Germany
bschrade@uos.de

**Abstract**

Word alignment of bilingual parallel corpora is usually generated using only statistical information. External linguistic information like e.g. a dictionary or linguistic structural annotation of the texts is used rarely, despite its usefulness. Additionally, it has to our knowledge never been examined systematically how linguistic information can be employed for word alignment improvement. In this paper, we present our experiments on finding out which linguistic information has which effect on word alignment quality, and we evaluate our experiments using precision and recall calculated for dictionaries that were generated after word alignment. The experiments show that information on e.g. lemmas and word category is useful to increase recall without lowering precision. Additionally, we discuss whether linguistic information can be used to compensate weak points of standard word alignment systems, and which features an ideal procedure should possess.

## 1. Introduction

Word alignment is an important technique in the exploitation of bilingual parallel corpora for lexicography, statistical machine translation, and cross-linguistic information retrieval (CLIR). It is used to automatically detect word pairs of translational equivalence, i.e. it computes which word in target language L2 is a translation of a word in source language L1.

Different word alignment techniques have been developed (cf. (Brown et al., 1990)), usually based on statistical information. Additionally, several researchers have experimented with combining linguistic and statistical information (Nießen and Ney, 2000). Still, the usefulness of linguistic information for word alignment has to our knowledge never been examined systematically.

The purpose of the experiments we are presenting here is to find out which linguistic information, whether on lemmas, word category or systactic constituency, can be used efficiently for word alignment. Additionally, we investigate which flaws standard alignment techniques have, and how they can be compensated. Experiments are evaluated using precision and recall calculated for 50-60 sample word pairs per corpus taken from automatically generated dictionaries after word alignment was done.

The paper is organized as follows: First, we give an overview on standard approaches to word alignment. Secondly, we introduce our the corpora and describe which linguistic tools were used for linguistic preprocessing. Then, we report on the experiments conducted and discuss their results.

## 2. Standard approaches to word alignment

Standard word alignment approaches like the ones by (Brown et al., 1990), (Brown et al., 1993), (Vogel et al., 1999), or (Hiemstra, 1996) make use of statistical models to derive word alignments.

(Brown et al., 1990) have been the first to publish a word alignment procedure. It consists of a cascade of five statistical translation models of increasing complexity. The first model of (Brown et al., 1990), IBM-1, treats every sentence as a *bag of words*, where the position of a word in a sentence does not have any influence on its translation probability. IBM-2 to IBM-5 refine this notion by introducing statistical weights such as *distortion* and *fertility* to account for word order phenomena and 1-to-many alignments.

The two competing standard alignment models, by (Vogel et al., 1999) and (Hiemstra, 1996) correspond most closely to the IBM-1 model: The HMM-model by (Vogel et al., 1999) treats a sentence mainly as a bag of words, but the probality of an alignment is influenced by the preceding alignment. (Hiemstra, 1996) uses a pure bag of words model. In contrast to (Brown et al., 1990) and (Vogel et al., 1999), he doesn't focus on the translation model, but instead uses word alignment as a means to generate a dictionary for CLIR.

All three approaches to word alignment do not use explicit linguistic knowledge, whether in form of a dictionary or in form of linguistic structural information, because these approaches are set up to be language independent, i.e. they are supposed to work equally well for each possible language pair. Researchers have, however, found it necessary to experiment on improving word alignment systems with linguistic knowledge: (Nießen and Ney, 2000) e.g. *manipulate* their parallel corpora: word order in one language e.g. is changed to resemble more closely word order in L2, in order to circumvent distortion problems caused by syntactic differences between L1 and L2.

## 3. Corpora

Three parallel German-English corpora were used for the experiments: debate protocols of the European Parliament (MLCC), a subset of the Linux manpages (MANPAGES), and a small corpus consisting of patent abstracts (PATENTE).

All corpora were tokenized, POS-tagged, and lemmatized using the tree-tagger by (Schmid, 1994). Two corpora were chunked using an extension of the tree-tagger (Schmid, unpublished) for the English, and the tool by (Kermes, 2003) for the German texts. All corpora were sentence aligned using an aligner that was developed as part of the IMS corpus workbench and that combines various sentence alignment strategies (Evert, , p.c.).

Only *secure* sentence pairs from all corpora were used in the experiments, and manipulated to include only the kind of linguistic information that was necessary. For one experiment e.g., tokens were included in the input only if they formed part of a nominal or prepositional chunk. Sentence pairs were considered secure if they occurred in a sequence of at least three 1:1-alignments. This condition was applied to ensure that the text used in the experiments was 100% correct.

For word alignment, we used the alignment tool by Hiemstra, 1996), as it automatically generates a bilingual dictionary in easy-to-read format.

### 3.1. MLCC

This parallel text is part of the corpus *Multilingual and Parallel Corpora for Cooperation* (MLCC) provided by ELRA[1] and consists of debate protocols of the European Parliament between 1992 and 1994. They were preprocessed and added to the IMS corpus workbench independently of our experiments.

After sentence alignment and restricting the data set to secure sentence pairs, it consists of 1,713,796 tokens in 78,130 sentence pairs. In the course of the experiments, the set of sentence pairs has been reduced further to a random sample of 2500 sentences due to software restrictions.

### 3.2. MANPAGES

The MANPAGES corpus consists of texts from the Linux online help for shell commands that are available in English and German. They have been reformatted removing all paragraphs except the sections NAME / NAME, BESCHREIBUNG / DESCRIPTION and ÜBERSICHT / ZUSAMMENFASSUNG / SYNOPSIS as only these sections consist of coherent text. After preprocessing and applying the restriction on secure alignments, the MANPAGES consist of 14,759 tokens in 860 sentence pairs.

### 3.3. PATENTE

The smallest corpus consists of patent abstracts in German and English that were provided by courtesy of the German Patent Office. After preprocessing and reduction to secure alignments, the corpus is made up of only 125 sentence pairs with 3,204 tokens. Although this size is much too small for a statistical alignment method, it is used for the experiments as the translations provided are very good and close to the original texts.

## 4. Experiments

We test in several experiments how information on word category, lemmas and syntactic costituency influences word alignment quality. Two experiments and the baseline are carried out on all three corpora, while the other experiments are done on only one or two of the corpora for reasons given in each experiment description.

### 4.1. Baseline

To be able to compare the experiment results to what a pure word alignment procedure is capable, a baseline has been created: all corpora have been word aligned using only the sentence aligned text, i.e. no linguistic information has been used.

### 4.2. Functional Class Words

First, we removed all words belonging to a functional class such as determiner or preposition from the texts. Words of the lexical classes nouns, adjectives, and verbs, remained in the corpus. POS-tags are used to distinguish between both groups of words.

The reason for removing function words is that they are uninteresting from a lexicographic point of view as they don't carry lexical meaning. Additionally, the number of function words per language is fixed, so that they are probably listed in any existing dictionary, and can be aligned easily using one.

### 4.3. Lemmas

In morphologically rich languages, words may only differ from each other due to their inflections, while their meaning stays the same. If such word forms are aligned, each of them will be treated as unique and will be aligned as such, i.e. two word forms of the same lemma in L1 can be set into translational equivalence with two tokens from L2 that may or may not share the same lemma. This happened e.g. in the baseline for German *Verhandlung/ Verhandlungen* (English: *negotiation/ negotiations*): With this consideration in mind, we should not align word forms but rather abstract away from inflections and use lemmas for alignment.

| Verhandlung | | Verhandlungen | |
|---|---|---|---|
| translation | probability | translation | probability |
| you | 0.65 | negotiations | 0.98 |
| followed | 0.31 | process | 0.02 |
| All | 0.03 | | |

Table 1: Baseline dictionary excerpt: MLCC corpus

In morphologically poor languages, on the other hand, favouring lemmas does not influence word alignment as much. We therefore refrained from lemmatizing the English texts. We have, however, lemmatized the German texts and aligned it with the unlemmatized English texts. Additionally, function words have been removed.

### 4.4. Lexicon

We also tested whether alignment quality is improved if we add data from an English-German dictionary, in this case the (Langenscheidts Handwörterbuch, 1991). For each corpus, a vocabulary list was compiled containing all nouns that occurred both in the corpus and in the dictionary, and the list was appended to the corpus. This procedure was necessary as the word aligner did not support direct lexicon lookup during the alignment process.

This experiment was carried out on the two corpora PATENTE and MANPAGES, only. MLCC proved too big for the addition of vocabulary in initial tests.

[1] http://www.icp.inpg.fr/ELRA/index.html

### 4.5. Morphology

Correctly aligning German compounds with their English equivalents is a problem for word alignment as German compounds usually correspond to English multi word units, i.e. they do not stand in a 1:1 relationship. The German compound "Dämpfungsscheibenanordnung" e.g. corresponds to the three subsequent tokens "dampening disk assembly" in English.

Splitting the compound in its components would solve this problem, however: "Dämpfungsscheibenanordnung" consists of the three elements "Dämpfung", "scheibe", and "anordnung" that can easily be aligned with the three elements of the corresponding English expression in three 1:1 alignments[2]. Therefore, we decomposed all German complex nouns of the PATENTE corpus using the morphological tool DEKO (Schmid et al., 2001) and replaced them by their decomposed sequence of elements before aligning the corpus.

This experiment was carried out only on the smallest corpus, the PATENTE corpus, as compound decomposition is a very time-consuming task.

