MATMT2008 workshop

Mixing Approaches to Machine Translation

Donostia-San Sebastian

February 14th, 2008

proceedings

editors:

Iñaki Alegria

Lluís Màrquez

Kepa Sarasola

Ixa group, University of the Basque Country

Elhuyar Fundazioa

ixa

eman ta zabal zazu

Universidad del País Vasco
Euskal Herriko Unibertsitatea
The University of the Basque Country

ELHUYAR fundazioa

Acknowledgements:

We gratefully acknowledge the financial support from the OpenMT project and from the Government of the Basque Country. The OpenMT project is funded by the Spanish Ministry of Education and Science (OpenMT: Open Source Machine Translation using hybrid methods, TIN2006-15307-C03-01) and the Local Government of the Basque Country (AnHITZ 2006: Language Technologies for Multilingual Interaction in Intelligent Environments, IE06-185).

Published by the University of the Basque Country

Cover design and picture: Xabier Artola

Printed by Fotocopias Zorroaga

MATMT2008 workshop

Mixing Approaches to Machine Translation

Donostia-San Sebastian

February 14th, 2008

Programme committee

Iñaki Alegria (University of the Basque Country, Donostia)
Kutz Arrieta (Vicomtech, Donostia)
Núria Castell (Technical University of Catalonia, Barcelona)
Arantza Diaz de Ilarraza (University of the Basque Country, Donostia)
David Farwell (Technical University of Catalonia, Barcelona)
Mikel Forcada (University of Alacant, Alicante)
Philipp Koehn (University Of Edinburgh, UK)
Lluís Màrquez (Technical University of Catalonia, Barcelona) (Co-chair)
Hermann Ney (Rheinisch-Westfälische Technische Hochschule, Aachen)
Kepa Sarasola (University of the Basque Country, Donostia) (Co-chair)

Local organization

IXA Group, University of the Basque Country

Iñaki Alegria, Arantza Casillas, Arantza Díaz de Ilarraza, Jon Igartua, Gorka Labaka, Mikel Lersundi, and Kepa Sarasola.

Elhuyar Fundazioa

Antton Gurrutxaga, Igor Leturia, and Xabier Saralegi.

Introduction

Welcome to the MATMT2008 Workshop on "*Mixing Approaches to Machine Translation*", held at the University of the Basque Country on February 14, 2008.

"Mixing Approaches To Machine Translation" aims at promoting practical hybrid approaches to MT, combining resources and algorithms coming from rule-based, example-based or statistical approaches, and helping to disseminate the work developed in OpenMT project.

The boundaries between the three principal approaches to MT (rule-based, example-based and statistical) are becoming narrower:

- Phrase based SMT models are incorporating morphology, syntax and semantics into their systems.
- Rule based systems are using parallel corpora to enrich their lexicons and grammars, and to create new disambiguation methods.
- Previous ASR/ALT projects have shown that in a MT system benefits can be realized by a simple combination of different MT approaches in a Rover architecture.

Data-driven Machine Translation (example-based or statistical) is nowadays the most prevalent trend in Machine Translation research. Translation results obtained with this approach have now reached a high level of accuracy, especially when the target language is English. But these data-driven MT systems base their knowledge on aligned bilingual corpora, and the accuracy of their output depends heavily on the quality and the size of these corpora. Large and reliable bilingual corpora are unavailable for many language pairs.

Three invited speakers will show us a wide perspective of the latest developments in machine translation:

- Philipp Koehn (University Of Edinburgh, UK)
  Moses: Moving Open Source MT towards Linguistically Richer Models
- Marcello Federico (Fundazione Bruno Kessler, Trento, Italy)
  Recent Advances in Spoken Language Translation
- Andy Way (Dublin City University, Ireland).
  Combining Approaches to Machine Translation: the DCU Experience

Twelve papers were submitted to our call for papers in November 2007, they were from Europe, Asia and America. The program committee selected eight of them for presentation at the conference.

Those three talks and eight presentations will provide us some points for the final discussion and conclusion that will be chaired by David Farwell (Technical University of Catalonia)

We wish you a pleasant and inspiring day.

Iñaki Alegria Lluís Màrquez Kepa Sarasola

About OpenMT project

http://ixa.si.ehu.es/openmt

One of the aims of the MATMT2008 Workshop on "*Mixing Approaches to Machine Translation*" concerns the dissemination of the work developed within the OpenMT research project. OpenMT is a project for machine translation that is been partially funded by the Spanish Ministry of Education and Science (OpenMT: Open Source Machine Translation using hybrid methods, TIN2006-15307-C03-01) and the Local Government of the Basque Country (AnHITZ 2006: Language Technologies for Multilingual Interaction in Intelligent Environments, IE06-185).

The main goal of OpenMT project is the development of open source machine translation architectures based on hybrid models and advanced semantic processors. These architectures will be open-source systems combining the three main Machine Translation frameworks into hybrid systems. Defined architectures and results of the project will be open source, so it will allow rapid development and adaptation of new advanced Machine Translation systems for other languages. We are testing the functionality of this system with different languages: English, Spanish, Catalan, and Basque. Corpora are easily available for English and Spanish, but not so for the remaining languages. While the structure of some of those languages is very similar (Catalan and Spanish), others are very different (English and Basque). Basque is an agglutinative language with a very rich morphology, unlike English, Catalan and Spanish.

The main innovative points of the OpenMT project are:

- The design of hybrid systems combining traditional linguistic rules, example-based methods, and statistical methods.
- Open Source Initiative
- The use of advanced syntactic and semantic processing in MT

SELECTED PAPERS

INVITED TALKS

A Method of Automatically Evaluating Machine Translations Using a Word-alignment-based Classifier

Katsunori Kotani
Kansai Gaidai University/NICT
3-5 Hikaridai, Seika-cho,
Soraku-gun, Kyoto, Japan
kat@khn.nict.go.jp

Takehiko Yoshimi
Ryukoku University/NICT
3-5 Hikaridai, Seika-cho,
Soraku-gun, Kyoto, Japan

Takeshi Kutsumi
Sharp Corporation
492 Minosho-cho, Yamato-koriyama, Nara, Japan

Ichiko Sata
Sharp Corporation
492 Minosho-cho, Yamatokoriyama, Nara,
Japan

Hitoshi Isahara
National Institute of Information and Communications Technology (NICT)
3-5 Hikaridai, Seika-cho, Soraku-gun, Kyoto,
Japan

Abstract

Constructing a classifier that distinguishes machine translations from human translations is a promising approach to automatically evaluating machine-translated sentences. We developed a classifier with this approach that distinguishes translations based on word-alignment distributions between source sentences and human/machine translations. We used Support Vector Machines as machine-learning algorithms for this classifier. Our experimental results revealed that our method of evaluation had a weak correlation with human evaluations. We further found that our method outperformed well-known automatic-evaluation metrics with respect to correlation with the manual evaluation, and that it could identify the qualitative characteristics of machine translations, which greatly help improve their quality.

1 Introduction

Previous research has proposed various automatic methods of evaluating machine-generated translations (MTs). Some methods have examined the similarity of MTs to human-generated translations (HTs), i.e., BLEU (Papineni et al. 2001), NIST (Doddington 2002), METEOR (Banerjee & Alon 2005), Kulesza & Shieber (2004), Paul et al. (2007), and Blatz et al. (2004). These methods would be rather expensive due to the need to prepare multiple-reference HTs for evaluation. To resolve this problem, Corston-Oliver et al. (2001) and Gammon et al. (2005) proposed methods of evaluation, which did not employ multiple reference HTs in evaluating MTs.[1] Instead of evaluating MTs by comparing them with HTs, evaluation was carried out with a machine-learning algorithm that classified MTs either into "good" or "bad" translations. A "good" translation is a translation that is indistinguishable from HTs, whereas a "bad" translation is a translation that is judged to be an MT. Although this method of classification might require reference HTs to construct a classifier as training data, it does not need any reference HTs for evaluation. Hence, once a classifier is constructed, this method can be applied to any translations without reference HTs. This is an advantage of classifier-based evaluation methods.

This new method also reveals what sorts of errors are involved in MTs, while others such as BLEU (Papineni et al. 2001) cannot, as Corston-Oliver et al. (2001) suggested. The primary goal of BLEU (Papineni et al. 2001) was to determine the superiority of translation systems, and hence, the method outputs numerical values in terms of BLEU scores. When one tries to improve a translation system, it is necessary to identify the problems with it. On these grounds, we surmised that a clas-

[1] Albrecht & Hwa (2007) also proposed MT evaluation metrics without using reference HTs. Their method employed the regression-trained metric.

sifier-based scheme would be a promising approach to evaluating MTs.

Although source sentences need to be referred to in order to evaluate the adequacy of MTs, these previous methods have only examined the linguistic properties of MTs but not those of source sentences. Hence, they have focused on the fluency of translation but not on the adequacy of translation. Adequacy is defined as to what extent a translated sentence conveys the meaning of the original sentence. Fluency is defined as the well-formedness of a translated sentence, which can be evaluated independently of adequacy.

This paper discusses our examination of a classifier, which can evaluate MTs from both viewpoints of fluency and adequacy. In evaluating translations from English to Japanese, for instance, not only the translation fluency but also its adequacy should be carefully assessed, because translations between these languages involve greater linguistic problems than those between European languages, e.g., English and French. European languages belong to the same language class, whereas English and Japanese do not. Thus, English and Japanese vary greatly with respect to various linguistic properties such as anaphoric systems (see Section 3.3).

This linguistic divergence makes evaluations of adequacy significant for MTs of English into Japanese. We propose employing a classification feature that reveals the linguistic correspondences between source sentences and translations to evaluate adequacy with the classification method. Incidentally, unlike reference translations, source sentences are necessary to obtain MTs. Thus, we constructed a classifier that would distinguish translations based on word-alignment distributions between source sentences and translations, assuming that the word-alignment distributions exhibited linguistic correspondences between these source sentences and translations. We then assessed our method by comparing its evaluation results with those of human evaluations.

2 Method of Machine Learning

Our method uses Support Vector Machines (SVMs), which are well known learning algorithms that have high degrees of generalization. We used SVMs to build a classifier based on word-alignment distributions as machine-learning features.

Our method employs parallel corpora to construct the classifier and requires neither manually labeled training examples (unlike Albrecht) nor multiple reference translations to evaluate new sentences. Due to these properties, our method should be a relatively inexpensive but effective automatic evaluation metric.

2.1 Evaluation Metric Obtained by SVMs

SVMs are learning algorithms based on maximum margin strategy (Vapnik 1998). We train an SVM classifier by taking HTs as positive training examples and MTs as negative. Consequently, the SVMs produce a hyperplane that separates the examples. As Kulesza & Shieber (2004) noted, the distance between the separating hyperplane and a test example can serve as an evaluation score. Based on this idea, our classifier not only distinguishes the MTs from HTs but also evaluates the MTs with this metric.

2.2 Features

As we noted in Section 1, word-alignment distribution should constitute classification features examining translation adequacy. We further presumed that word-alignment distribution could also be used to examine translation fluency.

Good, natural translations differ from poor, unnatural translations such as word-for-word translations, because superior translations involve various translation techniques. For instance, there is a technique for translating the English nominal modifier “some” into a Japanese existential construction, as in (1b) below. The meaning of the English nominal modifier is conveyed in the existential verb *i-ta* “existed”. The translation of (1a) without this technique, i.e., by word-for-word translation, is presented in (1c), where the English nominal modifier “some” is translated into the Japanese nominal modifier *ikuraka-no* “some”. Translation (1c) is perfectly grammatical but less natural than (1b). Actually, sentence (1c) was obtained with a state-of-the-art MT system. If this translation technique were implemented on this system, the system would produce a more natural sentence.

As example (1) illustrates, because MTs are literally translated, they often sound unnatural. Therefore, we decided to compare MTs and HTs regarding the degree of word-for-word translation. To identify word-for-word translation, we used the word-alignment distribution between source sentences and translations, i.e., MTs or HTs, because literally translated words should be more easily aligned than non-literally translated words. Literal translations maintain lexical features such as parts of speech, as can be seen in (1). By contrast, non-literal translations usually lack parallel lexical features.

(1)
a. Some students came.
b. Ki-ta gakusei-mo i-ta
come-PST student-also exist-PST
“Some students came”.
c. Ikuraka-no gakusei-wa ki-ta
some-GEN student-TOP come-PST
“Several students came”.
(GEN: Genitive case marker,
PST: Past tense marker, TOP:
Topic marker)

Figure 1. Translation Example (1)

Let us illustrate the difference in alignment distribution between MTs and HTs.

(2)
a. Today, the sun is shining.
b. Kyoo taiyoo-wa kagayai-teiru
today the-sun-TOP shine-BE-ING
“Today the sun is shining”.
c. Kyoo-wa seiten-da
today-TOP fine-BE
“It’s fine today”.
(TOP: Topic marker, BE: Copular
verb, ING: Gerundive verb
form)

Figure 2. Translation Example (2)

Sentence (2a) below is a source sentence both for the word-for-word translation in (2b), i.e., MT, and the natural translation in (2c), i.e., HT. Table 1 lists the word-alignment distribution attained with our alignment tool. In Tables 1 and 2, “align(A, B)” means that an English word “A” and a Japanese word “B” compose an aligned pair, “non-align_eng(C)” means that an English word “C” remains unaligned, and “non-align_jpn(D)” means that a Japanese word “D” remains unaligned. From the alignment distribution in Tables 1 and 2, we see that the rate of alignment and non-alignment varies between HTs and MTs. That is, non-aligned words often appear in HTs, and more aligned pairs are observed in MTs. Thus, non-aligned words should exhibit HT-likeness, while aligned pairs should exhibit MT-likeness. We constructed a classifier using these aligned pairs and non-aligned words in Tables 1 and 2 as classification features. Since word-alignment properties reveal the lexical correspondences between a source sentence and its counterpart, our classifier can take adequacy into account.

Table 1. Alignment Distribution of MTs

MT (2b)
align(today, kyoo [today])
align(is, teiru [BE-ING])
align(sun, taiyoo [sun])
align(shining, kagayai [shine])
nonalign_jpn(wa [TOP])
nonalign_eng(the)

Table 2. Alignment Distribution of HTs

HT (2c)
align(today, kyoo-wa [today-TOP])
align(is, da [BE])
nonalign_jpn(seiten [fine])
nonalign_eng(the)
nonalign_eng(sun)
nonalign_eng(shining)

3 Experiments

This section describes the design and results of our experiment, and discusses our findings.

3.1 Design

A parallel corpus was prepared for constructing classifiers in the experiment. The corpus consisted of Reuters’ news articles in English and their Japanese translations (Utiyama & Isahara 2003). Since some source sentences and translations appeared repeatedly in our corpus, we deleted these repetitions. The MTs for this corpus were obtained with a commercially available MT system. Word-alignment distributions between the source sentences and the MTs and HTs were obtained with an experimental word-alignment tool.[2] A total of

[2] Experiments with a free alignment tool (Och & Ney 2003) have yet to be done.

258,000 examples were obtained (129,000 HT-alignment examples and 129,000 MT-alignment examples).

We randomly chose 44 sentences from this corpus for a preliminary evaluation of our method.[3] These sentences were assessed by three human evaluators, who had been involved in developing MT systems (not the authors). The evaluators assessed both the adequacy and fluency of MTs, and scored them on a scale from 1 to 4. (See Section 1 for the definitions of adequacy and fluency.)

Machine learning was carried out with an SVM algorithm implemented on the TinySVM software.[4] The linear was taken as a type of kernel function, and the other settings were taken as default settings.

We first appraised the accuracy of classification with our classifier in this experiment. Then, we investigated the correlation between the human-assessment results obtained by our three evaluators to determine the upper bounds for our classification-based method. Finally, we investigated and tested its validity by examining how well the scores computed by the SVMs correlated with the adequacy and fluency scores awarded by the human evaluators.

3.2 Results

Before reporting the experimental results, let us briefly confirm the word-alignment distributions in MTs and HTs. As Table 3 shows, the number of aligned pairs constituted 35% of alignment distributions in MTs. By contrast, the aligned pairs made up 24% in HTs. In Table 3, the number refers to the sum of the aligned pairs and non-aligned words between the 129,000 source sentences and the MTs/HTs. Thus, MTs contain more aligned pairs than HTs. We tested the differences in alignment distributions between HTs (control sample) and MTs with a Fisher exact test. The results revealed that the alignment rate for MTs was significantly greater than that for HTs ($p<0.05$). Based on these results, we concluded that MTs and HTs differed with respect to word-alignment distributions.

Table 3. Alignment Distributions

	N	Aligned pairs (%)	Non-aligned words (%)	Alignment rate (%)
MT	521102	35.7	64.3	55.5
HT	568259	24.1	75.9	31.7

Next, we examined the robustness of our method for machine translation systems by comparing the classification accuracy of three commercially available state-of-the-art translation systems in a five-fold cross validation test. Our method of classification yielded high classification accuracy (98.7, 99.7%, and 99.8%). From these results, we concluded that our method is robust for MT systems.

Now, let us return to the results from the experiment. First, we examined the classification accuracy of our classifier. Its accuracy was obtained through the five-fold cross validation test. Our method of classification achieved a high accuracy of 98.7%. It is difficult to find benchmark methods to compare with our classifier, because previous methods often require multiple reference translations or manually labeled training examples. Since the previous studies used syntactic properties to construct classifiers (Corston-Oliver et al. 2001, Gamon et al. 2005, Mutton et al. 2007), we decided to compare our alignment-distribution-based classifier with a classifier based on syntactic properties, i.e., dependency relations. Although this comparison was not that rigorous, we believe it suggested that our method was valid. HTs and MTs were parsed with the CaboCha parser (Kubo & Matsumoto 2002), and the dependency pairs of a modifier and a modified phrase were used as classification features. This baseline method achieved an accuracy of 83.1%. Our proposed method outperformed the baseline, exhibiting a superiority of 18.8%. Based on these results, we concluded that our word-alignment-based classifier more accurately distinguishes MTs and HTs than a dependency-relation-based classifier.

We next checked the correlation of assessment results between the three human evaluators (I-III). The results for both adequacy and fluency exhibited strong correlations as listed in Table 4. The correlation coefficients for adequacy evaluation varied from .68 to .76, and those for fluency evaluation ranged from .40 to .61. We determined

[3] The number of test sentences should be increased in future experiments to enable more rigorous evaluations of our method. We are now preparing a larger-scale experiment.

[4] The packaging tool is available at the following URL: http://chasen.org/~taku/software/TinySVM/

the upper bounds for our classifier as the mean values of human evaluation. That is, the bound for adequacy evaluation was .73, the bound for fluency evaluation was .53, and the bound for the entire evaluation was .74. The entire evaluation was derived by summing up both adequacy and fluency evaluation scores.

Table 4. Correlation of Human Evaluation Results

	I-II	I-III	II-III	Mean
Adequacy	.76	.74	.68	.73
Fluency	.58	.40	.61	.53
Entire	.76	.70	.75	.74

Finally, we moved on to evaluating the performance of our method. We examined to what extent our classifier-based evaluation results were correlated with the human-evaluation results. The correlations were examined at the sentence level. The MT sentences were evaluated with our method using a score provided by the SVM classifier as described in Section 2.1. The human evaluation consisted of three types of evaluation scores: (i) adequacy, (ii) fluency, and (iii) entire. We assessed our evaluation method (W-A classifier) by comparing it with human evaluations. In addition, we evaluated three other methods: (i) a dependency-based classifier (D-classifier), (ii) NIST (Doddington 2002), (iii), and METEOR (Banerjee & Alon 2005). The correlations were assessed in terms of Spearman's rank-correlation coefficient.

Table 5. Correlation of Automatic-evaluation Results and Human-evaluation Results

	Adequacy	Fluency	Entire
W-A classifier	.44	.43	.47
D-classifier	.33	.37	.37
NIST	.40	.45	.46
METEOR	.20	.19	.20

Table 5 lists the correlation coefficients. In obtaining the evaluation results for NIST (Doddington 2002) and METEOR (Banerjee & Alon 2005), we used HTs of the parallel corpus as reference translations.

3.3 Discussion

Our classification-based method of evaluation, which employed word-alignment distributions as learning features, exhibited a weak correlation with the human-evaluation results for adequacy, fluency, and the entire evaluation, as listed in Table 5. Our method did not surpass the upper bound coefficients, i.e., the mean correlation coefficients between the human-evaluation results in Table 4.

Compared with the other three automatic methods, our classifier outperformed the D-classifier and METEOR (Banerjee & Alon 2005) in the three evaluation criteria, and our method achieved similar results to NIST (Doddington 2002). Our method had a lower correlation coefficient with human fluency evaluation than NIST (Doddington 2002), but it outperformed NIST (Doddington 2002) with respect to adequacy and the entire evaluation. The D-classifier-based method of evaluation did not achieve as high a correlation as NIST (Doddington 2002). From these results, we assumed that our method was tenable as an automatic method of evaluation without the use of reference translations. In addition, our method seems to account for evaluations of adequacy as we assumed that these need to be examined with features covering linguistic correspondences between source sentences and translations, i.e., word alignments (as discussed in Section 2.2). The correlation coefficient of adequacy evaluation for the D-classifier-based evaluation was lower than that of fluency evaluation. By contrast, the adequacy evaluation achieved a higher correlation than the fluency evaluation in the W-A-classifier-based evaluation. This suggests that the W-A-classifier-based evaluation appropriately assessed the adequacy evaluation. We intend to test and verify this conclusion in future studies.

We further appraised the experimental results by comparing them for our method and human evaluation. We consequently found that while fluency evaluation decreased in human evaluation, automatic-evaluation methods (including ours) did not exhibit such drops. All the automatic-evaluation methods exhibited similar correlations between adequacy and fluency evaluations. Hence, unlike human evaluation, automatic evaluation seems stable for evaluating fluency. This constitutes one advantage of automatic evaluation.

Using word-alignment distribution as classification features, we can construct three types of classifiers: (i) a classifier based on aligned pairs (AL), (ii) a classifier based on non-aligned words (n-AL), and (iii) a classifier based on both aligned pairs and non-aligned words (AL & n-AL). We compared the evaluation accuracy with these classifiers by comparing it with human evaluation. As listed in Table 6, the classifier using both aligned and non-aligned words achieved the highest correlation. Hence, this led us to employ both aligned and non-aligned distribution as classification features.

Table 6. Correlation of Classifier-evaluation and Human-evaluation Results

	Adequacy	Fluency	Entire
AL	.28	.32	.33
n-AL	.28	.27	.28
AL & n-AL	.44	.43	.47

In addition, our method could reveal problems with MT systems by enabling weights given to all features in training the SVM classifier to be assessed. The weight of a feature indicates its MT-likeness or HT-likeness with our method. The MT/HT-like properties are proportional to the absolute value of the weight.

Through investigating the weights of features, we found that well-known translation problems in MTs could be detected. As Yoshimi (2001) noted, the translation of English pronouns into non-pronominal Japanese expressions is an MT problem that needs to be resolved. This arises from the linguistic difference between English and Japanese. English is a language that frequently uses pronouns, whereas Japanese uses fewer pronouns. In investigating the weights, we found aligned English pronouns for MT-likeness features and non-aligned English pronouns for HT-likeness features.

Table 7. Weights for HT-like Features

Rank	HT-like	Weight
1	**nonalign_jpn(doo [the same])**	**1.134**
2	nonalign_eng(just)	0.884
3	**nonalign_jpn(doo-si [the same person])**	**0.846**
4	nonalign_jpn(kono [this])	0.805
5	nonalign_jpn(akiraka [clear])	0.727

Table 8. Weights for MT-like Features

Rank	MT-like	Weight
1	nonalign_jpn(paasento [percent])	-0.982
2	**align(and, sosite [and])**	**-0.915**
3	Align(delay, okure [delay])	-0.874
4	**align(and, oyobi [and])**	**-0.796**
5	nonalign_jpn(u [?])	-0.780

Tables 7 and 8 list the five most HT-like features and MT-like features, respectively. As we can see from Table 7, HTs involve "non-align_jpn(doo [the same])" and "non-align_jpn(doo-si [the same person])". These expressions remained non-aligned due to the application of a translation technique to HTs. Here, the meaning of an English pronoun seems to be conveyed with a non-pronominal suffix, "doo- [the same]". Based on how the features are weighted, we can see that this translation technique can be applied to HTs but not to MTs. This is illustrated by example (3). Here, the English pronoun "he" is translated into the Japanese pronoun "kare [he]" in MT (3b). In HT (3c), the English pronoun "he" is translated into "doo-si [the same person]", which conveys an anaphoric meaning more naturally than a pronoun in this context.

(3)
a. He said the policy would
increase textile exports
both in terms of value and
quantity.
b. kare-wa itt-ta. Sono-hooosin-wa
he-TOP say-PST this policy-TOP
kati-no-aru-kikan-ni sosite mata
value-GEN-exist-span-DAT and also
ryoo-de senni-no yusyutu-wo
quantitiy-DAT textile-GEN exports-ACC
zooka-suru-de-aroo-to
increase-will-COMP
c. doo-si-wa sin-booeki-
the-same-person-TOP new-export-
seisaku-ga doonyuu-sareru-to
policy-NOM introduce-PASS-COMP
senni-yusyutu-wa kakaku-to-
textile-exports-TOP value-and-
ryoo-no ryoomen-de
quantities-GEN both-side-DAT
zooka-suru-to katat-ta
increase-COMP tell-PST
(TOP: Topic marker, PST: Past
tense marker, GEN: Genitive
case marker, DAT: Dative case
marker, ACC: Accusative case
marker, COMP: Complementizer,
NOM: Nominative case marker,
PASS: Passive marker)

Figure 3. Translation Example (3)

In addition to translating pronouns, we found that MTs and HTs differed in translating coordinating conjunctions. The English conjunction “and” can conjoin any categorial phrases such as noun phrases, verb phrases, and sentences. Japanese has both a categorially restricted free conjunction, i.e., “sosite [and]” and a restricted conjunction, i.e., “-to [and]”. The latter conjunction can only conjoin nominals. Thus, conjunctions constitute another linguistic discrepancy between Japanese and English. As Fujita (2000) suggests, the translation of the English conjunction “and” into Japanese conjunctive expressions is a translation problem that needs to be resolved. HTs seem to apply another translation rule to conjunctions. While HTs have no alignment features concerning conjunctions, MTs involve aligned pairs for conjunctive expressions, i.e., “align(and, sosiste [and])” and “align(and, oyobi [and])”, as listed in Table 8. This difference in translating conjunctions is also illustrated in example (3). In MT (3b), the English conjunction “and” is translated into “sosite [and]”, while a conjunction is translated into the conjunction suffix “-to” in HT (3c). Noun phrases are more naturally conjoined with the conjunction “-to [and]” than the other conjunction “sosite [and]”.

4 Conclusion

We proposed an automatic method of evaluating MTs, which does not employ reference translations for evaluation of new sentences. Our evaluation metric classifies the results of MTs into either “good” translations (HTs) or “bad” translations (MTs). The classifier was constructed based on the word-alignment relations between source sentences and HTs/MTs, assuming that the alignment distribution reflected MT-likeness and HT-likeness. The classification accuracy in our experiment was 98.7%. We found that this classification-based method of evaluation exhibited a weak correlation with human-evaluation results and that it was more highly correlated with human evaluations than NIST (Doddington 2002) or METOR (Banerjee 2005) metrics. Our examination of how features were weighted revealed problems that studies on MTs should contend with, e.g., translation anaphoric expressions and conjunctive expressions. Our method, which employs parallel corpora, is relatively inexpensive but is an effective automatic evaluation metric.

This paper leaves several problems unsolved. First, we must examine to what extent the alignment features account for the difference between MTs and HTs. Second, we plan to investigate and test the validity of the new method in more detail by comparing our evaluation results with the more extended results attained by human evaluators.

References

Albrecht, J. S. and R. Hwa. 2007. Regression for sentence-level MT evaluation with pseudo references. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL07)*. 296–303.

Banerjee, S. and L. Alon. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the 43th Annual Meeting of the Association for Computational Linguistics (ACL05)*. 65–72.

Blatz, J., E. Fitzgerald, G. Foster, S. Gandrabur, C. Goutte, A. Kulesza, A. Sanchis, and N. Ueffing. 2004. Confidence estimation for machine translation. *Final Report, JHU /CLSP Summer Workshop*.

Corston-Oliver, S., M. Gamon, and C. Brockett. 2001. A machine learning approach to the automatic evaluation of machine translation. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL01)*. 148–155.

Doddington, G. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the 2nd Human Language Technologies Conference (HLT02)*. 128–132.

Fujita, N. 2000. *Nihongo-bunpoo*. Aruku, Tokyo.

Gamon, M., A. Aue and M. Smets. 2005. Sentence-level MT evaluation without reference translations: Beyond language modeling. In *Proceedings of the 10th the European Association for Machine Translation Conference (EAMT05)*. 103–111.

Kubo, T. and Y. Matsumoto. 2002. Japanese dependency analysis using cascaded chunking. In *Proceedings of the 6th Conference on Natural Language Learning 2002 (CoNLL 2002)*. 63–69.

Kulesza, A. and S. M. Shieber. 2004. A learning approach to improving sentence-level MT evaluation. In *Proceedings of the 10th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI04)*. 75–84.

Mutton, A., M. Dras, S. Wan, and R. Dale. 2007. GLEU: Automatic evaluation of sentence-level fluency. In *Proceedings of the 45th Annual Meeting of* *the Association for Computational Linguistics (ACL01)*. 344–351.

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

Papineni, K. A., S. Roukos, T. Ward, and W.-J. Zhu. 2001. Bleu: A method for automatic evaluation of machine translation. *Technical Report RC22176 (W0109–022)*, IBM Research Division, Thomas J. Watson Research Center.

Paul, M., A. Finch, and E. Sumita. 2007. Reducing human assessment of machine translation quality to binary classifiers. In *Proceedings of the 11th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI 2007)*. 154–162.

Utiyama, M. and H. 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 (ACL)*. 72–79.

Vapnik, V. 1998. *Statistical Learning Theory*. Wiley-Interscience, New York.

Yoshimi, T. 2001. Improvement of translation of pronouns in an English-to-Japanese MT system. *Journal of Natural Language Processing* 8 (3): 87–106.

Diagnosing Human Judgments in MT Evaluation: an Example based on the Spanish Language

Olivier Hamon
ELDA
55-57 rue Brillat-Savarin
75013 Paris, France, and
LIPN, U. of Paris XIII
99av. J.-B. Clément
93430 Villetaneuse, France
hamon@elda.org

Djamel Mostefa
ELDA
55-57 rue Brillat-Savarin
75013 Paris, France
mostefa@elda.org

Victoria Arranz
ELDA
55-57 rue Brillat-Savarin
75013 Paris, France
arranz@elda.org

Abstract

This paper aims at providing a methodology for analyzing the reliability of human evaluation in MT. In the scope of the second TC-STAR evaluation campaign, during which a human evaluation on English-to-Spanish was carried out, we first demonstrate the reliability of the evaluation. Then, we define several methods to detect judges who could bias the evaluation with judgments which are too strict, too permissive or simply incoherent.