### 4.6. Chunks

In our final experiment, we tested whether shallow syntactic information is useful for word alignment, too. For this reason, the corpora MLCC and MANPAGES were chunked, and all tokens that did not belong to a nominal or prepositional chunk were deleted.

The MANPAGES corpus has proven too sloppily translated to allow for successfull chunking, so that we have not run this experiment on this corpus.

## 5. Evaluation

For the evaluation, we constructed tokenlists and compared them to the dictionaries generated during word alignment. Precision and recall were chosen as evaluation measures, and we examined only the translation direction German → English.

For each corpus, we compiled a tokenlist containing the 50-60 most frequent nouns of the corpus[3]. and translated them manually. This sample size is small enough to allow for manually examining the data, and sufficiently big to allow an analysis of the experiment results. We restricted the tokenlists to nouns, because new words are often created as such. We defined precision and recall such that:

$$\text{precision} = \frac{\text{\# correct translations}}{\text{\# suggested translations}}$$

and

$$\text{recall} = \frac{\text{\# correct translations}}{\text{\# manually assigned translation}}$$

The number of translations is given by the number of words of the English translation. In the case of a multi word unit like “child process”, each element is counted as correct translation candidate, i.e. “child process” counts with two correct translations.

Translation candidates of the dictionaries were ignored if their translational probability was below 10%.

| **Precision (%)** | MLCC | MANPAGES | PATENTE |
|---|---|---|---|
| Baseline | 59 | 64 | 35 |
| Function words | 54 | 58 | 43 |
| Lemmatization | 50 | 46 | 46 |
| Lexicon | – | 47 | 53 |
| Morph. Decomposition | – | – | 37 |
| Chunks | 55 | – | 42 |

Table 2: Precision values for all experiments

As can be seen in the tables, the precision of the dictionaries created during the experiments is lower than the value of the baseline. The only exception is the results of the PATENTE corpus, where all experiment precisions are higher than in the baseline.

Recall, on the other hand, is higher in all experiments on all corpora and increases up to 98%.

| **recall (%)** | MLCC | MANPAGES | PATENTE |
|---|---|---|---|
| Baseline | 90 | 84 | 67 |
| Function words | 95 | 84 | 91 |
| Lemmatization | 95 | 87 | 88 |
| Lexicon | – | 90 | 89 |
| Morphology | – | – | 76 |
| Chunks | 98 | – | 71 |

Table 3: Recall values for all experiments

To find out why precision values for the experiments are lower than the precision of the baseline, the dictionaries were more closely examined: We found out that the number of translation candidates per token is higher in the experiment dictionaries than in the baseline. Additionally, the baseline dictionary has a lower coverage than the other experiment dictionaries.

Precision as calculated here obviously does not describe dictionary quality completely enough: For once, it punishes alternatives - the more translation candidates are given per token, the lower precision will be. Secondly, precision is higher if a word is missing from the dictionary then if it is listed with at least one wrong suggestion (See example in table 4), i.e. differences in coverage are not taken into account.

| Headword: Ergebnis (result) | | | | |
|---|---|---|---|---|
| Experiment | word | probability | word | probability |
| Baseline | no suggestions | | | |
| Function Words | | | | |
| Lemmatization | results | 0.97 | portable | 0.03 |
| Lexicon | result | 1.00 | | |

Table 4: Dictionary excerpts: Manpages corpus

If we take the problems with calculating precision into account, we assume that linguistic processing does not influence precision negatively despite evaluation numbers.

[2] Linking elements have been omittd for this example.

[3] 50 tokens each were chosen for PATENTE and MANPAGES; the tokenlist for the corpus MLCC contains 60 items as it is bigger than the other two corpora.

The analysis of the experiments shows as well, however, that some word alignment problems remain: Using linguistic information by means of text manipulation always means restricting oneself to one kind of knowledge, as a statistical model like that by (Hiemstra, 1996) allows for only one level of linguistic description – chunks e.g. can be used iff sentences are made up only of chunk material. Hence there is always some loss of information. Additionally, information on sentence-internal structure, like e.g. chunk boundaries, cannot be preserved and used as alignment clues: If we restrict the input to words occurring in noun or prepositional chunks and mark chunk boundaries, the alignment tool treats chunk boundaries in the same way as words.

Finally, a simple bag of words model is not able to align single words with multi word units correctly, as is necessary in the case of German compounds and their corresponding English multi word units. Even a morphological decomposition of compounds does not help much, as is seen in the experiments. The reason is that we cannot expect that the equivalent of a compound is a complex expression in itself - the German compound "Abstandselement" e.g. is equivalent to simplex "spacer". Additionally, even if the translation of a compound is morphologically complex, it need not be compositional as well: German "Schutzelement" is translated by "shield cushion" - where there is no correspondence between German "Element" and English "cushion" ( "cushion" translated to German means "Kissen", "pillow").[4]

## 6. Conclusion

In this paper, we systematically investigated which linguistic information can be used for improving word alignment quality: Lexical information and information on lemmas, word category, morphology and syntactic constituency were used to manipulate three parallel corpora before aligning them. Afterwards results were evaluated calculating precision and recall for the dictionaries generated during word alignment, and the dictionaries were examined in more detail. Experiment results show that linguistic information is useful in increasing recall. Precision as calculated here is not sufficient to determine the influence of linguistic information on word alignment in terms of correctness of the established translation correspondences. We have reason to assume, however, that precision was not decreased during the experiments.

However, using linguistic information for sophisticated text manipulation does not compensating flaws of a standard word alignment approach: Using it means loss of information elsewhere, and sentence-internal structure cannot be used as alignment clues.

A word alignment system should be able to parse linguistically annotated text, so that one level of linguistic description, e.g. lemma information, can be used to align while preserving all other information, e.g. on word forms. Additionally, it should be able to parse and preserve sentence-internal structure, e.g. chunks: if two chunks c1 and c2 are equivalent toeach other, then the words in c1 and c2 will be equivalent to each other as well. Concerning multi word units, it should be possible to align across levels, so that a word in L1 (e.g. a German compound noun) is aligned with its corresponding chunk in L2 (an English multi word expression).

## 7. Acknowledgements

We thank the German Patent Office for supplying the texts for the PATENTE corpus, and Langenscheidt for the permit to use the electronic version of the Langenscheidts Handwörterbuch, 1991).

## 8. References

Brown, Peter F., John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin, 1990. A statistical approach to machine translation. *Computational Lingusitics*, 16:79–85.

Brown, Peter F., Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer, 1993. The mathematics of machine translation: Parameter estimation. *Computational linguistics*, 19(2):263–311.

Evert, Stefan. personal communication.

Hiemstra, D., 1996. *Using statistical Methods to create a bilingual Dictionary*. Master's thesis, Universiteit Twente.

Kermes, Hannah, 2003. *Off-line (and On-line) Text Analysis for Computational Lexicography*. Ph.D. thesis, IMS, University of Stuttgart. Arbeitspapiere des Instituts für Maschinelle Sprachverarbeitung (AIMS), volume 9, number 3.

Langenscheidts Handwörterbuch, 1991. Langenscheidts Handwörterbuch Deutsch / Englisch, Englisch / Deutsch. 3. Auflage.

Manning, Christopher D. and Hinrich Schütze, 1999. *Foundations of statistical natural language processing*. Cambridge, Massachusetts, London: MIT Press.

Nießen, Sonja and Hermann Ney, 2000. Improving SMT quality with morpho-syntactic analysis. In *Proceedings of the 18th International Conference on Computational Linguistics (COLING 2000)*. Saarbruecken, Germany.

Schmid, Helmut, 1994. Probabilistic part-of-speech tagging using decision trees. In *International Conference on New Methods in Language Processing*. Manchester, England.

Schmid, Helmut, unpublished. The IMS Chunker. Unpublished manuscript.

Schmid, Tanja, Anke Lüdeling, Bettina Säuberlich, Ulrich Heid, and Bernd Möbius, 2001. DeKo: Ein System zur Analyse komplexer Wörter. In *GLDV - Jahrestagung 2001*.

Vogel, Stephan, Hermann Ney, and Christoph Tillmann, 1999. HMM-based word alignment in statistical translation. In *Proceedings of the International Conference on Computational LInguistics*. Copenhagen, Denmark.

[4] All examples are taken from the PATENTE corpus.

# Using Comparable Corpora for Discovering Universals in Surface Structure

**Dr. John Elliott**
Computational Intelligence Research Group,
School of Computing,
Leeds Metropolitan University,
Leeds, LS6 3HE, England
j.elliott@leedsmet.ac.uk

## Abstract

Many aspects of linguistic research, whatever their aims and objectives, are reliant on cross-language analysis for their results. In particular, any research into generic attributes, universals, or inter-language comparisons, requires samples of languages in a readily accessible format, which are 'clean' and of adequate size for statistical analysis. Implicit in such understanding and detection of 'universal' attributes of language, is the need to study and analyse a representative set of the human language chorus. So, as an ongoing process during recent years, many raw text samples, in electronic format, have been collected to create a suitably diverse repository. Predominantly, the texts attained are freely available on a variety of sites over the Internet and cover all of the major language groups. These comprise Austro-Asiatic, Amerindian, Sino-Tibetan, Indo-European (Indo-Iranian, Hellenic, Celtic, Italic, Germanic and Slavic) Austroesian, Attaic, Uralic, Niger-Congo and independents and currently total over fifty language scripts.