1 Introduction

For a quarter of a century, many evaluation campaigns involving human evaluation in Machine Translation (MT) have been carried out and surely even more evaluations have taken place outside such campaigns. DARPA, and then NIST MT campaigns[1], among others, were certainly the most influential in human evaluation. However, recent evaluation campaigns such as IWSLT (Fordyce, 2007), TC-STAR (Mostefa et al., 2006), CESTA (Hamon et al., 2007) or WMT (Callison-Burch et al., 2007) have also highlighted the importance of human evaluation in MT. The results are checked carefully so as to assess system quality, especially due to the weakness of the automatic or semi-automatic metrics. However, what is not always highlighted are the inconsistencies of the human evaluation process, given that this remains the result of subjective judgments. It is particularly important to observe in detail how human judges react according to what they evaluate. Some measures have been defined to estimate a judge's consistency (Blanchon et al., 2004) or the number of judgments which are needed to have a relevant evaluation campaign (Koehn, 2007). It is well-known that inter-judge agreement is generally far from perfect (Ye and Abney, 2006), and even professional human translators disagree through different cases of translation. If this was not the case, one unique reference translation would be sufficient. However, how do judges evaluate a segment, depending on whether it is low or high quality? What are the difficulties met by judges which cause such lack of consistency among them?

Most of the previous evaluation campaigns have been carried out with English as a target language. However, some others have used languages with a richer morphology, such as Spanish or French. The answers we try to get in this experiment could help to improve the human evaluation set up, in particular when using morphologically richer languages like Spanish.

After describing the framework of our experiments we try to determine a methodology to find judges consistency and, if need be, to delete judges who would have done random evaluation. Finally, we draw some conclusions on our experiments.

2 Framework and General Results

The experiment presented here is done using the material from the TC-STAR second evaluation campaign (Mostefa et al., 2006). During this campaign, a human evaluation was carried out on

[1] http://www.nist.gov/speech/tests/mt/

English-to-Spanish direction, with data coming from European Parliament Plenary Sessions. The vocabulary used in these data belongs to the political and diplomatic domains.

The experiment involves three kinds of input: automatic transcriptions from Automatic Speech Recognition (ASR) systems, manual transcriptions (Verbatim) and Final Text Edition (FTE) data provided by the European Parliament. Each input has its own attributes and difficulties: the ASR input contains sentences with errors deriving from ASR systems; the sentences in the Verbatim input include spontaneous speech phenomena such as hesitations, corrections or false-starts; the FTE input sentences have been rewritten and do not include spontaneous speech phenomena.

Although we distinguish systems for ASR, Verbatim and FTE in the following results, we do not separate them, and thus we obtain a large range of scores, from the presumed lower quality ones (ASR) to the presumed better ones (FTE). 26 systems were evaluated within this human evaluation as a whole, which can be split up into 6 ASR systems, 9 Verbatim systems and 11 FTE systems. A subset of around 400 segments for each system output was used for the evaluation. Since each segment was evaluated twice, an overall of 20,360 segments were evaluated by 125 judges, corresponding to around 163 segments per judge. Judges were native Spanish speakers and did the evaluation through an interface available on Internet.

Each segment was evaluated in relation to both *adequacy* and *fluency* measures (White et al., 1994). For fluency, the quality of the language is evaluated and the judges had to answers to the question *“Is the text written in good Spanish?”*. A five-point scale was provided ranging from *“Spotless Spanish”* to *“Non understandable Spanish”*. For adequacy, automatic translations and corresponding reference segments were compared and the judges had to answer to the following question: *“How much of the meaning expressed in the reference translation is also expressed in the target translation?”*. A five-point scale was also provided to the judges ranging from *“All the meaning”* to *“Nothing in common”*. For both fluency and adequacy only extreme points were proposed on the scale, the rest of the points were unconstrained and then dependent on the judges’ opinion.

The judges evaluated all their segments firstly according to fluency, and then according to adequacy. Thus, the fluency measure is applied independently and judges are not influenced by the reference translation. Both evaluations per segment are done by two different judges and no judge evaluates the same segment coming from two different systems. Finally, the segments are presented randomly.

The general results are shown in Figure 1.

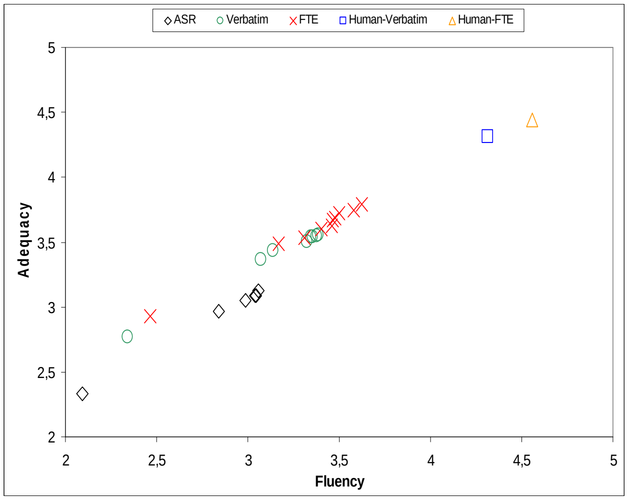

Figure 1: General results for fluency and adequacy.

Both Verbatim and FTE outputs located in the top right-hand side corner are coming from the human reference translations (“Human-Verbatim” and “Human-FTE”, respectively) and are clearly higher than the automatic translations (“ASR”, “Verbatim”, “FTE”). Scores are better for FTE systems, then for Verbatim ones and, finally, ASR systems get the lowest results. This allows us to use a large set of sentence qualities and observe how judges evaluate accordingly.

3 Methodology and the Problem of the Human Evaluation

A main step when using human evaluation in MT is to define a protocol and a methodology to perform the test. Once the evaluation has been finalized by the judges, looking at the results is not sufficient. It is also important to know how reliable these judges are. Several methods can be used to determine the reliability of the evaluation, not giving the same information, but giving an indication about the performance of judges.

However, judgments are at any rate subjective. In this experiment, judges are not experts but end users and they react differently according to their condition, culture or knowledge. One of our goals is to determine how their judgments can be subjective. Then, we would like to define the kinds of segments that can pose a problem when reliability is low.

Then, the question we ask is: Are the Judgments “Correct”?

There are several ways to compute the agreement between judges. We present two of them here, a variation of the inter-judge agreement and the Kappa coefficient (Miller and Vanni, 2005). To go further, we try to detect whether some judges have particularly unfair results. This does not necessarily mean that judges are wrong, but that some of them could be too strict in comparison with the other judges.

3.1 Inter-judge agreement

Instead of computing a strict inter-judge agreement based on a binary agreement (two evaluators agree or disagree on a single segment), we have decided to measure an *n*-agreement, for which *n* is the upper difference between two scores of a same segment. For *N* segments, this is defined as follows:

$$n - agreement(n) = \frac{1}{N}\sum_{i=1}^{N}\delta(\left|S_i^a - S_i^b\right| \leq n)$$

$$where:$$

$$\delta(\left|S_i^a - S_i^b\right| \leq n) = \begin{cases} 1 \; if \; \left|S_i^a - S_i^b\right| \leq n \\ 0 \; if \; \left|S_i^a - S_i^b\right| > n \end{cases}$$

n-agreement is described as the ratio of the number of segments for which the difference between the first evaluation of segment *S*, S_i^a, and its second evaluation, S_i^b, is lower or equal to *n*.

The results for the fluency and adequacy evaluations inter-judge agreement are presented in Table 1.

Evaluation	Input	*n*-agreement				
		0	1	2	3	4
Fluency	FTE	.34	.70	.88	.97	1
	Verb.	.34	.69	.87	.96	1
	ASR	.29	.63	.85	.95	1
	Cumul.	.33	.69	.87	.96	1
Adequacy	FTE	.35	.68	.88	.97	1
	Verb.	.33	.67	.87	.96	1
	ASR	.30	.66	.84	.95	1
	Cumul.	.33	.66	.87	.96	1

Table 1: Inter-judge *n*-agreement for the different types of data input.

Inter-judge agreement is quite similar whatever the data input or criteria of evaluation. Judges give exactly the same score for a third of the evaluated segments. This is quite low and demonstrates the relative subjectivity of the evaluation. However, around 70% of the evaluations do not differ in more than 1 point. Therefore, it seems more reasonable to use a 3 point scale instead of a 5 point scale.

We have observed that ASR input seems slightly harder to judge than Verbatim input, which is also slightly harder to judge than FTE input.

3.2 Calculation of the Kappa Coefficient

In addition to the inter-judge agreement we measure the global Kappa coefficient (Landis and Koch, 1977a), which allows to measure the agreement between *n* judges with *k* criteria of judgment. The measure goes further, taking into account the chance factor that judges give identical judgment on a same segment. For *N* segments, it is defined as:

$$\kappa = \frac{\overline{P_o} - \overline{P_e}}{1 - \overline{P_e}}$$

Where

$$\overline{P_o} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{n(n-1)}\sum_{j=1}^{k} n_{ij}(n_{ij} - 1)$$

and

$$\overline{P_e} = \sum_{j=1}^{k}\left(\frac{1}{Nn}\sum_{i=1}^{N} n_{ij}\right)^2$$

The amount of judges who evaluate the ith segment in the jth is represented by n_{ij}.

In other words, $\overline{P_o}$ is the proportion of observed agreement and $\overline{P_e}$ is the proportion of random agreement (also called *chance agreement*).

The values we obtain are shown in Table 2.

Evaluation	$\overline{P_o}$	$\overline{P_e}$	K
Fluency	.331	.209	.155
Adequacy	.326	.222	.135

Table 2: Global Kappa coefficient values for fluency and adequacy.

According to (Landis and Koch, 1977a), K values for both fluency and adequacy mean that judges agree slightly. But (Feinstein and Cicchetti, 1990) presented the limit of Kappa for low values even when agreement was high. This allows us to draw here some weakness of the Kappa coefficient at a practical level. It is representative of the exact comparison of the judgments, without taking into account the closeness of the judgments. One of its limitations is precisely that two judges who have close results would be penalized, as opposed to two judges with distinct results. This kind of case is particularly common in MT evaluation. Moreover, systematic errors between judges cause better coefficients since $\overline{P_e}$ would be lower.

One of the reasons for this low K value can also be the number of judgments per segment, or the number of judges. But according to (Feinstein and Cicchetti, 1990), the minimal ratio is 6 evaluators for 30 segments, which seems impossible regarding our 10,380 segments for this experiment!

Finally, computing the Kappa coefficient does not provide better information about the inter-judge *n*-agreement, which informs more precisely about the reliability of the evaluation regarding different aspects of precision.

3.3 Methods for Detecting Outliers

When an evaluation is done, it is not easy to know whether judges do their evaluations seriously or not. This is particularly so if the judgments are not done by experts and with a large number of people. Judges can be more or less familiar with the tool they used to evaluate, some of them may be tired, or even not feeling well, etc. We should bear in mind that an overall evaluation can take around 2 or 3 hours, with or without pauses, which could cause a drop in the judge's attention.

To reduce the unavoidable subjectivity of the judgments, we try to locate outliers whose judgments are badly evaluated, if there are any. If these judges were detected, it may be useful to delete them from the evaluation set in order to homogenize the results and have a fair evaluation of systems. Three methods have been defined in order to detect those outliers.

Mean score by Judge. Each judge evaluates a subset of around 163 segments. This subset has been built randomly and should be representative of the whole set of segments (10,180 segments). Since each segment has been evaluated twice, we can compute the mean score of the judge on his subset and compare it with the score of the same subset obtained with other judges.

Figure 2 and Figure 3 show the mean score for fluency and for adequacy, respectively. Judges are ranked increasingly, so as to have judges' scores sorted from the lowest to the highest.

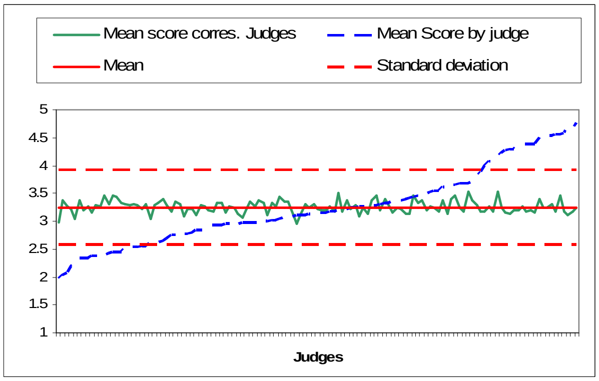

Figure 2: Mean score by judge for fluency and mean score for corresponding judgments from other judges.

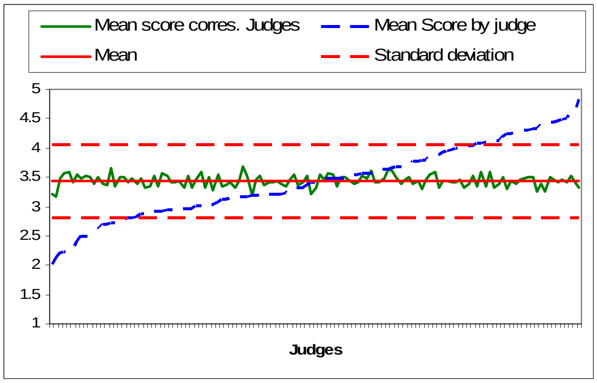

Figure 3: Mean score by judge for adequacy and mean score for corresponding judgments from other judges.

The variation of mean score by judge is similar for both fluency and adequacy. As expected, the score of each subset (plain peaky curve, Figures 2 and 3) is close to the general score of the whole set of segments (plain straight line). So each judge's subset is a representative sample of the whole data.

What is more surprising is the curve of the mean score by judge (dashed curved lines, Figures 2 and 3). We can see that some judges gave very low or very high scores compared to the other judgments on the same subset of segments.

We suspect that these evaluators misunderstood the 5-point scale or did not pay enough attention to the evaluation, or are either too strict or not strict enough. Judges above and behind the standard deviation are deviant and could probably be turned down to homogenize the judgment or be asked to redo their evaluation and thus obtain a more objective evaluation.

This method allows us to compare the score of each judge with the score of his subset of segments. But of course, for a given judge, we can have a mean score that is very close to the mean score of his subset with big differences for each segment. This is why we investigated the mean agreement by judge.

Mean agreement by judge. For each judge, a distance score is computed between his own judgment on a segment and the corresponding judgments from the other judges on the same segment. In other words, a mean agreement is measured for each judge comparing his own judgments to those of the other judges who assessed the same segments.

Figure 4 and Figure 5 present the mean agreement for fluency and for adequacy, respectively. Once again, judges's scores are ranked in an increasing manner.

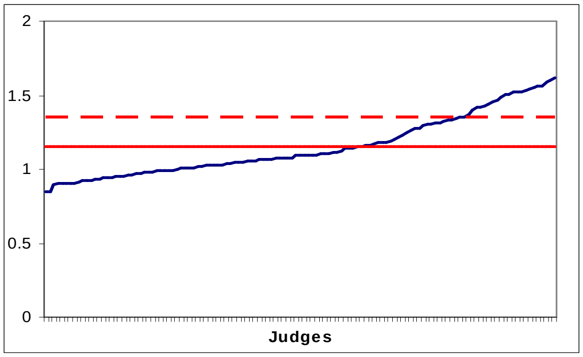

Figure 4: Mean agreement for fluency.

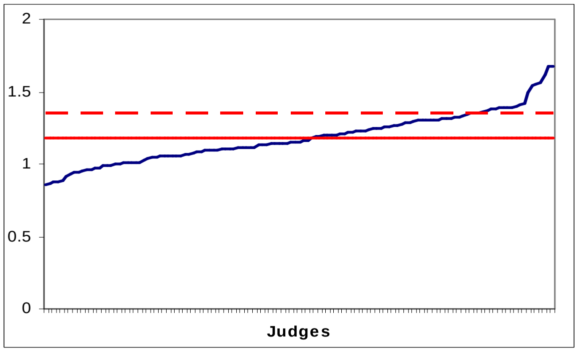

Figure 5: Mean agreement for adequacy.