## Introduction

The goal for my research is to derive structural language universals and unsupervised techniques for language discovery from a representative set of the human language chorus In addition to analysing raw text from over 50 languages, I am also endeavouring to apply these same principles to how the hidden layer of parts-of-speech interact, to glean a more complete and generic comparative picture linguistically. So, an additional goal has been to compile a repository of an equivalent representative set of freely available, or donated, tagged corpora. To date, in addition to acquiring suitable English corpora, which are almost ubiquitous and freely available to the Corpus Linguistics community, part-of-speech annotated corpora have been collected for Romanian, Arabic, Bulgarian, Czechoslovakian, Cuban-Spanish, Chinese, Japanese, Dutch, German, Hindi, Hungarian, Portuguese, Thai, and Turkish. Acquiring these non-English tagged corpora has been a difficult and long process. Nevertheless, over recent years, corpora for the languages listed above have gradually become available in varying states of readiness and size. A significant problem encountered during analysis, is the plethora of grammatical mark-up schemes, which are employed across the range of corpora available to the academic community. Whether inter or intra-language, each author adapts to varying degrees of granularity their own set of parts-of-speech and formats for mark-up.

To expedite comparative analysis of these corpora, it was imperative that a set of tools were created to facilitate 'cleaning up' these annotation schemes, to provide both a common, comparable set of tags and a system that can cope with the many formats evolved: e.g. extended ASCII and Big 5. The ergodic assumption that underpins the rationale for representative samples providing entropic stability for robust analysis is an important premise, as with it relatively small samples provide meaningful statistical results, where source scripts are limited. Utilising this premise, a multilingual parallel corpus of over 50 languages was created to support comparative analysis.

## Sample Size

Sample size is often a hotly debated topic, and answers to this particular question usually gravitate towards 'the bigger the better' (EAGLES, 1996) particularly with respect to issues of data sparseness and word prediction accuracy derived from inter-word statistics (Lesher, Moulton & Higginbottom, 1999), but just as often for pragmatic reasons. However, normal theoretical principles of statistical sampling and inference do not apply, as it is often impossible to delimit the total population in any rigorous way. There is also no obvious unit of language, which is to be sampled and which can be used to define the population (Atkins et al, 1992).

Given that most analyses for this research looks at the physical make-up of the language surface structure, as opposed to its semantics, letter ngrams were chosen to ascertain what constitutes a representative sample. An additional consideration imposed on any sample's minimum length, is the need to filter out random events: a plastic cup, blown bouncing down the road, can sound so like an approaching horse, as to fool any hearer or sound recognition system. Randomness therefore has the potential to mimic structure on occasion but it is an attribute unlikely to persist for more than just a short period of time. With this consideration, texts of varying length, both natural language and randomly generated letters were analysed for their coverage of bigram and trigram combinations to ascertain when convergence occurs. This was hypothesised to provide a method for calculating both the convergence of natural language bigram and trigram letter combinations, for minimum representative text length, and the point when randomly generated text displays non language-like behaviour to the point that its presence is transparent. To add further weight to this rationale, preliminary statistical analyses shows that after approximately 14,000 words, reliable scores can be obtained in machine translation output evaluations (when comparing several systems), and that any additional sampling only serves to confirm results (Elliott, D et al, 2003).

Results from analysing bigram and trigram occurrences show that the length of a sample can significantly affect the percentage of ngrams 'discovered'. Most dramatic of all is the increase in trigrams for randomly generated text. From this, it can be seen that the orthotactic constraints of natural language restrict ngram combinations considerably: letter bigrams remain below 70% and letter trigrams below 20%. However, randomly generated text has no such constraints and rapidly converges towards 100% coverage of all possible combinations: bigrams within 5000 letters and trigrams within 47000 letters. It is also observed that random trigrams exceed natural language constraints after a text length of 25000 letters. Therefore, it can be inferred that as little as 3000 characters can reliably indicate the presence of a random event, due to the immediacy of bigram convergence to 100% coverage. However, to err on the side of caution (a belt and braces approach) trigram convergence parameters were taken into consideration. This sets the sample minimum at approximately 47,000 letters or 10,000 words: wherever possible, this minimum is raised to 100,000 letters or 20,000 words, to cater for any minor increases in natural language orthotactic representations obtained by greater sampling, together with their representative probabilities.

## Corpus Mark-Up

The grammatical mark-up of constituent languages across comparable corpora is a major issue when conducting inter-linguistic analysis, due to both the diversity of the tag-sets adopted, often created for individual corpus aims, and linguistic interpretation of the authors. Even at the level of lexico-grammatical word-class annotation (Part-of-Speech word tagging), which corresponds to layers 'a' and 'b' outlined in the EAGLES report (EAGLES, 1996), there is a great diversity of schemes and models available. Here an example sentence is tagged according to several alternative tagging schemes and vertically aligned (Atwell 1996).

| | Brown | ICE | LLC | LOB | PARTS | POW |
|---|---|---|---|---|---|---|
| *select* | VB | V(montr,imp) | | VA+0 | VB | adj |
| *the* | AT | ART(def) | | TA | ATI | art |
| *text* | NN | N(com,sing) | | NC | NN | noun |
| *you* | PPSS | PRON(pers) | | RC | PP2 | pron |
| *want* | VB | V(montr,pres) | | VA+0 | VB | verb |
| *to* | TO | PRTCL(to) | | PD | TO | verb |
| *protect* | VB | V(montr,infin) | | VA+0 | VB | verb |
| . | . | PUNC(per) | | . | . | . |

EAGLES layers of syntactic annotation:

(a) Bracketing of segments
(b) Labelling of segments
(c) Showing dependency relations
(d) Indicating functional labels
(e) Marking sub-classification of syntactic segments
(f) Deep or 'logical' information
(g) Information about the rank of a syntactic unit
(h) Special syntactic characteristics of spoken language

In Jan Cloeren's paper, which evaluates schemes for a cross-linguistic tagset (Cloeren, 1993), the tagging of Germanic languages is considered in detail, with the following conclusion as its basic cross-linguistic tagset:

Noun
Pronoun
Article
Adjective
Numeral
verb
Adverb
Preposition
Conjunction
Particle
Interjection
Formulaic expression

Erjavec, Ide and Tufis compare, by language, the number of attributes in each part of speech for six central and eastern European languages (Erjavec et al, 98). Their conclusions, illustrate both granularity and, perhaps more importantly, the absence of features in individual languages in accordance with morpho-syntactic descriptions (MSDs) developed from proposals in the EAGLES project, with subsequent modifications for the Multi-East project. A zero in their analysis indicates that it distinguishes no features for that part-of-speech. However, interpretation of grammatical tokens suffers from a lack of classification universality and devices indicated as absent or rare in a language may well exist.

This issue becomes crucial, when inheriting annotation schemes and expert lexico-grammatical word-class annotation classification rationales, for the interpretation of what criteria constitute the allocation of a word-tag pairing: is a verb a verb in every language regardless of the fact that the word describes an action? To illustrate this point, the following words were entered into an online translator for single words entries of Thai-English to ascertain whether the parts-of-speech allocated agreed with the English interpretation.

| *Word* | *Thai PoS classification* |
|---|---|
| Beautiful | V[8], N[1] |
| Sweet | V[1] |
| Old | N[3], V[5] |
| Tall | V[1] |
| Happy | V[5] |
| Blue | N[2] |
| White | V[2], N[1] |
| Fat | N[2] |
| Ugly | V[3] |
| Clever | N[1], V[1] |
| Quick | V[4] |
| Slow | V[4], N[1] |
| Big | V[1] |

Results for these predominantly adjectival words in the English language illustrate how classification can differ markedly, in addition to any labelling or granularity issues. This simple exercise demonstrates that any meta-tagging, in addition to providing a consistent, comparative baseline, will ideally need to consider such differences in interpretation during the mapping process. A further issue is that of morphology: case markers, word-type determiners and the concatenation of lexical information in agglutinative languages to mention a few. These, of course, complicate matters further and require a considerable investment of resources to assure consistent inter-language mapping across lexical elements, metaphors, clichés and word translation granularity for subsequent robust analysis.

An example of such mapping is illustrated here between two of the more closely related languages (English – French) e.g. He saw her duck

French: Il a $\underbrace{\text{vu son}}$ canard<br>
He saw her *(possessive)* duck *(noun)*

Il l' $\underbrace{\text{a vue}}$ $\underbrace{\text{se baisser vivement}}$<br>
He her saw duck *(verb)*<br>
*(direct object)* *[lower herself quickly]*

This kind of sentence can easily be misinterpreted by a human, let alone a cross-linguistic or Machine Translation system.

It has been observed that there is a certain level of agreement between languages for such syntactic labelling. However, grammar is not indigenous to many languages such as Chinese, and the notion of parts-of-speech were most likely transplanted, and are a modified version of Western grammar, as originally devised by classical grammarians, such as Pannini and Thrax.