As for the *Mean Score*, fluency and adequacy curves follow the same trend, although the adequacy one increases faster when agreement is higher. For certain judges, mean agreement is above 1.5, which is quite high since the largest mean agreement possible is 4. This means that these judges disagree with other judges in 1.5 in mean. It does not necessarily mean that these judges have not done their judgments correctly (what is even more, they are close to the other judges), but simply that some judges are stricter than others.

Easy sentences. Some sentences are easier to translate than others, because of their length, their simpler lexical or syntactic content, etc. We decided to make a selection of those “easy sentences” in order to observe the judgments done. In theory, those sentences should not represent any problem for automatic systems: these systems should not make any mistakes, and then judgments should be “perfect”. Thus, a lower judgment draws our attention to the judge who has done it if the automatic system actually managed to translate the sentence correctly. Easy sentences can be described as containing few words, or easy words to translate like “gracias”. They can also be sentences which occur frequently in the data (even in the development data to train the systems). Figure 6 illustrates some of those sentences.

¿ podría hacer más la Comisión ?
gracias , Presidente .
la respuesta es compleja .
gracias .

Figure 6: Examples of "easy sentences".

A total of 80 segments have been manually identified, which should allow us to identify evaluators for whom evaluations are incoherent.

The study of the translated sentences and their judgments shows some segments that are not correctly assessed, which does not mean that the general results of the judge who assessed them are not correct either. Such segments are localized and reasons for such erroneous assessment could be fatigue, lack of attention, or others things which are more linked to the activity than to the judge's competence itself.

For detecting incorrect segments, a study should be done at a segment level. But currently it is hard to provide such a study since there are only two evaluations per segment and, more particularly, because of the tedious and time-consuming work to be done. Moreover, the proportion of such segments seems to be very low and in any case, these segments are drowned in the whole volume of segments.

However, even if our analysis is quite subjective, some judges seem to evaluate incorrectly a significant amount of easy sentences.

3.4 Removing Outliers

The standard deviation allows to observe the statistical dispersion of judges away from the mean. Thus, we can remove judges who are above the positive standard deviation (for mean score and mean agreement) and under the negative standard deviation (only for mean score). Then, Table 3 can be drawn to compare the judges deleted for each method.

	Mean Score	Mean Agreement	Easy Sentences
Fluency	45	23	9
Adequacy	45	18	10
Fluency + Adequacy	30	10	4

Table 3: Number of judges deleted with the three methods.

Moreover, for fluency, 20 judges are common to both *Mean Score* and *Mean Agreement*, while for adequacy there are only 15. Most of them are included in the upper part of the *Mean Score* (17 for fluency, 5 for adequacy), the others in the lower part (3 for fluency, 10 for adequacy). It seems that outliers are too permissive for fluency, but on the contrary they are too strict for adequacy. Indeed, for fluency, judges have only the translated segment to evaluate, they have nothing to compare with and then are more flexible regarding the different possibilities of judgments. However, when comparing to the reference segment for adequacy, judges are then able to try to match exactly both segments. Another possible reason for being more permissive regarding fluency could be that an MT user's expectations are always higher with regard to content transmission than with regard to syntactic perfection, or rather, that a system's user will mind less having a syntactically imperfect output than a semantically inaccurate one.

Should we decide to delete those judges, that means that more than a third of the judgments would be deleted for *Mean Score*, that the number of judgments for *Mean Agreement* would be divided by 6, and finally divided by 13 for *Easy Sentences*.

This experiment may not mean that we delete "bad judges", but rather that we only delete judges who diverge from the set of judges. Thus we homogenize the evaluation.

After deleting judges and their judgments, we have computed again the scores of the human evaluation. Table 4 shows the Pearson correlations between the scores of the official evaluation presented above, and the scores after deleting judges, for the three methods of identification.

	Mean Score	Mean Agreement	Easy Sentences
Fluency	.98	.99	.98
Adequacy	.99	1.00	.99

Table 4: Pearson correlations between official scores and scores after deleting judges.

Spearman's rank correlation is up to .99 for all the methods and criteria.

The results are not really surprising for the *Mean score*: judges who have higher and lower mean scores have been deleted and they about complement each other.

However, this is more surprising for the *Mean agreement*. Deleted judgments are in strong disagreement with the judgments from other judges, so scores should be from the boundaries and they bias strongly the results. In fact, comparing "good judges" with outliers, scores are not identical but very close: For fluency, mean scores are 3.47 against 3.24 and 3.26 against 3.91 for adequacy mean scores, respectively. Values for deleted judges are still low regarding other judges, and they are still representative for the whole evaluation set.

Regarding *Easy Sentences*, the amount of judges removed is probably not sufficient to affect the scores, all the more so, as according to the previous comments, there are no real divergent judgments.

Although general results are higher, or lower, the trend of results is identical, like the systems ranking, as shown in Figure 7.

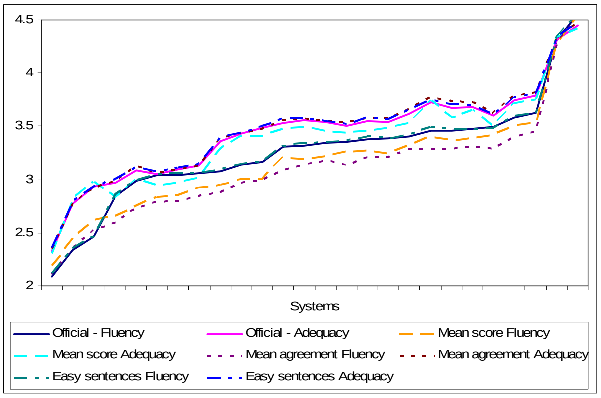

Figure 7: System scoring for Official results and the methods of judges deletion.

Another interesting point would be to know whether low agreement means wrong scores. In the same way as above, we have computed the Pearson correlation between official results and scores from the deleted judges only. Pearson correlation for fluency is .84, and for adequacy .96. This does not make a strong difference between the results for adequacy. However, this difference is more important for fluency. As mentioned earlier, that reflects bigger differences between judges for fluency (but not particularly higher difficulties for evaluating it), because of the absence of comparison to a reference. Since judges are typically free in their evaluation (i.e. there is no detailed guideline), they are more heterogeneous than for the adequacy evaluation, during which they refer to a single reference (and, which, in a certain way, serves the purpose of a guideline).

However, keeping all the judges does not really affect the systems ranking either for fluency or for adequacy, although, in general, scores are slightly lower.

4 Conclusion and Further Work

Our experiment is based on three general points in order to diagnose the performance of human judges in machine translation in two ways: a statistical observation of the judgment, and a linguistic study of the evaluated sentences.

First, we observed the agreement between judges to estimate the reliability of the evaluation. We drew the conclusion with the inter-judge *n*-agreement that this experiment contains an extremely detailed scale of judgments (5 points), which seems to confuse the evaluators, and we propose to limit the criteria to three. It would be interesting to make the same observations taking into account three criteria, for instance by merging criteria "1" and "2" , and "4" and "5", and then studying the difference. Using the Kappa coefficient has proved its limitations in a practical case, since it does not take into account the variation of the agreement.

Then we tried to define a protocol and methods for detecting outliers, i.e., judges who are too subjective regarding other judges. In that experiment, deleting this kind of judges did not clearly change scores and ranking. Moreover, the number of judgments done does not allow to change the score so easily when deleting several judges.

Our future work will consist in applying this method to the third evaluation campaign of the TC-STAR project (under the same conditions but different judges), and to French corpora from the CESTA evaluation campaigns. This should allow us to observe the consistency of judges and perform intra-judge agreement too.

We also would like to do the same kind of study, but this time according to segments, systems and data criteria. Although it is important to measure the reliability of human evaluation, we also need to find how to improve the methodology and, most of all, to understand why judges evaluate sentences in such a way. This is directly linked to a currently-ongoing linguistic study of the evaluated segments and how these reflect the judges' criteria and skills.

Acknowledgement

This work was supported by the TC-STAR project (grant number IST-506738).

References

Blanchon H., Boitet C., Brunet-Manquat F., Tomokio M., Hamon A., Hung V. T. and Bey Y. 2004. *Towards Fairer Evaluation of Commercial MT Systems on Basic Travel Expressions Corpora.* Proc. IWSLT 2004. Kyoto, Japan.

Callison-Burch C., Fordyce C. and Koehn P., Monz C. and Schroeder J. 2007. (Meta-) Evaluation of Machine Translation Proceedings of the Second Workshop on Statistical Machine Translation, June 2007, Prague, Czech Republic.

Feinstein A.R., Cicchetti D.V. 1990. *High agreement but low kappa: I. The problems of Two Paradoxes.* J. Clin. Epidemiol., 43, 543-548.

Fordyce C. 2007. Overview of the IWSLT 2007 *campaign.* In Proceedings of IWSLT 2007, Trento, Italy.

Hamon O., Hartley A., Popescu-Belis A. and Choukri K. 2007. *Assessing Human and Automated Quality Judgments in the French MT Evaluation Campaign CESTA.* In Proceedings of MT Summit XI, September, 2007, Copenhagen, Denmark.

Koehn P. 2007. *Evaluating Evaluation Lessons from the WMT 2007 Shared Task.* MT Summit XI Workshop on Automatic Procedures in Machine Translation Evaluation, September 2007, Copenhagen, Denmark.

Landis J.R., Koch G.G. 1977a. *The Measurement of Observer Agreement for Categorical Data.* In Biometrics, 33, 159-174.

Landis J.R., Koch G.G. 1977b. *A one-way components of variance model for categorical data.* Biometrics, 33, 671-679.

Miller K.J., Vanni M. 2005. Inter-rater Agreement *Measures and the Refinement of Metrics in the PLATO MT Evaluation Paradigm.* In Proceedings of MT Summit X, Phuket, Thailand.

Mostefa D., Hamon O. and Choukri K. 2006. *Evaluation of Automatic Speech Recognition and Spoken Language Translation within TC-STAR: results from the first evaluation campaign*, Proc. Language Resources and Evaluation Conference, Genoa, Italy.

Mostefa D., Garcia M-N., Hamon O. and Moreau N. 2006. *Evaluation report, Technology and Corpora for Speech to Speech Translation (TC-STAR) project.* Deliverable D16.

White J. S. and O'Connell T. A. 1994. *The ARPA MT Evaluation Methodologies: Evolution, Lessons, and Future Approaches.* Proceedings of AMTA Conference, 5-8 October 1994, Columbia, MD, USA.

Ye Y. and Abney S. 2006. *How and Where do People Fail with Time: Temporal Reference Mapping Annotation by Chinese and English Bilinguals.* In Proceedings of Frontiers in Linguistically Annotated Corpora 2006, a Merged Workshop with 7th International Workshop on Linguistically Interpreted Corpora (LINC-2006) and Frontiers in Corpus Annotation III at ACL, Sydney, Australia, pp13-20.

Mixing Approaches to MT for Basque: Selecting the best output from RBMT, EBMT and SMT

I. Alegria, A. Casillas, A. Diaz de Ilarraza, J. Igartua, G. Labaka, M. Lersundi, A. Mayor, K. Sarasola
Ixa taldea. University of the Basque Country.
X. Saralegi
Elhuyar Fundazioa
B. Laskurain
Eleka S.L.
i.alegria@ehu.es

Abstract

We present the first steps in the definition of a mixing approach to MT for Basque based on combining single engines that follow to three different MT paradigms. After describing each engine we present the hierarchical strategy we use in order to select the best output, and a first evaluation.

1 Introduction and Basque Language

Basque is a highly inflected language with free order of sentence constituents.

It is an agglutinative language, with a rich flexional morphology. In fact for nouns, for example, at least 360 word forms are possible for each lemma. Each of the declension cases such as absolutive, dative, associative… has four different suffixes to be added to the last word of the noun phrase. These four suffix variants correspond to indefinite, definite singular, definite plural and "close" definite plural. Basque syntax and word order is very different compared with other languages as Spanish, French or English.

Machine translation is both, a real need, and a test bed for our strategy to develop NLP tools for Basque. We have developed corpus based and rule based MT systems, but they are limited.

On the one hand, corpus based MT systems base their knowledge on aligned bilingual corpora, and the accuracy their output depends heavily on the quality and the size of these corpora. When the pair of languages used in translation have very different structure and word order, obviously, the corpus needed should be bigger.

Being Basque a lesser resourced language, nowadays large and reliable bilingual corpora are unavailable for Basque. Domain specific translation memories for Basque are not bigger than two-three millions words, so they are still far away from the size of the present corpora for languages; e.g., Europarl corpus (Koehn, 2005), that is becoming a quite standard corpus resource, has 30 million words. So, the results obtained in corpus based MT to Basque are promising, but they are still not ready for public use.

On the other hand, the Spanish->Basque RBMT system Matxin's performance, after new improvements in 2007 (Labaka et al., 2007), is becoming useful for assimilation, but it is still not suitable enough to allow unrestricted use for text dissemination.

Therefore it is clear that we should combine our basic hes for MT (rule-based and corpus-based) in order to build a hybrid system with better performance. As the first steps on that way, we are experimenting with two simple mixing alternative approaches used up to now for languages with huge corpus resources:

- Selecting the best output in a multi engine system (MEMT, Multi-engine MT), in our case combining RBMT, EBMT and SMT approaches.
- Statistical post-editing(SPE) after RBMT.

This paper deals with the first approach. Our design has been carried out bearing in mind the following concepts:

- Combination of MT paradigms.
- Reusability of previous resources, such as

translation memories, lexical resources, morphology of Basque and others.

- Standardization and collaboration: using a more general framework in collaboration with other groups working in NLP.
- Open-source: this means that anyone having the necessary computational and linguistic skills will be able to adapt or enhance it to produce a new MT system,

Due to the real necessity for translation in our environment the involved languages would be Basque, Spanish, French and English.

The first strategy we are testing when we want to build a MT engine for a domain, is translating each sentence using each of our three single engines (rule-based, example-based and statistical) and then choosing the best translation among them (see section 4).

In section 2 we present the corpus that we will use in our experiments, while in section 3 we explain the single engines built up for Basque MT following the three traditional paradigms: rule-based, example-based and statistical. In section 4, we report on our experiment to combine those three single engines. We finish this paper with some conclusions.

2 The corpus

Our aim was to improve the precision of the MT system trying to translate texts from a domain. We were interested in a kind of domain where a formal and quite controlled language would be used and where any public organization or private company would be interested in.

Finally the domain related to *labor agreements* was selected. The Basque Institute of Public Administration (IVAP[1]) collaborated with us in this selection, by examining some possible domains, parallel corpora available and their translation needs. The Labor Agreements Corpus is a bilingual parallel corpus (Basque and Spanish) with 585,785 words for Basque and 839,003 for Spanish. We automatically aligned it at sentence level and then manual revision was performed.

As said before, our goal is to combine different MT approaches: Rule-Based (RBMT), Example Based (EBMT) and Statistical (SMT). Once we had the corpus, we split it in three for SMT (training, development and test corpus) and in two for EBMT (development and test corpus).

To build the test corpus the full text of several labor agreements was randomly chosen. We chose full texts because we wanted to ensure that several significant but short elements as the header or the footer of those agreements would be represented, and because it is important to measure the coverage and precision we get when translating the whole text in one agreement document and not only some sentences of parts of it. System developers are not allowed to see the test corpus.

In SMT we use the training corpus to learn the models (translation and language model); the development corpus to tune the parameters; and the test corpus to evaluate the system.