Nevertheless, irrespective of these often-transplanted notions of grammar, the information we all communicate consists of the same physics and basic necessary building blocks to describe our environment and thought processes. As a human race, our mechanism for language processing and generation – the Brain – functions, using the same physiology: areas of the Brain are dedicated to storing and accessing particular words classified by their parts-of-speech, such as the frontal lobe for verbs and the temporal lobe for verbs (Frederici et al, 2000). These neural constraints do not vary according to some linguistically geographical accident. So taking a theoretical stance akin to Chomsky, the *principles* should be detectable as long as the *parameters* are mapped accurately. This rationale provides the baseline for such design criteria and the notion of a 'universal' base-set across which annotation can operate.

## Results

Results to date have provided many significant findings for modelling 'universal' features of the surface structure of language. These currently range from the entropic evaluation of surface and dependency structure to the probabilistic modelling of cognition and visualising the bi-directional cohesion of linguistic objects, at varying distances of grammatical collocation (Elliott, J. 1999, 2000, 2001, 2002, 2003). Observations from these comparable and parallel corpora now comprise unsupervised mathematical models and algorithms that detect and identify linguistic objects without prior knowledge of their encoding strategies. In summary these comprise methods for detecting: language-like structure from all other structured and semi-structured phenomena; the internal structure of language; orthotactic constraints of phonetically based scripts; word boundaries; the identification of 'names', verbs and function words; identifying phrase-like chunks; determining the cohesive bonding of linguistic objects; the consistency of ratios between core parts-of-speech and the inappropriateness of using sentence structure to 'guide' unsupervised learning.

An interesting example of inter-language analysis at part-of-speech level was the comparison of Chinese (Piao, 2000a;b) and English (Johansson et al, 1986). These two seemingly incongruent languages were chosen as exemplar comparators to ascertain if the behaviour of core parts-of-speech, irrespective of their encoding strategies, display evidence of a generic cohesive template. This then provided an opportunity to compare the two very different orthographic systems of a Sino-Tibetan logographic script and Indo-European (West Teutonic) alphabetic language.

The Chinese corpus comprised a tag-set, which closely adheres to the 'core' part-of-speech, which classical grammarians originally devised and my meta-tag-set used to supplant the many diverse tag-sets discussed earlier. This assisted the immediacy of data analysis, without the need for extensive re-assignment of existing fine-grained annotation. All parts-of-speech pairs were then analysed for combinational constraint behaviour over a window of ten words, using a visualisation tool created for this purpose (Elliott, 2001).

Results using these metrics indicate that the interactive behaviour between their core parts-of-speech is in fact remarkably similar. These results therefore support the hypothesis, that the way we weave our respective ontological descriptions of the world around us, when communicating, are in fact constrained to general binding rules. The one area where differences are seen to occur is with immediate bonding of articles with some of the descriptive parts-of-speech: specifically, with verbs, adjectives and adverbs, when preceded by articles and where conjunctions and adverbs are immediate priors. It is believed the restricted set of articles used in the Chinese language is the root cause of this effect.

## Conclusions

A major hurdle, that has prevented the inclusion of many potentially interesting scripts, such as Tamil, is that freely available sources were only obtainable as scanned images, during the time period of acquisition. Samples with this format restriction, which effectively present language as an image, could not therefore be usefully analysed automatically, for the transparent investigation of their structure: a limitation that pervades the otherwise potentially useful Rosetta project (Rosetta Project, 2002). Other resources have either been unavailable, because of awaiting corpus development, or the current plethora of formats have posed too great a hurdle for the duration of this project. Currently, Unicode initiatives are underway to address this problem, but developments of such resources are slow and as yet too sparse for any practical benefits. However, it is hoped that in the not too distant future, a single Unicode-type format will emerge, providing the necessary platform to expedite computational analysis across all languages.

Nevertheless, as test samples do include many disparate languages, it is submitted that the resources gathered have been as comprehensive a test set as practically possible, given the time frame and fiscal limitations. It is also submitted that these samples present a credible representation of the human language condition to test subsequent hypotheses. The process continues.

A proposed format for future annotation is a bracketed, hierarchical tagset, comprising 4 potential word classification layers: e.g. Corpus [N] {com; sg} <NN1> “open info”]: (1) Generic base-set; (2) added information; (3) original annotation scheme: scheme inherited from donated annotated corpus; (4) open: additional information added by user. Morphological information is also an important element for segmenting grammatical tokens, especially when mapping the syntactic content across agglutinative and inflectional languages to isolating and mixed morphologies. It is therefore intended to incorporate morphological information by concatenating grammatical tags, where words contain more than one grammatical element, to expedite content transparency and prevent misinterpretation of ‘true’ lexico-grammatical comparisons.

Ultimately, the aim is to provide a single corpus that will expedite such inter-language analysis by incorporating languages from all the major language families, comprising all typologies, morphology and word order. Parts-of-speech classification and granularity will be consistent across all languages within this corpus and will conform more closely to the main parts-of-speech originally conceived by Dionysius Thrax than to the fine-grained systems used by the British National Corpus (BNC) and Lancaster-Oslo-Bergen (LOB) corpora. This will then enable cross-language analysis without the need for cross-mappings between differing annotation systems, or for writing/adapting software each time a different language or corpus is analysed. (Elliott, J & Elliott, D. 2003).

## References

Atwell E 1996 *Comparative Evaluation of Grammatical Annotation Models* in Sutcliffe R, Koch H-D, and McElligott A (editors), pages 25-46, Rodopi, Amsterdam.

Cloeren, Jan (1993): "Towards a cross-linguistic tagset", *Proceedings of the ACL Workshop on Very Large Corpora*, Ohio State University, Columbus (OH), 1993.

EAGLES (1996), WWW site for European Advisory Group on Language Engineering Standards, http://www.ilc.pi.cnr.it/EAGLES96/home.html

Elliott, J. (2002b) Detecting Languageness: in proceedings of *6th World Multi-Conference on Systemics, Cybernetics and Informatics (SCI2002)*, Orlando, Florida, USA: volume XI, pp 323-328.

Elliott, J & Atwell, E. (2001) Visualisation of long distance grammatical collocation patterns in language in: IV2001: *Proceedings of 5th International Conference on Information Visualisation*, pp.297-302. 2001. ISBN 0-7695-195-

Elliott, J. Atwell, E & Whyte, B. (2000). Language identification in unknown signals: in Proceeding of *COLING'2000*, pages 1021-1026, Association for Computational Linguistics (ACL) and Morgan Kaufmann Publishers, San Francisco. ISBN: 1-55860-717-X (2 volumes).

Elliott, J. & Atwell, E. (1999) Language in signals: the detection of generic species-independent intelligent language features in symbolic and oral communications in: Proceedings of the *50th International Astronautical Congress*, International Astronautical Federation: IAA-99-IAA.9.1.08

Elliott, J & Elliott, D. (2003) *The* Human Language Chorus Corpus in: proceedings of *CL2003, vol. 16 part 2 pp. 201-210* Archer D, Rayson P, Wilson A and McEnery T (eds.) Proceedings of CL2003: International Conference on Corpus Linguistics.

Elliott, D., Hartley, A. & Atwell, E. (2003) Rationale for a multilingual aligned corpus for machine translation evaluation In: Archer D, Rayson P, Wilson A and McEnery T (eds.) Proceedings of CL2003: International Conference on Corpus Linguistics.

Erjavec T, Ide N & Tufis D. 1998 Development And Assessment Of Common Lexical Specifications For Six Central And Eastern European Languages. *LREC'98*.

Friederici, A. D., Opitz, B. & von Cramon, D.Y. (2000) Segregating Semantic and Syntactic Aspects of Processing in the Human Brain in: *Cerebral Cortex (Journal)* 10, pp698-705.

Lesher G W, Moulton B J & Higginbotham D J.1999. Effects of ngram order and training text size on word prediction. *Proceedings of the RESNA '99 Annual Conference*, 52-54, Arlington, VA: RESNA Press.

Piao Scott Songlin, 2000. Sentence and Word Alignment between Chinese and English (PhD Thesis), Lancaster University. (b): Piao, Scott Songlin ,2000. Chinese Corpus adapted from CEPC Corpus, Sheffield University, Sheffield UK.

*Rosetta Project* (2002) [online] Available on World Wide Web: <http://www.rosettaproject.org/live>

# Exploiting Parallel Corpora for Monolingual Grammar Induction —A Pilot Study

**Jonas Kuhn**

The University of Texas at Austin
Department of Linguistics
Austin, TX 78712, USA
jonask@mail.utexas.edu

## Abstract

This paper presents results from a pilot study on ways of exploiting statistical word alignment for grammar induction. Following a scheme proposed in (Kuhn, 2004), we use GIZA++-word alignment from the multiple parallel texts in the Europarl corpus for the identification of string spans that cannot be constituents in one of the languages. This information is exploited in monolingual PCFG grammar induction for that language. Besides the aligned corpus, no other resources are required.

## 1. Introduction

There have been a number of recent studies exploiting parallel corpora in bootstrapping of monolingual analysis tools. In the "information projection" approach (e.g., (Yarowsky and Ngai, 2001)), statistical word alignment is applied to a parallel corpus of English and some other language $F$ for which no tagger/morphological analyzer/chunker etc. (henceforth simply: analysis tool) exists. A high-quality analysis tool is applied to the English text, and the statistical word alignment is used to project a (noisy) target annotation to the $F$ version of the text. Robust learning techniques are then applied to bootstrap an analysis tool for $F$, using the annotations projected with high confidence as the initial training data. (Confidence of both the English analysis tool and the statistical word alignment is taken into account.) The results that have been achieved by this method are very encouraging.