In RBMT and EBMT there are not parameters to optimize, and so, we consider only two corpora: one for the development (joining the training and development ones) and one for the test.

The size of each subset is shown in Table 1 (eu= Basque, es = Spanish).

		Doc	Sentences	Words
Training	es	81	51,740	839.393
	eu	81		585,361
Development	es	5	2,366	41,508
	eu	5		28,189
Test	es	5	1,945	39,350
	eu	5		27,214

Table 1. Labor Agreements Corpus

3 Single MT engines for Basque

In this section we present three single engines for Spanish-Basque translation following the three traditional paradigms: rule-based, example-based and statistical. The first one has been adapted to the domain corpus, and the other two engines have been trained with it.

3.1 The rule-based approach

In this subsection we present the main architecture of an open source MT engine, named *Matxin* (Alegria et al., 2007), the first implementation of which translates from Spanish into Basque using the traditional transfer model and based on shallow and dependency parsing. Later on, in a second step, we have specialized it to the domain.

[1] http://www.ivap.euskadi.net

The design and the programs of Matxin system are independent from the pair of languages, so the software can be used for other projects in MT. Depending on the languages included in the adaptation, it will be necessary to add, reorder and change some modules, but this will not be difficult because a unique XML format is used for the communication among all the modules.

The project has been integrated in the *OpenTrad*[2] initiative, a government-funded project shared among different universities and small companies, which include MT engines for translation among the main languages in Spain. The main objective of this initiative is the construction of an open, reusable and interoperable framework.

In the *OpenTrad* project, two different but coordinated architectures have been carried out:

- A shallow-transfer based MT engine for similar languages (Spanish, Catalan and Galician).
- A deeper-transfer based MT engine for the Spanish-Basque and English-Basque pair. It is named *Matxin* and it is stored in *matxin.sourceforge.net.* It is an extension of previous work in IXA group.

In the second engine, following the strategy of reusing resources, another open source engine, *FreeLing* (Carreras et al., 2004)*,* was used for analysis.

The transfer module is divided into three phases dealing at the level of the three main objects in the translation process: words or nodes, chunks or phrases, and sentences.

- First, lexical transfer is carried out using a bilingual dictionary compiled into a finite-state transducer.
- Then, structural transfer at sentence level is applied, some information is transferred from some chunks to others, and some chunks may disappear. For example, in the Spanish-Basque transfer, person and number information of the object and the type of subordination are imported from other chunks to the chunk corresponding to the verb chain.
- Finally the structural transfer at chunk level is carried out. This process can be quite simple (e.g. noun chains between Spanish and Basque) or more complex (e.g. verb chains between these same languages).

The XML file coming from the transfer module is passed on the generation module.

- In the first step, syntactic generation is performed in order to decide the order of chunks in the sentence and the order of words in the chunks. Several grammars are used for this purpose.
- Morphological generation is carried out in the last step. In the generation of Basque, the main inflection is added to the last word in the phrase (in Basque, the declension case, the article and other features are added to the whole noun phrase at the end of the last word), but in verb chains other words need morphological generation. A previous morphological analyzer/generator for Basque (Alegria et al., 1996) has been adapted and transformed to the format used in *Apertium*.

	BLEU	Edit-distance TER
Corpus1 (newspapers)	9.30	40.41
Corpus2 (web magazine)	6.31	43.60

Table 2. Evaluation for the RBMT system

The results for the Spanish-Basque system using *FreeLing* and *Matxin* are promising. The quantitative evaluation uses the open source evaluation tool IQMT and figures are given using Bleu and NIST measures (Giménez et al., 2005). An additional user based evaluation has been carried out too, using Translation Error Rate (Snover, 2006). The results using two corpora without very long sentences are shown in Table 2 (Mayor, 2007).

We have to interpret the results having in mind that the development of this RBMT system was based on texts of newspapers.

Adaptation to the domain

The adaptation to the domain has been out in three main ways:

[2] www.opentrad.org

- Terminology. Semiautomatic extraction of terminology using Elexbi, a bilingual terminology extractor for noun phrases (Alegria et al., 2006). Additionally, an automatic format conversion to the monolingual and bilingual lexicons is carried out for the selected terms. More than 1,600 terms were extracted from the development corpus, manually examined, and near to 807 were selected to be included in the domain adapted lexicon.
- Lexical selection. Matxin does not face the lexical selection problem for lexical units (Matxin only does it for the preposition-suffix translation); just the first translation in the dictionary is always selected (the other possible lexical translations are stored for the post-edition). For the domain adaptation, a new order for the possible translations has been calculated in the dictionary, based on the parallel corpus and using GIZA++.
- Resolution of format and typographical variants which are found frequently in the administrative domain.

After this improvements this engine is ready to process the sentences from this domain.

3.2 The example-based approach

In this subsection we explain how we automatically extract translation patterns from the bilingual parallel corpus and how we exploit it in a simple way.

Translation patterns are generalizations of sentences that are translations of each other in that various sequences of one or more words are replaced by variables (McTait, 1999).

Starting from the aligned corpus we carry out two steps to automatically extract translation patterns.

First, we detect some concrete units (entities mainly) in the aligned sentences and then we replace these units by variables. Due to the morphosyntactic differences between Spanish and Basque, it was necessary to execute particular algorithms for each language in the detection process of the units. We have developed algorithms to determine the boundaries of dates, numbers, named entities, abbreviations and enumerations.

After detecting the units, they must be aligned, to relate the Spanish and Basque units of the same type that have the same meaning. While in the case of numbers, abbreviations and enumerations the alignment is almost trivial, in the case of named entities, the alignment algorithm is more complex. It is explained in more detail in (Martinez et al., 1998). Finally, to align the dates, we use their canonical form.

Table 3 shows an example of how a translation pattern is extracted.

Once we have extracted automatically all the possible translation patterns from the training set, we store them in a hash table and we can use them in the translation process. When we want to translate a source sentence, we just have to check if that sentence matches any translation pattern in the hash table. If the source sentence matches a sentence of the hash table that has not any variable, the translation process will immediately return its translation. Otherwise, if the source sentence does not exactly match any sentence in the hash table, the translation process will try to generalize that sentence and will check again in the hash if it finds a generalized template. To generalize the source sentence, the translation process will apply the same detection algorithms used in the extraction process.

In a preliminary experiment using a training corpus of 54.106 sentence pairs we have extracted automatically 7.599 translation patterns at sentence level.

Aligned sentences	**Aligned sentences with generalized units**	**Translation pattern**
En Vitoria-Gasteiz, a 22 de Diciembre de 2003.	En <rs type=loc> Vitoria-Gasteiz </rs> , a <date date=22/12/2003> 22 de Diciembre de 2003</date> .	En <rs1> , a <date1>.
Vitoria-Gasteiz, 2003ko Abenduaren 22.	<rs type=loc> Vitoria-Gasteiz </rs> , <date date=22/12/2003> 2003ko Abenduaren 22</date>.	<rs1>, <date1>.

Table 3. Pattern extraction process

These translation patterns cover 35.450 sentence pairs of the training corpus. We also think that an aligned pair of sentences can be a translation pattern if it does not have any generalized unit but it appears at least twice in the training set.

As this example based system has a very high precision but quite low coverage (see Table 6 and Table 7), it is very interesting to combine with the other engines specially in this kind of domain where a formal and quite controlled language is used.

3.3 The SMT approach

The corpus-based approach has been carried out in collaboration with the National Center for Language Technology in Dublin.

The system exploits SMT technology to extract a dataset of aligned chunks. We have conducted Basque to English (Stroppa et al., 2006) and Spanish to Basque (Labaka et al., 2007) translation experiments, based on a quite large corpus (270,000 sentence pairs for English and 50,000 for Spanish).

Freely available tools are used to develop the SMT systems:

- GIZA++ toolkit (Och and H. Ney, 2003) is used for training the word/morpheme alignment.
- SRILM toolkit (Stolcke, 2002) is used for building the language model.
- Moses Decoder (Koehn et al., 2007) is used for translating the sentences.

Due to the morphological richness of Basque, in translation from Spanish to Basque some Spanish words, like prepositions or articles, correspond to Basque , and, in case of ellipsis, more than one of those suffixes can be added to the same word. In order to deal with this features a morpheme-based SMT system has been built.

Adapting the SMT system to work at morpheme level consists on training the basic SMT on the segmented text. The system trained on these data will generate a sequence of morphemes as output. In order to obtain the final Basque text, we have to generate words from those morphemes.

To obtain the segmented text, Basque texts are previously analyzed using *Eustagger* (Aduriz and Díaz de Ilarraza, 2003). After this process, each word is replaced with the corresponding lemma followed by a list of morphological tags. The segmentation is based on the strategy proposed on (Agirre et al., 2006).

Both systems (the conventional SMT system and the morpheme based), were optimized decoding parameters using a Minimum Error Rate Training. The metric used to carry out the optimization is BLEU.

The evaluation results in a quite general domain (for the same type of texts) are in Table 4.

	BLEU	NIST	WER	PER
SMT	9.51	3.73	83.94	66.09
morpheme-based SMT	8.98	3.87	80.18	63.88

Table 4. Evaluation for SMT systems

Details about the system and its evaluation can be consulted in (Díaz de Ilarraza et al., 2008).

4 Combining the approaches and evaluation

van Zaanen and Somers (2005) and Matusov et al. (2006) review a set of references about MEMT (Multi-engine MT) including the first attempt by Frederking and Nirenburg (1994), Macherey and Och (2007)

All those papers reach the same conclusion: combining the outputs results in a better translation.

Most of the approaches generate a new consensus translation using different language models. They have to train the system on those language models. Some of the approaches require confidence scores for each of the outputs. This approach is being used in several works (Macheret&Och, 2007; Sim et al., 2007), and some of them are used inside the GALE research program.

MEMT for Basque

Bearing in mind that huge parallel corpora for Basque are not available we decided to combine the different methods in a domain where translation memories were available.

Because confidence scores are not still available for the RBMT engine, we decided, for a first attempt, to combine the three approaches in a very simple hierarchical way: processing each sentence by the three engines (RBMT, EBMT and SMT) and then trying to choose the best translation among them.

In a first step the text is divided into sentences, then each sentence is processed using each engine (parallel processing is possible). Finally one of the translations is selected.

In order to make this selection the facts we can deal with are the followings:

- Precision for the EBMT approach is very high, but its coverage low.
- The SMT engine gives a confidence score.
- RBMT translations are more adequate for human postedition than those of the SMT engine, but SMT gets better scores when BLEU and NIST are used with only one reference (Labaka et al., 2007).

	BLEU RBMT	BLEU SMT	HTER RBMT	HTER SMT
EiTB corpus (news)	9.30	9.02	40.41	71.87
Consumer (magazine)	6.31	8.03	43.60	57.97

Table 5. Evaluation using Bleu and HTER for RBMT and SMT (Labaka et al., 2007)

We can see in Table 5 that automatic evaluation (BLEU) with one reference and user-driven evaluation (HTER) yield different results.

Bearing this in mind, in this first attempt, we decided to apply a hierarchical strategy:

- If the EBMT engine covers the sentence its translation is selected.
- Else we chose the translation from the SMT engine if its confidence score is higher than a given threshold.
- Otherwise the output from the RBMT engine will be taken.

The results on the development corpus appear in Table 6.

The best results, evaluated using automatic metrics with only one reference, are obtained combining EBMT and SMT. But bearing in mind our previous evaluation trials with human translators (Table 5), we think that a deeper evaluation is necessary.

Table 7 shows the results on the test corpora.

	Coverage	BLEU	NIST
RBMT (domain adapted)	100%	7.97	3.21
SMT	100%	14.37	4.43
EBMT+RBMT	EBMT 42% RBMT 58%	26.85	5.15
EBMT+SMT	EBMT 42% SMT 58%	30.44	5.93
EBMT+SMT+ RBMT	EBMT 42% SMT 33% RBMT 25%	29.41	5.68

Table 6. Results for the MEMT system using the development corpus

	Coverage	BLEU	NIST
RBMT (domain adapted)	100%	5.16	3.08
SMT	100%	12.71	4.69
EBMT+RBMT	EBMT 58% RBMT 42%	26.29	5.40
EBMT+SMT	EBMT 58% SMT 42%	29.11	6.25
EBMT+SMT+ RBMT	EBMT 58% SMT 28% RBMT 14%	28.50	6.02

Table 7. Results for the MEMT system using the test corpus

5 Conclusions

We have presented a hierarchical strategy to select the best output from three MT engines we have developed for Spanish-Basque translation.

In this first attempt, we decided to apply a hierarchical strategy: First application of EBMT (translation patterns), then SMT (if its confidence score is higher than a given threshold), and then RBMT.

The results of the initial automatic evaluation showed very significant improvements. For example, 129% relative increase for BLEU when comparing. EBMT+SMT combination with SMT single system. Or 124% relative increase for BLEU when comparing. EBMT+SMT+RBMT combination with SMT single system.

Anyway the best results, evaluated using automatic metrics with only one reference, are obtained combining just EBMT and SMT.

The consequence of the inclusion of a final RBMT engine (to translate just the sentences not covered by EBMT and with low confidence score for SMT) has a small negative contribution of 2% relative decrease for BLEU. But based on previous evaluations we think that a deeper evaluation based on human judgements is necessary.

For the near future we plan to carry out new experiments using combination of the outputs based on a language model. We are also plan defining confidence scores for the RBMT engine (penalties when suspicious or very complex syntactic structures are present in the analysis, penalties for high proportion of ignored word senses, promoting translations that recognize multiword lexical units, …)

Acknowledgments

This work has been partially funded by the Spanish of Education and Science (OpenMT: Open Source Machine Translation using hybrid methods,TIN2006-15307-C03-01) and the Local Government of the Basque Country (AnHITZ 2006: Language Technologies for Multingual Interaction in Intelligent Environments., IE06-185). Gorka Labaka is supported by a PhD grant from the Basque Government (grant code, BFI05.326)

Reference

Aduriz, I. and Díaz de Ilarraza, A. 2003. Morphosyntactic disambiguation ands shallow parsing in Computational Processing of Basque. In *Inquiries into the lexicon-syntax relations in Basque. Bernarrd Oyharabal (Ed.), Bilbao.*

Agirre, E., D´ de Ilarraza, A., Labaka, G., and Sarasola, K. (2006). Uso de información morfológica en el alineamiento Español-Euskara. In *XXII Congreso de la SEPLN.*

Alegria I., Artola X., Sarasola K. 1996. Automatic morphological analysis of Basque. *Literary & Linguistic Computing* Vol. 11, No. 4, 193-203. Oxford University Press. Oxford. 1996.

Alegria I., Gurrutxaga A., Saralegi X., Ugartetxea S. 2006. ELeXBi, A Basic Tool for Bilingual Term Extraction from Spanish-Basque Parallel Corpora. *Proc. of the 12th EURALEX International Congress*. pp 159-165

Alegria I., Díaz de Ilarraza A., Labaka G., Lersundi M., Mayor A., Sarasola K. 2007. Transfer-based MT from Spanish into Basque: reusability, standardization and open source. *LNCS 4394*. 374-384. Cicling 2007.

Carreras X.,, Chao I., Padró L., Padró M. 2004. FreeLing: An open source Suite of Language Analyzers, in P*roceedings of the 4th International Conference on Language Resources and Evaluation* (LREC'04).

Díaz de Ilarraza A., Labaka G., Sarasola K.. 2008. Spanish-Basque SMT system: statistical translation into an agglutinative language. (Submitted to LREC 2008)

Frederking R., Nirenburg S. 1994. Three heads are better than one. *Proc. of the fourth ANLP*. Stuttgart,

Giménez J., Amigó E., Hori C. 2005. Machine Translation Evaluation Inside QARLA. In *Proceedings of the International Workshop on Spoken Language Technology* (IWSLT'05)

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., Herbst, E. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Annual Meeting of the ACL*, Prague.

Koehn, P. 2005. Europarl: A parallel corpus for statistical machine translation. *MT Summit X.* Phuket.

Labaka G., Stroppa N., Way A., Sarasola K. 2007 Comparing Rule-Based and Data-Driven Approaches to Spanish-to-Basque Machine Translation *Proc. of MT-Summit XI,* Copenhagen

Macherey W., Och F, 2007. An Empirical Study on Computing Consensus Translations from Multiple Machine Translation Systems. *Proc. of the EMNLP and CONLL* 2007. Prague.

Martínez R., Abaitua J., Casillas A. Alingning Tagged Bitext. *Proceedings of the Sixth Workshop on Very Large Corpora,* 1998.

Mayor A. 2007. *Matxin: erregeletan oinarritutako itzulpen automatikoko sistema*. Ph. Thesis. Euskal Herriko Unibertsitatea.

Matusov E., Ueffing, N, Ney H. 2006. Computing Consensus Translation from Multiple Machine Translation Systems Using Enhanced Hypotheses Alignment. *Proc. of EACL 2006*, Trento.

McTait K. A Language-Neutral Sparse-Data 1999. Algorithm for Extracting Translation Patterns". *Proceedings of 8th International Conference on Theorical and Mathodological Issues in Machine Translation,.*

Och F. and Ney H. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics,* 29(1): 19–51.

Sim K., Byrne W., Gales M., Sahbi H. 2007., Woodland P. Consensus network decoding for statistical machine translation system combination. Proc. of ICASSP, 2007

Snover M., Dorr B., Schwartz R., Micciulla L., and Makhoul J.. 2006. A study of translation edit rate with targeted human annotation. In Proceedings of AMTA-2006, Cambridge, USA.

Stolcke, A. (2002). SRILM - An Extensible Language Modeling Toolkit. In P*roc. Intl. Conf. Spoken Language Processing*, Denver, Colorado.

Stroppa N., Groves D., Way A., Sarasola K. 2006.Example-Based Machine Translation of the Basque Language*. 7th conf. of the AMTA.*

van Zaanen M. and Somers H. 2005. DEMOCRAT: Deciding between Multiple Outputs Created by Automatic Translation. *MT Summit X*. Phuket.

Way A. and Gough N. 2005. Comparing Example-Based and Statistical Machine Translation. *Natural Language Engineering*, 11(3):295–309.

Statistical Post-Editing: A Valuable Method in Domain Adaptation of RBMT Systems for Less-Resourced Languages

A. Diaz de Ilarraza, G. Labaka, K. Sarasola
Ixa taldea
Informatika Fakultatea
University of the Basque Country
{jipdisaa, jiblaing, jipsagak}@ehu.es

Abstract

We present two experiments with Basque to verify the improvement obtained for other languages by using statistical post editing. The small size of available corpora and the use a morphological component in both RBMT and SMT translations make different our experiments from hose presented for similar works. Our results confirm the improvements when using a restricted domain, but they are doubtful for more general domains.

1 Introduction

Corpus based MT systems base their knowledge on aligned bilingual corpora, and the accuracy of their output depends heavily on the quality and the size of these corpora. When the two languages used in translation have very different structure and word order, the corpus needed to obtain similar results should be bigger.

Basque is a highly inflected language with free constituent order. Its structure and word order is different compared with languages as Spanish, French or English.

Being Basque a lesser used language, nowadays large and reliable bilingual corpora are unavailable. At present, domain specific translation memories for Basque are not bigger than two-three millions words, so they are still far away from the size of the corpora used for other languages; for example, Europarl corpus (Koehn, 2005), that is becoming a quite standard corpus resource, has 30 million words. So, although domain restricted corpus based MT for Basque shows promising results, it is still not ready for general use.

Moreover, the Spanish>Basque RBMT system Matxin's performance, after new improvements in 2007 (Alegria et al., 2007), is becoming useful for content assimilation, but it is still not suitable enough to allow unrestricted use for text dissemination.

Therefore, it is clear that we should experiment combining our basic approaches for MT (rule-based and corpus-based) to get a better performance. As the first steps on that way, we are experimenting with two simple alternative approaches to combining RBMT, SMT and EBMT:

- Selecting the best output in a multi engine system combining RBMT, EBMT and SMT approaches. (Alegría, et al., 2008)
- Statistical post-editing (SPE) on RBMT systems.

This paper deals with the second approach, where significant improvements have been recently published (Dugast et al., 2007; Ehara, 2007; Elming, 2006; Isabelle et al., 2007; Simard et al., 2007a and 2007b).

We don't have large corpus on post editing for Basque as proposed in (Isabelle et al., 2007), because our RBMT system has recently been created. However, we could manage to get parallel corpus on some domains with a few million of words,

We will show that the issue of domain adaptation of the MT systems for Basque can be performed via the serial combination of a vanilla RBMT system and a domain specific statistical post-editing system even when the training corpus is not very big (half a million words). Unfortunately, we could not show that RBMT +SPE combination improves the result of RBMT systems when the corpus used is not related to a restricted domain.

The rest of this paper is arranged as follows: In section 2, we position the present work with respect to our ongoing research on SMT and SPE. In section 3 we present the corpora that will be used in our experiments. Section 4 describes the basic RBMT and statistical translation systems. In section 5, we report on our experiments comparing translation results under a range of different MT conditions: SMT versus RBMT, RBMT+SPE versus RBMT, and RBMT+SPE versus SMT. We finish this paper with some conclusions and future work.

2 Related work

In the experiments related by (Simard et al., 2007a) and (Isabelle et al., 2007) SPE task is viewed as translation from the language of RBMT outputs into the language of their manually post-edited counterparts. So they don't use a parallel corpus created by human translation. Their RBMT system is SYSTRAN and their SMT system PORTAGE. (Simard et al., 2007a) reports a reduction in post-editing effort of up to a third when compared to the output of the rule-based system, i.e., the input to the SPE, and as much as 5 BLEU points improvement over the direct SMT approach. (Isabelle et al., 2007) concludes that such a RBMT +SPE system appears to be an excellent way to improve the output of a vanilla RBMT system and constitutes a worthwhile alternative to costly manual adaptation efforts for such systems. So a SPE system using a corpus with no more than 100.000 words of post-edited translations is enough to outperform an expensive lexicon enriched baseline RBMT system.

The same group recognizes (Simard et al., 2007b) that this sort of training data is seldom available, and they conclude that the training data for the post-editing component does not need to be manually post-edited translations, that can be generated even from standard parallel corpora. Their new RBMT+SPE system outperforms both the RBMT and SMT systems again. The experiments show that while post-editing is more effective when little training data is available, it remains competitive with SMT translation even when larger amounts of data. After a linguistic analysis they conclude that the main improvement is due to lexical selection.

In (Dugast et al., 2007), the authors of SYSTRAN's RBMT system present a huge improvement of the BLEU score for a SPE system when comparing to raw translation output. They get an improvement of around 10 BLEU points for German-English using the Europarl test set of WMT2007.

(Ehara, 2007) presents two experiments to compare RBMT and RBMT+SPE systems. Two different corpora are issued, one is the reference translation (PAJ, Patent Abstracts of Japan), the other is a large scaled target language corpus. In the former case, RBMT+SPE wins, in the later case RBMT wins. Evaluation is performed using NIST scores and a new evaluation measure NMG that counts the number of words in the longest sequence matched between the test sentence and the target language reference corpus.

Finally, (Elming, 2006) works in the more general field called as Automatic Post-Processing (APE). They use transformation-based learning (TBL), a learning algorithm for extracting rules to correct MT output by means of a post-processing module. The algorithm learns from a parallel corpus of MT output and human-corrected versions of this output. The machine translations are provided by a commercial MT system, PaTrans, which is based on Eurotra. Elming reports a 4.6 point increase in BLEU score.

3 The corpora

Our aim was to improve the precision of the MT system trying to translate texts from a restricted domain. We were interested in a kind of domain where a formal and quite controlled language would be used and where any public organization or private company would be interested in automatic translation on this domain. We also wanted to compare the results between the restricted domain and a more general domain such as news.

Specific domain: Labor Agreements Corpus

The domain related to *Labor Agreements* was selected. The Basque Institute of Public Administration (IVAP[1]) collaborated with us in this selection, after examining some domains, available parallel corpora and their translation

[1] http://www.ivap.euskadi.net

needs. The Labor Agreements Corpus is a bilingual parallel corpus (Basque and Spanish) with 640,764 words for Basque and 920,251 for Spanish. We automatically aligned it at sentence level and then manual revision was performed.

To build the test corpus the full text of several labor agreements was randomly chosen. We chose full texts because we wanted to ensure that several significant but short elements as the header or the footer of those agreements would be represented. Besides it is important to measure the coverage and precision we get when translating the whole text in one agreement document and not only those of parts of it. System developers are not allowed to see the test corpus.

In SMT we use the training corpus to learn the models (translation and language model); the development corpus to tune the parameters; and the test corpus to evaluate the system.

The size of each subset is shown in Table 1.

		Sentences	Words
Training	Spanish	51,740	839.393
	Basque		585,361
Development	Spanish	2,366	41,508
	Basque		28,189
Test	Spanish	1,945	39,350
	Basque		27,214

Table 1. Statistics of Labor Agreements Corpus

General domain: Consumer Eroski Corpus

As general domain corpus, we used the *Consumer Eroski* parallel corpus. The *Consumer Eroski* parallel corpus is a collection of 1,036 articles written in Spanish (January 1998 to May 2005, Consumer Eroski magazine, http://revista.consumer.es) along with their Basque, Catalan, and Galician translations. It contains more than one million Spanish words for Spanish and more than 800,000 Basque words. This corpus is aligned at sentence level.

In order to train the data-driven systems (both SMT and SPE systems), we used approximately 55,000 aligned sentences extracted from the Consumer dataset. Two additional sentence sets are used; 1501 sentences for parameter tuning and 1515 sentences for evaluation (see Table 2).

		Sentences	Words
Training	Spanish	54,661	1,056,864
	Basque		824350
Development	Spanish	1,501	34,333
	Basque		27,235
Test	Spanish	1,515	32,820
	Basque		34,333

Table 2. Statistics of Consumer Eroski corpus

4 Basic translation systems

Rule based system: Matxin

In this subsection we present the main architecture of an open source MT engine, named *Matxin* (Alegria et al., 2007). the first implementation of Matxin translates from Spanish into Basque using the traditional transfer model and based on shallow and dependency parsing.

Matxin is a classical transfer system consisting of three main components: (i) analysis of the source language into a dependency tree structure, (ii) transfer from the source language dependency tree to a target language dependency structure, and (iii) generation of the output translation from the target dependency structure. These three components are described in more detail in what follows.

The analysis of the Spanish source sentences into dependency trees is performed using an adapted version of the Freeling toolkit (Carreras et al., 2004). The shallow parser provided by Freeling is augmented with dependency information between chunks.

In the transfer module the Spanish analysis tree is transformed into Basque dependency tree. In this step, a very simple lexical selection is carried out, the Spanish lemma is translated by most frequent equivalent.

Finally, the dependency tree coming from the transfer module is passed on the generation module, in order to get the target language sentence. The order of the words in the final sentence is decided and morphological generation is carried out when it is necessary (in Basque: the declension case, the article and other features are added to the whole noun phrase at the end of the last word). We reused a previous morphological analyzer/generator developed for Basque (Alegria et al., 1996) adapted and transformed to our purposes.

Corpus based system

The corpus-based approach has been carried out in collaboration with the National Center for Language Technology in Dublin City University (DCU).

The system is based on a baseline phrase-based SMT system, but the dataset of aligned phrases is enriched with linguistically motivated phrase alignments. We have carried out Basque to English (Stroppa et al., 2006) and Spanish to Basque (Labaka et al., 2007) translation experiments.

Freely available tools are used to develop the SMT systems:

- GIZA++ toolkit (Och and H. Ney, 2003) is used for training the word/morpheme alignment.
- SRILM toolkit (Stolcke, 2002) is used for building the language model.
- Moses Decoder (Koehn et al., 2007) is used for translating the sentences.

Due to the morphological richness of Basque, when translating from Spanish to Basque some Spanish words, like prepositions or articles, correspond to Basque suffixes, and, in case of ellipsis, more than one of those suffix can be added to the same word. Example of concatenation of two case suffixes:

```
  puntuarenean =
= puntu  +  aren    +  ean  =
= point  +  of the  +  in the  =
= in the one(ellipsis) of the point
```

In order to deal with these features a morpheme-based SMT system was developed.

Adapting the SMT system to work at the morpheme level consists on training the basic SMT on the segmented text. The system trained on these data will generate a sequence of morphemes as output. In order to obtain the final Basque text, we have to generate words from those morphemes.

To get the segmented text, Basque texts are previously analyzed using Eustagger (Aduriz & Díaz de Ilarraza, 2003). After this process, each word is replaced with the corresponding lemma followed by a list of morphological tags. The segmentation is based on the strategy proposed on (Agirre et al., 2006).

Both systems (the conventional SMT system and the morpheme based), were optimized decoding parameters using a Minimum Error Rate Training. The metric used to carry out the optimization is BLEU.

The evaluation results for the general domain Consumer corpus (also used in this paper) are in Table 3. The morpheme based MT system gets better results for all the measures except BLEU.

	BLEU	NIST	WER	PER
SMT	**9.85**	4,28	82,72	63,78
Morpheme-based SMT	9,63	**4,43**	**80.92**	**62,27**

Table 3. Evaluation for SMT systems

RBMT and Statistical Post-Editing

In order to carry out experiments with statistical post-editing, we have first translated Spanish sentences in the parallel corpus using our rule-based translator (Matxin). Using these automatically translated sentences and their corresponding Basque sentences in the parallel corpus, we have built a new parallel corpus to be used in training our statistical post-editor.

The statistical post-editor is the same corpus-based system explained before. This system is based on freely available tools but enhanced in two main ways:

- In order to deal morphological richness of Basque, the system works on morpheme-level, so a generation phase is necessary after SPE is applied.
- Following the work did in collaboration with the DCU, the phrases statistically extracted are enriched with linguistically motivated chunk alignments.

5 Results

We used automatic evaluation metrics to assess the quality of the translation obtained using each system. For each system, we calculated BLEU (Papineni et al., 2002), NIST (Doddington, 2002), Word Error Rate (WER) and Position independent Error Rate (PER).

Besides, our aim was to evaluate performance using different corpora types, so we tested the output of all systems applied to two corpora: one domain specific (Labor Agreements Corpus), and a general domain corpus (Consumer corpus).

	BLEU	NIST	WER	PER
Rule-based	4,27	2,76	89,17	74,18
Corpus-based	12,27	4,63	77,44	58,17
Rule-based + SPE	**17,11**	**5,01**	**75,53**	**57,24**

Table 4. Evaluation on domain specific corpus

Results obtained on the Labor Agreements Corpus (see Table 4) shows that the rule-based gets a very low performance (rule-based system is not adapted to the restricted domain), and the corpus-based system gets a much higher score (8 BLEU points higher, a 200% relative improvement). But if we combine both systems using the corpus-based system as a statistical post-editor, the improvement is even higher outperforming corpus-based system in 4.48 BLEU point (40% relative improvement).