Will the information projection approach also work for less shallow analysis tools, in particular full syntactic parsers? An obvious issue is that one does not expect the phrase structure representation of English (as produced by state-of-the-art treebank parsers) to carry over to less configurational languages. Therefore, (Hwa et al., 2002) extract a more language-independent dependency structure from the English parse as the basis for projection to Chinese. From the resulting (noisy) dependency treebank, a dependency parser is trained using the techniques of (Collins, 1999). (Hwa et al., 2002) report that the noise in the projected treebank is still a major challenge, suggesting that a future research focus should be on the filtering of (parts of) unreliable trees and statistical word alignment models sensitive to the syntactic projection framework.

Our hypothesis is that the quality of the resulting parser/grammar for language $F$ can be significantly improved if the training method for the parser is changed to accomodate for training data which are in part unreliable. The experiments we report in this paper focus on a specific part of the problem: we replace standard treebank training with an Expectation-Maximization (EM) algorithm for PCFGs, augmented by weighting factors for the reliability of training data, following the approach of (Nigam et al., 2000), who apply it for EM training of a text classifier. The factors are only sensitive to the constituent/distituent status of each span of the string in $F$ (cp. (Klein and Manning, 2002)). The constituent/distituent status is derived from an aligned parallel corpus using the scheme of (Kuhn, 2004) (compare section 2.). We use the Europarl corpus (Koehn, 2002), and the statistical word alignment was performed with the GIZA++ toolkit (Al-Onaizan et al., 1999; Och and Ney, 2003).[1]

For the current experiments we assume no pre-existing parser for any of the languages, contrary to the information projection scenario. While better absolute results could be expected using one or more parsers for the languages involved, we think that it is highly informative to run a pilot study that isolates the effect of using crosslinguistic word order divergences as prior knowledge about the constituent structure of a language. This prior knowledge is exploited in an EM learning approach (section 3.). Not using a parser for some languages also makes it possible to compare various language pairs at the same level, and since we don't need English as the most reliable basis of projection, we can in particular run grammar induction experiments *for* English (section 4.), which facilitates evaluation against a treebank (section 5.).

## 2. Cross-language order divergences

The English-French example in figure 1 gives a simple illustration of the partial information about constituency that a word-aligned parallel corpus may provide. The en bloc reversal of subsequences of words provides strong evidence that, for instance, [ *moment the voting* ] or [ *aura lieu à ce* ] do *not* form constituents.

At first sight it appears as if there is also clear evidence for [ *at that moment* ] forming a constituent, since it fully covers a substring that appears in a different position in French. Similarly for [ *Le vote aura lieu* ]. However, from the distribution of contiguous substrings alone we cannot distinguish between the two types of situations sketched in (1) and (2): a string that is contiguous under projection, like $e_1e_2$ (1) may be a true constituent, but it may also be a non-constituent part of a larger constituent as in $L_1$ in (2).

[1] The software is available at http://www.isi.edu/˜och/GIZA++.html

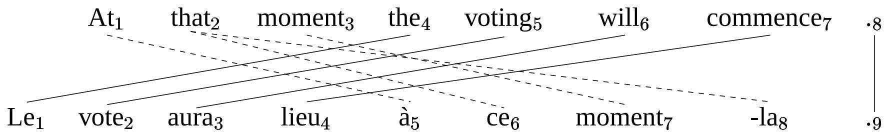
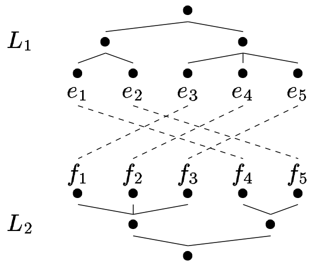
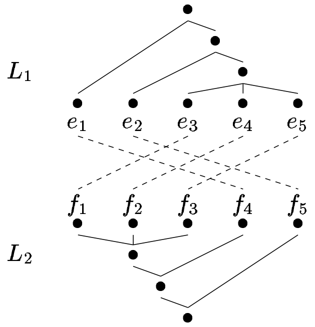

At1 that2 moment3 the4 voting5 will6 commence7 .8
Le1 vote2 aura3 lieu4 à5 ce6 moment7 -la8 .9

Figure 1: Alignment example

(1)



(2)



(Kuhn, 2004) provides a detailed discussion on the formal conditions for hypothesizing reliable non-consituency spans in a word-aligned corpus.

The core idea is to mark the boundary between contiguous word blocks (e.g., between $e_2$ and $e_3$ (1) or (2)). Then, spans of words crossing such boundaries without exhaustively covering one of the adjacent blocks are excluded from constituent status, i.e., we mark them as *distituents*.

**Mild divergences are best.** As should be clear, our scheme for detecting clues for non-constituency (i.e., information about distituents) relies on the occurrence of reorderings of constituents in translation. If two languages have the exact same structure (and no paraphrases whatsoever are used in translation), the approach does not gain any information from a parallel text. However, this situation does not occur realistically. If on the other hand, massive reordering occurs without preserving *any* contiguous subblocks, the approach cannot gain information either. The ideal situation is in the middleground, with a number of mid-sized blocks in most sentences.

## 3. EM grammar induction with weighting factors

The distituent identification scheme introduced in (Kuhn, 2004) and reviewed briefly in the previous section can be used to hypothesize a fairly reliable exclusion of constituency for many spans of strings from a parallel corpus. Besides a statistical word alignment, no further resources are required.

In order to make use of this scattered (non-)constituency information in grammar induction, a semi-supervised approach is needed that can fill in the (potentially large) areas for which no prior information is available. For the present experiments we decided to choose a conceptually simple such approach, with which we can build on substantial existing work in grammar induction: we construe the learning problem as PCFG induction, using the inside-outside algorithm, with the addition of weighting factors based on the (non-)constituency information. This use of weighting factors in EM learning follows the approach discussed in (Nigam et al., 2000).

For our pilot study, the conceptual simplicity and the availability of efficient implemented open-source systems of a PCFG induction approach outweighs the disadvantage of potentially poorer overall performance than one might expect from some other approaches.

The PCFG topology we use is a binary, entirely unrestricted X-bar-style grammar based on the Penn Treebank POS-tagset (expanded as in the TreeTagger by (Schmid, 1994)). All possible combinations of projections of POS-categories X and Y are included following the schemata in (3). This gives rise to 13,110 rules.

(3) a. XP → X

b. XP → XP YP

c. XP → YP XP

d. XP → YP X

e. XP → X YP

We tagged the English version of our training section from the Europarl corpus with the TreeTagger and used the strings of POS-tags as the training corpus for the inside-outside algorithm.[2]

We based our EM training algorithm on Mark Johnson's implementation of the inside-outside algorithm.[3] The initial parameters on the PCFG rules are set to be uniform. In the iterative induction process of parameter reestimation, the current rule parameters are used to compute the expectations of how often each rule occurred in the parses of the training corpus, and these expectations are used to adjust the rule parameters, so that the likelihood of the training data is increased. When the probablity of a given rule drops below a certain threshold, the rule is excluded from the grammar. The iteration is continued until the increase in likelihood of the training corpus is very small.

[2] Note that it is straightforward to apply our approach to a language for which no taggers are available if an unsupervised word clustering technique is applied first.

[3] http://cog.brown.edu/˜mj/

**Weight factors.** The inside-outside algorithm is a dynamic programming algorithm that uses a chart in order to compute the rule expectations for each sentence. We use the information obtained from the parallel corpus as discussed in section 2. (and more extensively in (Kuhn, 2004)) as prior information (in a Bayesian framework) to adjust the expectations that the inside-outside algorithm determines based on its current rule parameters. Note that this prior information is information about string spans of (non-)constituents – it does not tell us anything about the categories of the potential constituents affected. It is combined with the PCFG expectations as the chart is constructed. For each span in the chart, we get a weight factor that is multiplied with the parameter-based expectations. In the simplest model, we use the factor 0 for spans that are clear distituents, and factor 1 for all other spans; in other words, parses involving a distituent are cancelled out. We also used versions of the weight factors in which a number of levels is applied: distituents are assigned factor 0.01, likely distituents factor 0.1, neutral spans 1, and likely constituents factor 2.[4] The multi-level factor system turns out to outperform the simple distituent scheme.

## 4. Experiments

We applied GIZA++ (Al-Onaizan et al., 1999; Och and Ney, 2003) to word-align parts of the Europarl corpus (Koehn, 2002) for English and all other 10 languages. For the experiments we report in this paper, we only used the 1999 debates, with the language pairs of English combined with Finnish, French, German, Greek, Italian, Spanish, and Swedish.

For computing the weight factors we used a two-step process implemented in Perl, which first determines the location of boundaries between contiguous word blocks under cross-language word alignment. (5) shows the internal representation of the block structure for (4). `L` and `R` are used for the beginning and end of blocks, where it is unambiguous because there are no adjacent zero-fertility words (i.e., words for which the word alignment does not specify a correspondent). The notation `l` and `r` is used where zero-fertility word make the representation ambiguous. Words whose correspondents are in the same word order sequence are encoded as `*`, zero fertility words as `-`; `A` and `B` are used for the first block in a sentence instead of `L` and `R`, unless it arises from "relocation", which increases likelihood for constituent status (likewise for the last block: `Y` and `Z`).

(4)
```
la parole est à m. graefe zu baringdorf
pour motiver la demande
NULL ({ 3 4 11 }) mr ({ 5 }) graefe ({ 6
}) zu ({ 7 }) baringdorf ({ 8 }) has ({
}) the ({ 1 }) floor ({ 2 }) to ({ 9 })
explain ({ 10 }) this ({ }) request ({
12 })
```

(5)
```
[L**r-lRY*-*Z]
```

The second step for computing the weight factors creates a chart of all string spans over the given sentence and marks for each span whether it is a distituent, possible constituent or likely distituent, based on the location of boundary symbols. (For instance *zu Baringdorf has the* is marked as a distituent; *the floor* and *has the floor* are marked as likely constituents.) The tests are implemented as simple regular expressions. The chart of weight factors is represented as an array which is stored in the training corpus file along with the sentences. We combine the weight factors from various languages, since each of them may contribute distinct (non-)constituent information. The inside-outside algorithm reads in the weight factor array and uses it in the computation of expected rule counts.

We used the probability of the statistical word alignment as a confidence measure to filter out unreliable training sentences. Due to the conservative nature of the constituent/distituent information we extract from the alignment, the results indicate however that filtering is not necessary.

## 5. Evaluation

For evaluation, we used the PCFG resulting from the training described in section 4. in order to find the best parse for each test sentence according to the model. For this, we ran the trained grammar with the Viterbi algorithm[5] on parts of the Wall Street Journal (WSJ) section of the Penn Treebank and compared the predicted tree structure with the gold standard treebank annotation. The evaluation criteria we apply are unlabeled bracketing precision and recall (and crossing brackets). We follow an evaluation criterion that (Klein and Manning, 2002, footnote 3) discuss for the evaluation of a not fully supervised grammar induction approach based on a binary grammar topology: bracket multiplicity (i.e., non-branching projections) is collapsed into a single set of brackets (since what is relevant is the constituent structure that was induced).[6] For comparison, we provide baseline results that a uniform left-branching structure and a uniform right-branching structure (which encodes some non-trivial information about English syntax) would give rise to. As an upper boundary for the performance that a binary grammar can achieve on the WSJ, we present the scores for a minimal binarized extension of the gold-standard annotation.

The results we can report at this point are based on a comparatively small training set.[7] So, it may be too early for conclusive results. (An issue that arises with the small training set is that smoothing techniques would be required to avoid overtraining, but these tend to dominate the test application, so the effect of the parallel-corpus based information cannot be seen so clearly.) But we think that the

[4] The factor weights were chosen empirically; but it can be expected that in the future, a more systematic technique using a set of held-out data will lead to further improvements.

[5] We used the LoPar parser (Schmid, 2000) for this.

[6] Note that we removed null elements from the WSJ, but we left punctuation in place. We used the EVALB program for obtaining the measures, however we preprocessed the bracketings to reflect the criteria we discuss here.

[7] This is not due to scalability issues of the system; we expect to be able to run experiments on rather large training sets. Since no manual annotation is required, the available resources are practically indefinite.

| System | Unlab. Prec. | Unlab. Recall | $F_1$-Score | Crossing Brack. |
|---|---|---|---|---|
| Left-branching | 30.4 | 35.8 | 32.9 | 3.06 |
| Right-branching | 36.2 | 42.6 | 39.2 | 2.48 |
| Standard PCFG induction | 42.4 | 64.9 | 51.3 | 2.2 |
| PCFG trained with constituent/distituent weight factors from Europarl corpus | **47.8** | **72.1** | **57.5** | **1.7** |
| Upper limit | 66.08 | 100.0 | 79.6 | 0.0 |

Figure 2: Scores for test sentences up to length 10.

results are rather encouraging.

As the table in figure 2 shows, the PCFG we induced based on the parallel-text derived weight factors reaches 57.5 as the $F_1$-score of unlabeled precision and recall. We show the scores for an experiment without smoothing, trained on about 3,000 sentences. Since no smoothing was applied, the resulting coverage (with low-probability rules removed) on the test set is about 80%. It took 74 iterations of the inside-outside algorithm to train the weight-factor-trained grammar; the final version has 1005 rules.

For comparison we induced another PCFG based on the same X-bar topology without using the weight factor mechanism. This grammar ended up with 1145 rules after 115 iterations. The $F_1$-score is only 51.3 (while the coverage is the same as for the weight-factor-trained grammar).

## 6. Discussion

This paper presented a pilot study on ways of using parallel corpora as the only resource in the creation of a monolingual analysis tools. We believe that in order to induce high-quality tools based on statistical word alignment, the training approach for the target language tool has to be able to exploit islands of reliable information in a stream of potentially rather noisy data. We experimented with an initial idea to address this task, which is conceptually simple and can be implemented building on existing technology: using the notion of word blocks projected by word alignment as an indication for (mainly) impossible string spans. Applying this information in order to impose weighting factors on the EM algorithm for PCFG induction gives us a first, simple instance of the “island-exploiting” system we think is needed. More sophisticated models may make use some of the experience gathered in these experiments.

The conservative way in which cross-linguistic relations between phrase structure is exploited has the advantage that we don’t have to make unwarranted assumptions about direct correspondences among the majority of constituent spans, or even direct correspondences of phrasal categories. The technique is particularly well-suited for the exploitation of parallel corpora involving multiple languages like the Europarl corpus. Note that nothing in our methodology made any language particular assumptions; future research has to show whether there are language pairs that are particularly effective, but in general the technique should be applicable for whatever parallel corpus is at hand.

A number of studies are related to the work we presented, most specifically work on parallel-text based “information projection” for parsing (Hwa et al., 2002), but also grammar induction work based on constituent/distituent information (Klein and Manning, 2002) and (language-internal) alignment-based learning (van Zaanen, 2000). However to our knowledge the specific way of bringing these aspects together which we proposed in (Kuhn, 2004) is new.

## 7. References

Al-Onaizan, Yaser, Jan Curin, Michael Jahr, Kevin Knight, John Lafferty, Dan Melamed, Franz-Josef Och, David Purdy, Noah A. Smith, and David Yarowsky, 1999. Statistical machine translation. Final report, JHU Workshop.

Collins, M., 1999. A statistical parser for Czech. In *Proceedings of ACL*.

Hwa, R., P. Resnik, and A. Weinberg, 2002. Breaking the resource bottleneck for multilingual parsing. In *Proceedings of LREC*.

Klein, D. and C. Manning, 2002. A generative constituent-context model for improved grammar induction. In *Proceedings of ACL*.

Koehn, P., 2002. Europarl: A multilingual corpus for evaluation of machine translation. Ms., University of Southern California.

Kuhn, Jonas, 2004. Experiments in parallel-text based grammar induction. Ms., The University of Texas at Austin.

Nigam, K., A. K. McCallum, S. Thrun, and T. Mitchell, 2000. Text classification from labeled and unlabeled documents using EM. *Machine Learning*, 39(2/3):103–134.

Och, F. J. and H. Ney, 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Schmid, Helmut, 1994. Probabilistic part-of-speech tagging using decision trees. In *International Conference on New Methods in Language Processing*. Manchester, UK.

Schmid, Helmut, 2000. Lopar: Design and implementation. Arbeitspapiere des Sonderforschungsbereiches 340, No. 149, IMS Stuttgart.

van Zaanen, M., 2000. ABL: Alignment-based learning. In *COLING 2000 - Proceedings of the 18th International Conference on Computational Linguistics*.

Yarowsky, D. and G. Ngai, 2001. Inducing multilingual POS taggers and NP bracketers via robust projection across aligned corpora. In *Proceedings of NAACL*.

# A Multilingual Parallel Parsed Corpus as Gold Standard for Grammatical Inference Evaluation

**Menno van Zaanen†, Andrew Roberts*, Eric Atwell***

†Tilburg University
Postbus 90153, 5000LE, Tilburg, the Netherlands
mvzaanen@uvt.nl

*University of Leeds
Woodhouse Lane, LS2 9JT, Leeds, U.K.
{andyr,eric}@comp.leeds.ac.uk

**Abstract**

In this article we investigate how (computational) grammar inference systems are evaluated and how the evaluation procedure can be improved. First, we describe the currently used evaluation methods and look at the advantages and disadvantages of each method. The main problems of the methods are: the dependency on language experts, the influence of the annotation scheme of language data, and the language dependency of the evaluation. We then propose a new method that will allow for an evaluation independently of language and annotation scheme. This method requires (syntactically) structured corpora in multiple languages to test for language independency of the grammatical inference system and corpora structured using different annotation schemes to diminish the influence the annotation has on the evaluation.

## 1. Introduction

Grammar inference (GI) is focused on the task of inferring or learning grammatical descriptions of a language from a corpus of language examples. Research on grammar inference focuses on showing which (classes of) grammars can be learned and how this can be done. This includes formal learnability research, which identifies, for example, classes of grammars that can be learned within polynomial time and gives mathematical proofs for this. Additionally, linguists (including, among others, formal linguists, psycholinguists, cognitive linguists and computational linguists) concentrate more on natural languages. Discussions and cooperations between the different groups of researchers has led to interesting results (de la Higuera et al., 2003).

On the one hand, formal grammar inference research provides us with solid proof of the learnability of classes of grammars, which might not have any linguistic relevance. On the other hand, researchers from other fields have a harder time actually proving or even showing that a system or approach might actually learn a certain type of language.