	BLEU	NIST	WER	PER
Rule-based MT	6,78	3,72	81,89	66,72
Corpus-based MT	**9,63**	**4,43**	80,92	**62,27**
Rule-based + SPE	8,93	4,23	**80,34**	63,49

Table 5. Evaluation on general domain corpus

Otherwise, results on the general domain corpus (see Table 5) do not indicate the same. Being a general domain corpus, the vanilla rule-based system gets better results, and those approaches based on the corpus (corpus-based MT and RBMT +SPE) get lower ones. Furthermore, the improvement achieved by the statistical post-editor over the rule-based system is much smaller and it does not outperforms the corpus-based translator.

6 Conclusion

We performed two experiments to verify the improvement obtained for other languages by using statistical post editing. Our experiments differ from other similar works because we use a morphological component in both RBMT and SMT translations, and because the size of the available corpora is small.

Our results are coherent with huge improvements when using a RBMT+SPE approach on a restricted domain presented by (Dugast eta al., 2007; Ehara, 2007; Simard et al., 2007b). We obtain 200% improvement in the BLEU score for a RBMT+SPE system working with Matxin RBMT system, when comparing to raw translation output, and 40% when comparing to SMT system.

Our results also are coherent with a smaller improvement when using more general corpora as presented by (Ehara, 2007; Simard et al., 2007b).

We can not work with manually post-edited corpora as (Simard et al., 2007a) and (Isabelle et al., 2007) because there is no such a big corpus for Basque, but we plan to collect it and compare results obtained using a real post-edition corpus and the results presented here.

We also plan automatic extracting rules to correct MT output by means of a post-processing module (Elming, 2006).

Acknowledgments

This work has been partially funded by the Spanish Ministry of Education and Science (OpenMT: Open Source Machine Translation using hybrid methods,TIN2006-15307-C03-01) and the Local Government of the Basque Country (AnHITZ 2006: Language Technologies for Multilingual Interaction in Intelligent Environments., IE06-185). Gorka Labaka is supported by a PhD grant from the Basque Government (grant code, BFI05.326).

References

Aduriz, I. and Díaz de Ilarraza, A. (2003). Morphosyntactic disambiguation and shallow parsing in Computational Processing of Basque. In Inquiries into the lexicon-syntax relations in Basque. Bernard Oyharçabal (Ed.), Bilbao.

Alegria, I., Artola Zubillaga, X. and Sarasola, X. (1996). Automatic morphological analysis of Basque. Literary & Linguistic Computing 11(4):193—203.

Alegria I., Díaz de Ilarraza A., Labaka G., Lersundi M., Mayor A., Sarasola K. (2007) Transfer-based MT from Spanish into Basque: reusability, standardization and open source. Cicling.

Alegria, I., Casillas, A., Díaz de Ilarraza, A., Igartua, J., Labaka, G., Lersundi, M., Mayor, A., Sarasola, K., Saralegi, X., Laskurain, B. (2008). A Simple Mixing Approach to MT for Basque. To be presented in MATMT08 workshop: Mixing Approaches to Machine Translation. Donostia.

Agirre, E., Díaz de Ilarraza, A., Labaka, G., and Sarasola, K. (2006). Uso de información morfológica en el alineamiento Español-Euskara. In XXII congreso de la SEPLN, Zaragoza, Spain.

Carreras, X., Chao, I., Padró, L., Padró, M. (2004). FreeLing: An Open-Source Suite of Language Analyzers. In Proceedings of 4th LREC, Lisbon, Portugal.

Doddington, G. (2002). Automatic evaluation of Machine Translation quality using n-gram cooccurrence statistics. In Proceedings of HLT 2002, San Diego, CA.

Dugast, L., Senellart, J., & Koehn, P. (2007). Statistical post-editing on SYSTRAN's rule-based translation system. ACL 2007: proceedings of the Second Workshop on Statistical Machine Translation, 23, 2007, Prague, Czech Republic; pp. 220-223

Elming, J. (2006). Transformation-based correction of rule-based MT. 11th Annual Conference of the European Association for Machine Translation, Oslo, Norway.

Isabelle, P., Goutte, C., & Simard, M. (2007). Domain adaptation of MT systems through automatic post-editing. MT Summit XI, 10-14 September 2007, Copenhagen, Denmark. pp.255-261

Koehn, Ph. (2005). Europarl: A parallel corpus for statistical machine translation. Proc. of the MT Summit X ,pp. 79–86, September.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan N., Shen, W. Moran, C. Zens, R. Dyer, C. Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: Open Source Toolkit for Statistical Machine Translation. In Annual Meeting of the Association for Computational Linguistics (ACL), Prague, Czech Republic.

Labaka, G., Stroppa, N., Way, A. and Sarasola, K. (2007). Comparing Rule-based and Data-driven Approaches to Spanish-to-Basque Machine Translation. In Proceedings of the MT-Summit XI, Copenhagen, Denmark.

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

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

Simard, M., Goutte, C., and Isabelle, P.. (2007a). Statistical phrase-based post-editing. In Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference, pages 508–515, Rochester, USA, April. Association for Computational Linguistics.

Simard, M., Ueffing, N., Isabelle, P., & Kuhn, R. (2007b). Rule-based translation with statistical phrase-based post-editing. ACL 2007: proceedings of the Second Workshop on Statistical Machine Translation, June 23, 2007, Prague, Czech Republic; pp. 203-206

Stolcke, A. (2002). SRILM - An Extensible Language Modeling Toolkit. In Proc. Intl. Conf. Spoken Language Processing, Denver, Colorado.

Stroppa, N., Groves, D., Way, A., and Sarasola, K. (2006). Example-base Machine Translation of the Basque Language. In Proceedings of AMTA 2006, pp. 232—241, Cambridge, MA.

Terumasa E. (2007). Rule based machine translation combined with statistical post editor for Japanese to English patent translation. MT Summit XI Workshop on patent translation, 11 September 2007, Copenhagen, Denmark; pp.13-18.

From free shallow monolingual resources to machine translation systems: easing the task

Helena M. Caseli
NILC – ICMC
University of São Paulo
helename@icmc.usp.br

Maria das Graças V. Nunes
NILC – ICMC
University of São Paulo
gracan@icmc.usp.br

Mikel L. Forcada
DLSI
Universitat d'Alacant
mlf@ua.es

Abstract

The availability of machine-readable bilingual linguistic resources is crucial not only for machine translation but also for other applications such as cross-lingual information retrieval. However, the building of such resources demands extensive manual work. This paper describes a methodology to build automatically bilingual dictionaries and transfer rules by extracting knowledge from word-aligned parallel corpora processed with free shallow monolingual resources (morphological analysers and part-of-speech taggers). Experiments for Brazilian Portuguese–Spanish and Brazilian Portuguese–English parallel texts have shown promising results.

1 Introduction

Two of the main challenges in natural language processing (NLP) are (1) the production, maintenance and extension of computational linguistic resources and (2) the integration of these resources into NLP applications.

In particular, the availability of machine-readable bilingual linguistic resources is crucial not only for rule-based machine translation (RBMT) but also for other applications such as cross-lingual information retrieval. However, the building of such resources (bilingual single-word and multi-word correspondences, translation rules) demands extensive manual work. As a consequence, bilingual resources are usually more difficult to find than shallow monolingual resources such as morphological dictionaries or part-of-speech taggers.

In an attempt to overcome the lack of these bilingual resources, several methods have been proposed to build automatically translation grammars (McTait, 2003; Menezes and Richardson, 2001; Lavie et al., 2004; Carbonell et al., 2002) and bilingual dictionaries (Wu and Xia, 1994; Fung, 1995; Koehn and Knight, 2002; Langlais et al., 2001; Schafer and Yarowsky, 2002).

In line with some of these initiatives, this paper describes a methodology to build automatically both bilingual dictionaries and shallow-transfer rules. These resources are built by extracting knowledge from automatically word-aligned (or *lexically aligned*) parallel corpora which have been processed with shallow monolingual resources (morphological analysers and part-of-speech taggers). The free shallow monolingual resources used in these experiments are available as part of the `Apertium` open-source machine translation (MT) platform.[1]

This methodology is part of the ReTraTos project[2] which aims at inducing linguistic knowledge for Brazilian Portuguese (`pt`), Spanish (`es`) and English (`en`). The MT ex-

[1] `http://www.apertium.org`.

[2] `http://www.nilc.icmc.usp.br/nilc/projects/retratos.htm`

periments carried out for the pt–es and pt–en language pairs produced reasonable results as will be shown here.

There is a distinct advantage in the method proposed in this paper, as compared to other learning approaches to MT (such as statistical machine translation, SMT). It generates dictionaries and rules which may be edited by human experts to improve the performance of the resulting system, or even combined with data written by experts. In particular, there is an ongoing project to convert the data generated by our method to be freely used with Apertium. The induction software will also be distributed as open-source in a near future.

The main contribution of the proposed methods is a way to induce bilingual resources automatically from a parallel corpus using free monolingual resources and tools.

This paper is organized as follows. Section 2 presents related work on automatic induction of bilingual dictionaries and transfer rules. The proposed methods for inducing bilingual dictionaries and transfer rules are described in section 3. The experiments carried out with the pt–es and pt–en language pairs are described in section 4. The paper ends with some conclusions and proposals for future work (section 5).

2 Related work

In this section we present methods to induce automatically bilingual dictionaries (section 2.1) and transfer rules (section 2.2).

2.1 Induction of bilingual dictionaries

A bilingual dictionary —a bilingual list of words and multiword units that are mutual translations— is usually a by-product of a word alignment process (Brown et al., 1993; Och and Ney, 2000; Caseli et al., 2005).[3]

In (Wu and Xia, 1994), an English–Chinese dictionary was automatically induced by means of training a variant of the statistical model described in (Brown et al., 1993). This model was trained on a large corpus (about 3 million words) resulting in a set of about 6,500 English words (on average 2.33 possible Chinese translations for each English word). Evaluation through direct human inspection of a random set of 200 words showed an accuracy lying between 86.0% (completely automatic process) and 95.1% (manual correction).

By contrast, the method proposed by Fung (1995) uses a non-aligned Chinese–English parallel corpus (with about 5,760 English words) to induce bilingual entries for nouns and proper nouns based on co-occurrence (source and target) positions. Three judges evaluated 23.8% of the induced entries and the average accuracy was 73.1%.

This paper proposes a bilingual dictionary induction method based on automatic word alignment as explained in section 3.1.

2.2 Induction of translation rules

In the literature, methods for inducing transfer rules are based on many different approaches. However, all of them get a sentence-aligned parallel corpus (a set of translation examples) as input. The induced rules can, in turn, be used by the MT system to translate source sentences into target sentences.

The method proposed in (McTait, 2003) looks for transfer rules in two steps. In a monolingual step, the method looks for sequences of items that occur at least in two sentences by processing each side (source or target) separately —these sequences are taken as monolingual patterns. In the bilingual step, the method builds bilingual patterns following a co-occurrence criterion.[4] Finally, a bilingual similarity (distance) measure is used to set the alignment between source and target items that form a bilingual pattern. This method achieved 33.9% coverage, considering only full translations, in experiments with a training corpus of 2,500 and a test corpus of 1,000 pairs of en–fr (French) sentences.

The method proposed in (Menezes and Richardson, 2001) aligns the nodes of the

[3] An automatic word aligner is a tool for finding correspondences between words, and sometimes multiword units, in parallel texts.

[4] One source pattern and one target pattern occurring in the same pair of sentences are taken to be mutual translations.

source and target parse trees by looking for word correspondences in a bilingual dictionary. Then, following a best-first strategy (processing first the nodes with the best word correspondences), the method aligns the remaining nodes using a manually created alignment grammar composed of 18 bilingual compositional rules. After finding alignments between nodes of both parse trees, these alignments are expanded using linguistic constructs (such as noun and verb phrases) as context boundaries. Menezes and Richardson (2001) show that their system performed better than BabelFish[5] in 46.5% of test cases in experiments carried out with a training corpus of 161,606 and test corpora of 200-500 pairs of `es`–`en` sentences.

In (Lavie et al., 2004; Carbonell et al., 2002), the method infers hierarchical syntactic transfer rules, initially, on the basis of the constituents of both (manually) word-aligned languages. To do so, sentences from the language with more resources (English, in that case) are parsed and disambiguated. Value and agreement constraints[6] are determined from the syntactic structure, the word alignments and the source and target dictionaries. Lavie et al. (2004) show experiments carried out with RBMT and SMT systems trained with 17,589 lexically aligned sentences and phrases and tested with 258 sentences, for Hindi(`hi`)–`en`. The results show that the RBMT system scored better than the SMT: 11.2 BLEU and 5.32 NIST vs. 10.2 BLEU and 4.70 NIST.

Sánchez-Martínez and Forcada (2007) use an aligned parallel corpus to infer shallow-transfer rules based on the alignment templates approach by Och and Ney (2004). This research makes extensive use of the information in an existing manually-built bilingual dictionary to guide rule extraction. A training corpus composed of 100,834 pairs of `es`–`ca` (Catalan) sentences and a test corpus of about 10,000 words were used to evaluate the induced rules. The evaluation carried out via post-edited translation shows a word error rate (WER) of 8.1–8.5% for automatically induced rules vs. 6.5–6.7% for hand-coded rules.

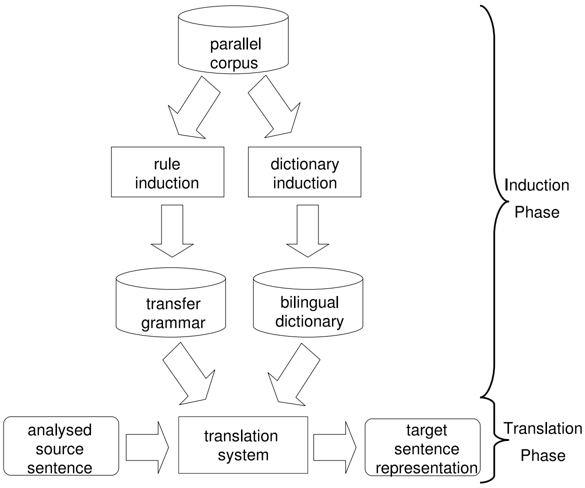

Figure 1: Scheme of proposed induction and translation systems

The method to induce transfer rules presented in this paper brings forth a new approach to induce and filter rules as described in section 3.2.

3 Induction and translation in the ReTraTos environment

The general scheme of the proposed induction and translation systems is shown in Figure 1. A PoS-tagged and word-aligned parallel corpus is given as input to our bilingual dictionary and transfer rule induction systems.

The induced sets of transfer rules (transfer grammar) and bilingual entries (bilingual dictionary) are then used by a shallow-transfer MT system to translate source sentences into target sentences.

The induction systems are introduced in the next two sections while the translation system is described in section 3.3.

[5] `http://babelfish.altavista.com`.

[6] Value and agreement constraints specify which values (value constraints) the morphological features of source and target words should have (for instance, masculine as gender, singular as number and so on) and whether these values should be the same (agreement constraints).

3.1 Inducing the bilingual dictionary

A brief description of the bilingual dictionary induction process is presented in this section. For a more complete description see Caseli and Nunes (2007).

The bilingual dictionary induction process comprises the following steps: (1) the compilation of two bilingual dictionaries, one for each translation direction (one source–target and another target–source); (2) the merging of these two dictionaries; (3) the generalization of morphological attribute values in the bilingual entries; and (4) the treatment of morphosyntactic differences related to entries in which the value of the target gender/number attribute has to be determined from information that goes beyond the scope of the entry itself.[7]

3.2 Inducing the transfer rules

The transfer rule induction process is briefly described in this section and in detail in Caseli et al. (2008).

In contrast with other rule induction methods, our method follows an *alignment block* based approach. Specifically, it does not learn rules from the whole pairs (source language, target language) of aligned sentences but from sequences of contiguous word-aligned items[8] in these pairs: the alignment blocks.