In this article we will take a look at the evaluation methods that are available for investigating the performance of grammar inference systems. We will describe the approaches currently in use and discuss their advantages and disadvantages. Based on this, we propose a new evaluation approach. This approach reduces the influence of a specific language or annotation scheme by testing on several different languages and on texts annotated with different schemes.

## 2. Current Evaluation Approaches

Several descriptions of grammar inference systems together with some evaluation have been published (see, for example, (Adriaans, 1992; Déjean, 2000; Grünwald, 1994; Nakamura and Ishiwata, 2000; Stolcke and Omohundro, 1994; Wolff, 1980)). These and other GI systems have been evaluated using different methods. The evaluation methods used can be divided into three large groups (van Zaanen, 2002).[1] These groups are described below.

### 2.1. Looks-Good-to-Me

The GI system is applied to unstructured data. This data can be, for example, linguistic data or it can be generated by a grammar. The output produced by the system is then checked manually for interesting aspects.

This approach as two main advantages. Firstly, only unstructured data is needed. This makes it easy to apply the system on different languages. Secondly, the evaluation can focus on certain specific syntactic constructions. Not only can the output of the GI system be easily searched for a given construction, the input can be tailored to learning it as well.

However, this approach will only provide a useful means of reference if it is done by an independent expert comparing outputs of rival systems. In practice most GI developers have applied *looks-good-to-me* evaluation to their own systems, rather than perform objectively quantifiable comparisons.

Human evaluation of output is accepted standard practice in Machine Translation evaluation, e.g. (Elliott et al., 2003), where a range of translations may be equally valid. However, this evaluation involves assessments by independent judges, who give an expert assessment of quality of output.

[1] A fourth method, which we call *language membership*, is being used in GI competitions as Abbadingo, Gowachin, and Omphalos. The learning system must indicate whether a test sentence is a member of the language or not. The correct answers are counted. We will not consider this approach any further, since no explicit grammatical properties are measured.

### 2.2. Rebuilding Known Grammars

In this evaluation approach, one or more "toy" grammars are selected beforehand. These grammars are used to generate data, which again is used as input for the GI system. The output (i.e. the grammar or the structured version of the input) is then compared to the original data.

The grammars can be chosen with known properties. These properties can, for example, reflect specific syntactic constructions or be more global, such as context-freeness. Additionally, the evaluation itself can be done automatically, without the need for a language expert.

Because the grammars are chosen or created by hand, this may work for small, artificial languages, but does not scale up to wide-coverage Natural Language grammars. A related problem is that the specific grammars might be tailored to the specific GI systems.[2] On a wider scale, different GI systems aim for different types of grammars or language models, making this an unfair test of systems not geared to generate, for example, small context-free grammars.

The generation part of this approach also poses interesting problems. One has to decide what probability distribution should be assigned to the grammar rules. This decision might influence the learning process. Additionally, all grammar rules should be applied at least once (otherwise the grammar rule cannot be learned) and restrictions may be necessary to limit the sentence length. With respect to emulating natural languages, this comes down to deciding on a language model.

Another problem is that comparing grammars in general is hard. With infinite languages, not all sentences in the language can be compared, which results in a need to compare the generative power of the grammars themselves, which in turn can be quite hard in practice. Note that when the goal is to learn the tree language, this problem is less hard (since the grammar rules themselves can be compared), but not necessarily trivial.

### 2.3. Compare Against Treebank

The final approach starts out with an annotated treebank which is selected as a "gold standard". The GI system then infers or rebuilds the structure of the plain sentences extracted from the annotated treebank. The learned, structured sentences are compared against the trees in the original treebank, which measures how well the GI system can find the original structure.

The gold standard is a treebank, that may contain natural language data or tree structures generated by a grammar. This allows for flexibility in the data or grammars used. Different natural languages or data from specific domains can be tested.

All GI systems can be adapted to generate structured versions of the input sentences, unlike with the *rebuilding known grammars* approach, where the output of the GI system needs to be a grammar. When a system generates a grammar, the sentences can be parsed, which still results in a structured version of the input sentences. This makes comparing trees a valid option for all systems.

The main problem with this approach is that structured corpora are needed. This may not be a problem when evaluating known grammars, but in the case of natural languages, the underlying grammar is not known. This means that natural language treebanks are needed, which need to be build by hand (or semi-automatically).

## 3. Problems with Current Approaches

Although the current approaches provide information on the effectiveness of GI systems and even some standard grammars and test treebanks (Clark, 2001; Klein and Manning, 2002; van Zaanen and Adriaans, 2001) arise, each approach has some problems as described above.

From the existing approaches, the *compare against treebank* approach has most potential. With the *looks-good-to-me* approach, objective evaluation is difficult (especially since often blind evaluation is not performed). The *rebuilding known grammars* approach is too limited because the underlying grammar of natural language data is not currently known. This restricts the application to relatively small artificial grammars.

One of the aims of GI is to achieve generic learning, across a wide range of source language data. Focusing on a specific treebank for comparative evaluations may result in over-training and/or a bias in favor of GI systems developed for a comparable language. Another bold aim of GI is the discovery of new concepts in grammar, or at least valid alternatives to "standard theory". Evaluation by comparison with "received wisdom" will not favor innovation.

Another problem is that, doing evaluation using treebanks is not as simple as one might expect from the discussion above. One needs to decide on several parameters. The metrics that will be used to compute similarity between trees have a huge impact on the final results. Currently, the PARSEVAL metrics[3] are often used (Black et al., 1991), but other measures are of course possible.

Furthermore, we have to keep in mind that to investigate and compare the effectiveness of the wide range of GI systems properly, a robust evaluation method is needed. GI systems are meant to be used on different (natural) languages (and domains), so the evaluation method needs at least to be robust with respect to language. Additionally, since we are considering structure, the annotation of this structure should not be a major factor in the evaluation results. Robustness with respect to annotation should, thus, also be taken into account.

## 4. Evaluation Using a Parallel Corpus

We propose the use of a parallel-parsed corpus as the new gold standard, as it offers a fairer approach to evaluation, and does not promote over-training as easily (Roberts and Atwell, 2003).

The idea of using a gold standard in itself is not new. There have been similar gold standard approaches to evaluation of parsers (Black et al., 1991), Machine Translation

[2] In practice, there are some grammars that are considered "standard" test grammars (Cook et al., 1976; Hopcroft et al., 2001; Nakamura and Matsumoto, 2002; Stolcke, 2003).

[3] The PARSEVAL metrics can compare simple phrase-structure bracket overlap between GI output and Gold Standard phrase-structure parses.

systems (Elliott et al., 2003), and other NLP systems. However, here we try to solve many of the problems of the existing approaches.

### 4.1. Different Languages

Non-English language resources are comparatively rare compared to English ones. We are not only referring to corpora, but to language tools, too. If we are to provide a multi-parsed corpus for each language selected, there must exist a variety of taggers and parsers to achieve this aim.

Fortunately, there are many sizable treebank creation projects under way: Dutch ALPINO treebank (van der Beek et al., 2001), Bulgarian BulTreebank (Osenova and Simov, 2003), UPenn Chinese treebank (Xue et al., 2004), UAM Spanish Treebank (Moreno and López, 1999), NEGRA German treebank (Skut et al., 1997), and many more. These would need to be expanded for our purposes to include parallel parses.

Another aspect to take into consideration is to select a broad range of languages, spanning a variety of language families. This should result in a well balanced corpus. For example, we will obviously have English as one of our candidate languages, which comes from the Germanic branch of the Indo-European family. It would therefore make sense not to include (much data of) another language from this branch such as Dutch or Afrikaans until other language families are represented for better coverage, e.g., Russian from the Slavic branch of the Indo-European family, Arabic from the Semitic branch of the Afro-Asiatic family, Japanese from the Altaic family, etc.

### 4.2. Different Domains

Related to the selection of data from several languages (and language families) is the selection of data from different domains. Current *compare against treebank* evaluations within the field of GI take the ATIS treebank (taken from the Penn Treebank) as gold standard.[4] The problem with this is that the treebank is taken from the limited domain of air travel. A fair evaluation should be done on a treebank taken from a much larger domain or a combination of domains.

### 4.3. Different Annotation Schemes

One of the largest and most complex tasks of compiling a parallel corpus (by cherry-picking the most appropriate existing treebanks) will be dealing with the large variety of annotation schemes. There is no standard tagset that is commonly adopted by corpus builders, and so each individual corpus is likely to have its own individual annotation scheme.[5]

For our corpus to be adopted by the GI community for evaluation purposes, these inner variances must be transparent, as few developers would have the patience, or resources, to create their own interfaces for each of the various treebanks within the evaluation corpus. We must ensure, that—at least from the end-users' point of view—there is only a single annotation scheme to deal with.

To achieve this, we must first decide upon the "best" annotation scheme for our entire corpus. For the purposes of grammar induction evaluation, a large and highly specific tagset is not necessary. Next, we must work upon a system for mapping original treebank annotation into the "GI evaluation" annotation. Such an approach has already been successfully applied on a small scale within the AMALGAM project (Atwell et al., 2000).

## 5. Future Work

Clearly, the construction of this corpus is still in its early design stages. It has the potential to be an enormous project in terms of resources required. We can use our current parallel-parsed treebank as a seed for future development. Perfecting the design and required skills for compiling a single language, large-scale, multi-treebank is an ongoing process, which entails selecting suitable candidate treebanks, parsers and an annotation scheme. Once this multi-treebank is complete, the next stage will be to apply the same principles for additional languages.

With respect to the practical evaluation using a multilingual, parallel corpus, one would like to allow easy access to this data. Preferably, an (operating system independent) software suite should be developed that applies the GI system to the plain sentences of the treebank and compares the output against the structures found in the treebank.

It may prove difficult to automatically compare GI output against Gold Standard trees in all cases, so a fall-back may be to use human "looks-good-to-me" assessment; but in this case the judges are constrained to assess how close the GI output is to the example parse, as in Machine Translation evaluation experiments (Elliott et al., 2003).

The suite should be flexible with respect to different languages, domain specific sub-corpora, annotation schemes and evaluation metrics. This flexibility is needed, for example, when a GI system is computationally intensive and can only be applied to a limited amount of data.

## 6. Conclusion

In this article, we have investigated the current evaluation approaches that are applied to grammatical inference systems. The approaches can be classified in three groups: *looks-good-to-me*, *rebuilding known grammars*, and *compare against treebank*. Each of these approaches have some advantages, but also disadvantages.

We propose to use a multi-lingual, parallel-parsed corpus as the basis of the evaluation. By applying the system to multiple languages within different domains, the language and domain independency of the GI system is evaluated, while the evaluation against the different parses of the sentences diminishes the impact of the used annotation scheme. In other words, it extends the *compare against treebank* approach in that it also measures the amount of language and annotation scheme independency of the GI system.

[4] Recently, people have started to use the WSJ treebank for evaluation, but this does not entirely solve the problem (Klein and Manning, 2002; van Zaanen, 2002).

[5] Different languages may, for example, have the need for different part-of-speech tags. Design issues like this influence the annotation of the corpus. Additionally, a treebank may be structured with respect to different syntactic phenomena.

## 7. References

Adriaans, Pieter Willem, 1992. *Language Learning from a Categorial Perspective*. Ph.D. thesis, University of Amsterdam, Amsterdam, the Netherlands.

Archer, D., P. Rayson, A. Wilson, and T. McEnery (eds.), 2003. *Proceedings of the Corpus Linguistics 2003 conference; Lancaster, UK*.

Atwell, E., G. Demetriou, J. Hughes, A. Schiffrin, C. Souter, and S. Wilcock, 2000. A comparative evaluation of modern english corpus grammatical annotation schemes. *ICAME Journal, International Computer Archive of Modern and medieval English*, 24:7–23.

Black, E., S. Abney, D. Flickinger, C. Gdaniec, R. Grishman, P. Harrison, D. Hindle, R. Ingria, F. Jelinek, J. Klavans, M. Liberman, M. Marcus, S. Roukos, B. Santorini, and T. Strzalkowski, 1991. A procedure for quantitatively comparing the syntactic coverage of English grammars. In *Proceedings of a Workshop—Speech and Natural Language*.

Clark, Alexander, 2001. Unsupervised induction of stochastic context-free grammars using distributional clustering. In *Proceedings of the Workshop on Computational Natural Language Learning held at the 39th Annual Meeting of the ACL and the 10th Meeting of the EACL; Toulouse, France*.

Cook, Craig M., Azriel Rosenfeld, and Alan R. Aronson, 1976. Grammatical inference by hill climbing. *Informational Sciences*, 10:59–80.

de la Higuera, Colin, Pieter Adriaans, Menno van Zaanen, and Jose Oncina (eds.), 2003. *Proceedings of the Workshop and Tutorial on Learning Context-Free Grammars held at the 14th European Conference on Machine Learning (ECML) and the 7th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD); Dubrovnik, Croatia*.

Déjean, Hervé, 2000. ALLiS: a symbolic learning system for natural language learning. In Claire Cardie, Walter Daelemans, Claire Nédellec, and Erik Tjong Kim Sang (eds.), *Proceedings of the Fourth Conference on Computational Natural Language Learning and of the Second Learning Language in Logic Workshop; Lisbon, Portugal*. Held in cooperation with ICGI-2000.

Elliott, Debbie, Anthony Hartley, and Eric Atwell, 2003. Rationale for a multilingual aligned corpus for machine translation evaluation. In (Archer et al., 2003), pages 191–200.

Grünwald, Peter, 1994. A minimum description length approach to grammar inference. In G. Scheler, S. Wernter, and E. Riloff (eds.), *Connectionist, Statistical and Symbolic Approaches to Learning for Natural Language*. Berlin Heidelberg, Germany: Springer-Verlag, pages 203–216.

Hopcroft, J.E., R. Motwani, and J.D. Ullman, 2001. *Introduction to automata theory, languages, and computation*. Reading:MA, USA: Addison-Wesley Publishing Company.

Klein, Dan and Christopher D. Manning, 2002. A generative constituent-context model for improved grammar induction. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL); Pennsylvania:PA, USA*. Association for Computational Linguistics (ACL). Yes.

Moreno, A. and S. López, 1999. Developing a spanish tree bank. In *Proceedings of the ATALA Treebank Workshop (Journés ATALA, Corpus annotés pour la syntaxe); Paris, France*.

Nakamura, K. and M. Matsumoto, 2002. Incremental learning of context-free grammars. In Pieter Adriaans, Henning Fernau, and Menno van Zaanen (eds.), *Grammatical Inference: Algorithms and Applications (ICGI); Amsterdam, the Netherlands*, volume 2482 of *Lecture Notes in AI*. Berlin Heidelberg, Germany: Springer-Verlag.

Nakamura, Katsuhiko and Takashi Ishiwata, 2000. Synthesizing context free grammars from sample strings based on inductive CYK algorithm. In Arlindo L. Oliveira (ed.), *Grammatical Inference: Algorithms and Applications (ICGI); Lisbon, Portugal*. Berlin Heidelberg, Germany: Springer-Verlag.

Osenova, Petya and Kiril Simov, 2003. The bulgarian hpsg treebank: Specialization of the annotation scheme. In *Proceedings of The Second Workshop on Treebanks and Linguistic Theories; Växjö, Sweden*.

Roberts, Andrew and Eric Atwell, 2003. The use of corpora for automatic evaluation of grammar inference systems. In (Archer et al., 2003), pages 657–661.

Skut, Wojciech, Brigitte Krenn, Thorsten Brants, and Hans. Uszkoreit, 1997. An annotation scheme for free word order languages. In *Proceedings of the Fifth Conference on Applied Natural Language Processing (ANLP-97); Washington:DC, USA*.

Stolcke, A., 2003. Boogie. `ftp://ftp.icsi.berkeley.edu/pub/ai/stolcke/software/boogie.shar.Z`.

Stolcke, Andreas and Stephen Omohundro, 1994. Inducing probabilistic grammars by bayesian model merging. In *Proceedings of the Second International Conference on Grammar Inference and Applications; Alicante, Spain*.

van der Beek, Leonoor, Gosse Bouma, Robert Malouf, and Gertjan van Noord, 2001. The Alpino Dependency Treebank. In Mariët Theune, Anton Nijholt, and Hendri Hondorp (eds.), *Computational Linguistics in the Netherlands 2001; Enschede, the Netherlands*. Amsterdam, the Netherlands: Rodopi.

van Zaanen, Menno, 2002. *Bootstrapping Structure into Language: Alignment-Based Learning*. Ph.D. thesis, University of Leeds, Leeds, UK.

van Zaanen, Menno and Pieter Adriaans, 2001. Alignment-Based Learning versus EMILE: A comparison. In *Proceedings of the Belgian-Dutch Conference on Artificial Intelligence (BNAIC); Amsterdam, the Netherlands*.

Wolff, J. Gerard, 1980. Language acquisition and the discovery of phrase structure. *Language and Speech*, 23(3):255–269.

Xue, Nianwen, Fei Xia, Fu-Dong Chiou, and Martha Palmer, 2004. The penn chinese treebank: Phrase structure annotation of a large corpus. *Natural Language Engineering*, 10(4):1–30.

# Author Index

**A**
Asunmaa Marita 1
Atwell Eric 18, 58
Aziz Normaziah Abdul 22

**B**
Bentivogli Luisa 38

**C**
Castel Núria 26
Cheung Percy 30

**D**
Dong Mingchui 42

**E**
Elliot Debbie 18
Elliot John 50

**F**
Fung Pascale 30

**H**
Hartley Anthony 18
Hu Dong Cheng 42

**K**
Kingsbury Paul 34
Kranias Lambros 1
Kuhn Jonas 54

**L**
Lambert Patrik 26

**M**
Magnusdottir Gudrun 1
Maia Belinda 14
Mao Yu Hang 42

**P**
Palmer Martha 34
Papadopoulos George 1
Pianta Emanuele 38

**R**
Rahman Suhaimi Ab 22
Ranieri Marcello 38
Roberts Andrew 58

**S**
Samiotou Anna 1
Santos Diana 14
Sarmento Luís 14
Schrader Bettina 46

**T**
Tang Chiwai 42
Tate Calandra 10
Thurmair Gregor 5

**V**
Vanni Michelle 10
van Zaanen Menno 58
Voss Clare 10

**W**
Wong Fai 42

**X**
Xue Nianwen 34