Figure 2 shows the three types of alignment blocks considered in this approach: omissions (type 0), alignments preserving item order in sentence (type 1) and reorderings (type 2). In this figure, source and target items are accompanied by their positions in the source and target sentences. For example, the source items *a* and *b* are aligned to *a'*, *a"* and *b'* in a way that preserves item order; therefore, they form an alignment block of type 1. Furthermore, they are also part of an alignment block of type 2, since the source item *c* has a cross-link to *c'*.[9]

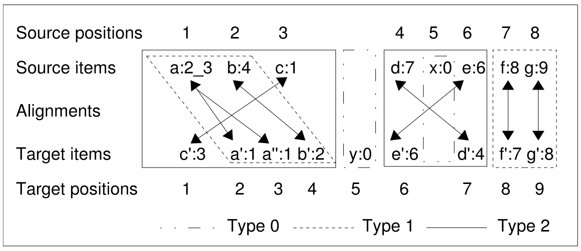

Figure 2: Types of alignment blocks

After building these alignment blocks, the rules are induced from each type separately, following four phases: (1) pattern identification, (2) rule generation, (3) rule filtering and (4) rule ordering.

First, analogously to (McTait, 2003), the bilingual patterns are extracted in two steps: monolingual and bilingual. In the monolingual step, source patterns are identified by an algorithm based on the Sequential Pattern Mining (SPM) technique and the `PrefixSpan` algorithm (Pei et al., 2004). In the bilingual step, the target items aligned to each source pattern are examined (in the parallel translation example) to form the bilingual pattern.

In pattern identification, the frequency threshold necessary to call a sequence of items a pattern is different for each type of alignment block (0, 1 or 2). This frequency threshold is calculated as a percentage *p* (an input parameter) of the total amount of blocks of each type.[10] The idea behind alignment-block-guided induction is that if we were using the same absolute frequency threshold for all types of alignments, very few relevant patterns coming from less frequent alignment types would be identified.

Second, the rule generation phase encompasses the building and the generalization of constraints between values on one (monolingual) or both (bilingual) sides of a bilingual

[7] For example, the **es** noun *tesis* (thesis) is valid for both number (singular and plural) and it has two possible **pt** translations: *tese* (singular) and *teses* (plural).

[8] For example, o/o<**det**><**def**><**m**>:5 is an item found in **pt** sentences where *o* (the) is the original word; o<**det**><**def**><**m**> is its lemma, PoS and morphological features; and 5 is the position of the word aligned with it. For details on how these information are obtained see section 4.1.

[9] Only alignment blocks of type 2 can include other alignment blocks (types 0 and 1).

[10] For example, suppose that we have 10,000 alignment blocks of type 1, 100 of type 2 and an input percentage of 15%. So, the frequency threshold for identifying patterns of type 1 is 1,500 while it is only 15 for patterns of type 2.

pattern. Constraints are derived from feature values (morphological information) in translation examples. Two kinds of constraints can be built —value constraints and agreement/value constraints— as in (Carbonell et al., 2002).[11]

Third, the induced rules are filtered to solve ambiguities. For all ambiguous rules —those with the same source side (sequence of source PoS tags) but different target sides— the filtering approach looks for source feature and lexical values which can distinguish the ambiguous rules.

Finally, the rule ordering specifies the order in which transfer rules should be applied by the MT system. It is done implicitly by setting the frequency and weight (the probability of its occurrence) of each rule, each target side and each constraint set.[12]

3.3 Translating sentences

The induced resources are used in the MT task by means of a simple translation system (see figure 1). The input of this system is an already analysed source sentence, that is, a sequence of source lexical forms (each one consisting of lemma, PoS tag and morphological inflection attributes).

The MT system implemented has two modes of translation: word-by-word and transfer. The former uses only the induced bilingual dictionary, while the latter uses both the induced dictionary and transfer rules.

In transfer mode, the system chooses and applies the best suitable transfer rules following a left-to-right longest-match procedure. The "best suitable rule" is the most frequent rule which: (i) matches the source sequence, (ii) matches a set of source constraints (there can be more than one) and (iii) this source constraint set is the most frequent.

Unlike in `Apertium`, a backtracking approach is used in transfer translation: if a source pattern *abcd* matches the input sentence but cannot be applied because it has no compatible constraint, the system will try to apply the sub-pattern *abc*. This backtracking goes on until the sub-pattern has just one item and, in this case, word-by-word translation is applied.

4 Experiments and results

The next sections describe the corpora used to induce the linguistic resources (4.1) and the evaluation settings and results (4.2).

4.1 Preprocessing of bilingual corpora

The experiments described in this paper were carried out using two training parallel corpora. One corpus consists of 18,236 pairs of `pt`–`es` parallel sentences with 503,596 tokens in `pt` and 545,866 tokens in `es`. The other corpus consists of 17,397 pairs of `pt`–`en` parallel sentences with 494,391 tokens in `pt` and 532,121 tokens in `en`. Both corpora contain articles from the online version of a Brazilian scientific magazine, *Pesquisa FAPESP*.[13] It contains parallel texts written in `pt` (original), `en` (version) and `es` (version).

These corpora were PoS-tagged using the morphological analyser and the PoS tagger available in `Apertium` (Armentano-Oller et al., 2006). The morphological analysis provides one or more lexical forms for each surface form (the form as it appears in the text) using a monolingual morphological dictionary. The PoS tagger chooses the best possible lexical form based on a first-order hidden Markov model (HMM).

The original morphological dictionaries available at `Apertium` were enlarged in the ReTraTos project with entries from `Unitex`[14] (`pt` and `en`) and from the `es`–`ca` MT system `InterNOSTRUM`[15] (`es`).[16] So, the original morphological dictionaries for `pt` and `es` available

[11] The value constraints here are the same as (Carbonell et al., 2002), but the agreement/value constraints are quite different from the agreement constraints used by them since, here, the morphological value is also specified in agreement/value constraint.

[12] The weight of a rule is calculated as its frequency divided by the total frequency of the whole set of rules. The weight of each target side and each constraint set are calculated in a similar way.

[13] `http://revistapesquisa.fapesp.br`.

[14] `http://www-igm.univ-mlv.fr/~unitex/`.

[15] `http://www.internostrum.com/`.

[16] The `es` new entries were provided by the Transducens group from the Universitat d'Alacant.

in the Apertium es–pt linguistic data package (version 0.9) were enlarged to cover 1,136,536 and 337,861 surface forms, respectively. The en morphological dictionary available in the Apertium en–ca linguistic data package (version 0.8) was enlarged to cover 61,601 surface forms.[17]

After PoS tagging, the translation examples were word-aligned using two different tools: LIHLA (Caseli et al., 2005) and GIZA++ (Och and Ney, 2000). Experiments have shown that LIHLA had a better alignment error rate (AER) performance than GIZA++ on pt–es parallel texts (5.39% AER vs. 6.35% AER). But GIZA++ had a better performance on pt–en (15.44% AER vs. 8.61% AER) (Caseli et al., 2008). The translation examples were aligned in both directions (source–target and target–source) and the alignments were merged using the union algorithm proposed by Och and Ney (2003).

4.2 Evaluation settings and results

The linguistic resources induced from the two parallel corpora described in section 4.1 were: one bilingual dictionary for each pair of languages and some sets of transfer rules induced using different input parameters.

The induced bilingual dictionaries have 23,450 pt–es entries and 19,191 pt–en entries. The best set of transfer rules was obtained using a percentage $p = 0.07\%$ to calculate the frequency thresholds for pattern identification of each block type. With this parameter, 1,421 pt–es transfer rules, 1,329 es–pt transfer rules, 647 pt–en transfer rules and 722 en–pt transfer rules were induced.

The corpus used to test/evaluate the induced resources consists of 649 parallel sentences from the same domain of the training corpus. The sentences in the test corpus were translated in the four possible directions (pt–es, es–pt, pt–en and en–pt). To evaluate the translations, a reference corpus was created consisting of the corresponding parallel sentences in the test corpus. For example, the reference corpus used to evaluate the translation from pt to es is composed by the es sentences in the test corpus.

The sentences translated automatically were compared automatically with those in the reference corpus by means of the indirect scores BLEU (Papineni et al., 2002) and NIST (Doddington, 2002).

In these experiments, we evaluated the sentences translated by the ReTraTos MT system (see section 3.3) by applying the induced resources in the word-by-word translation (RTT_word-by-word) and in the transfer translation (RTT_transfer). The word-by-word translation was used here with three purposes: (1) to be a baseline for comparison with other systems, (2) to evaluate the quality of the induced vocabulary, and (3) to measure the improvement brought by using transfer rules (RTT_transfer).

We also evaluated translations produced by other MT systems available for the studied languages. For pt–es–pt, we have used two versions of the es–pt data provided in the open-source MT platform Apertium: version 0.9.1, which will be called Apertium and version 0.9.2, using a larger dictionary, which will be called Apertium-P.[18] For pt–en–pt, we have used the MT systems: FreeTranslation,[19] Google[20] and BabelFish.

Table 1 shows the results of pt–es–pt translation. From these values, it is possible to notice that the ReTraTos MT system using only one (RTT_word-by-word) or both (RTT_transfer) the induced linguistic resources obtained scores that were slightly higher than Apertium's versions, with a more significant difference in the es–pt direction.

In the pt–es direction, when compared to Apertium-P, the RTT_transfer had an improvement of around 2 points in BLEU and 0.2 in NIST; while in the es–pt direction, this improvement was twice as large: 4 points in BLEU and 0.4 in NIST.

[17] Initially the pt, es and en morphological dictionaries covered 128,772, 116,804 and 48,759 surface forms, respectively.

[18] Version 0.9.2. was the one that could be tried online in April 2007 at http://xixona.dlsi.ua.es/prototype.

[19] http://www.freetranslation.com.

[20] http://translate.google.com.

Table 1: Evaluation of pt–es–pt MT

Lang.	System	BLEU	NIST
pt–es	RTT_transfer	65.13	10.85
	RTT_word-by-word	64.90	10.82
	Apertium	63.82	10.64
	Apertium-P	63.87	10.64
es–pt	RTT_transfer	66.66	10.98
	RTT_word-by-word	66.49	10.95
	Apertium	60.98	10.31
	Apertium-P	62.88	10.51

Table 2: Evaluation of pt–en–pt MT

Lang.	System	BLEU	NIST
pt–en	RTT_transfer	28.32	7.09
	RTT_word-by-word	26.06	6.77
	FreeTranslation	32.94	7.65
	BabelFish	31.61	7.46
	Google	32.95	7.61
en–pt	RTT_transfer	24.00	6.11
	RTT_word-by-word	23.24	6.02
	FreeTranslation	30.53	6.85
	BabelFish	36.66	7.68
	Google	31.21	6.88

The similar scores of the two versions of ReTraTos on pt–es–pt seem to be due to the greater coverage of the induced bilingual dictionary on the texts of the domain. This fact indicates that, for related languages such as pt and es, a greater coverage of the bilingual dictionary has a stronger impact in translation scores than the transfer rules.

Table 2 shows the results of pt–en–pt translation. In the evaluation for this pair of languages, the translations produced by the ReTraTos versions did not score so high as those for the pt–es pair. This result was already expected, since the transfer rule induction system was not designed to deal with more complex changes in the structure of translation, but simply agreement and position changes between close items.

However, it is worth noticing that the improvement attributed to the use of rules (RTT_transfer) compared to the word-by-word (RTT_word-by-word) translation in the pt–en–pt pair is greater (0.76–2.26 BLEU points and 0.09–0.32 NIST points) than in the pt–es–pt pair (less than 0.3 points in BLEU and 0.03 in NIST). This indicates that, albeit simple (in the sense that they perform only shallow changes), the induced rules may indeed improve word-by-word translation between more distant languages.

5 Conclusions and future work

In this paper we have described a methodology to build bilingual dictionaries and translation rules automatically from parallel corpora. The input corpora were processed using word aligners and shallow monolingual resources such as morphological analysers and PoS taggers.

One advantage of the method proposed here is that both the inferred dictionaries and the induced rules are written in formats that can be easily edited by humans or combined with manually written rules.

In particular, the rules can be easily converted to the formats used by the open-source MT platform Apertium, and the bilingual dictionary entries are already induced in the formalism used by Apertium. Thus, the induction systems presented in this paper can be used along with the tools and linguistic data distributed with Apertium to ease the task of building new MT systems.

As future work, we intend to finish an ongoing project to adapt the induced resources to Apertium and to implement a *open-source toolchain* for MT. This toolchain will join the already existing free resources from Apertium and from ReTraTos and make them freely available to produce new MT systems.

Other future work includes an evaluation by means of the WER using post-edited output as a reference. We also aim at testing different configurations of ReTraTos to determine to what extent changes in optional modules (rule filtering and rule ordering) affect translation quality. Experiments to compare the performance of the system presented here (using automatically induced transfer rules) and that of a SMT system trained and tested on the same corpora are already been carried out.

Acknowledgements

We thank the financial support of the Brazilian agencies FAPESP (02/13207-8 and 04/06707-0), CAPES (053804-3 and 407507-

2) and CNPq (550388/2005-2), and of the Spanish Ministry of Education and Science; in particular through MEC-CAPES project PHB2005-0052 (Spain) and 116/06 (Brazil).

References

C. Armentano-Oller, R. C. Carrasco, A. M. Corbí-Bellot, M. L. Forcada, M. Ginestí-Rosell, S. Ortiz-Rojas, J. A. Pérez-Ortiz, G. Ramírez-Sánchez, F. Sánchez-Martínez, and M. A. Scalco. 2006. Open-source Portuguese-Spanish machine translation. In *Proceedings of the VII PROPOR*, pages 50–59, Itatiaia-RJ, Brazil.

P. Brown, V. Della-Pietra, S. Della-Pietra, and R. Mercer. 1993. The mathematics of statistical machine translation: parameter estimation. *Computational Linguistics*, 19(2):263–312.

J. Carbonell, K. Probst, E. Peterson, C. Monson, A. Lavie, R. Brown, and L. Levin. 2002. Automatic rule learning for resource-limited MT. In *Proceedings of AMTA-02*, volume 2499 of *LNCS*, pages 1–10, London, UK.

H. M. Caseli and M. G. V. Nunes. 2007. Automatic induction of bilingual lexicons for machine translation. *International Journal of Translation*, 19:29–43.

H. M. Caseli, M. G. V. Nunes, and M. L. Forcada. 2005. Evaluating the LIHLA lexical aligner on Spanish, Brazilian Portuguese and Basque parallel texts. *Procesamiento del Lenguaje Natural*, 35:237–244.

H. M. Caseli, M. G. V. Nunes, and M. L. Forcada. 2008. Automatic induction of bilingual resources from aligned parallel corpora: application to shallow-transfer machine translation. *Machine Translation*. (in press).

G. Doddington. 2002. Automatic evaluation of machine translation quality using n-gram cooccurrence statistics. In *Proceedings of ARPA WHLT*, pages 128–132, San Diego.

P. Fung. 1995. A pattern matching method for finding noun and proper noun translations from noisy parallel corpora. In *Proceedings of ACL-95*, pages 236–243.

P. Koehn and K. Knight. 2002. Learning a translation lexicon from monolingual corpora. In *Proceedings of the Workshop of the ACL SIGLEX*, pages 9–16, Philadelphia.

P. Langlais, G. Foster, and G. Lapalme. 2001. Integrating bilingual lexicons in a probabilistic translation assistant. In *Proceedings of the 8th MT Summit*, pages 197–202, Santiago de Compostela, Spain.

A. Lavie, K. Probst, E. Peterson, S. Vogel, L. Levin, A. Font-Llitjós, and J. Carbonell. 2004. A trainable transfer-based machine translation approach for languages with limited resources. In *Proceedings of EAMT-04*, pages 1–8, Valletta, Malta.

K. McTait. 2003. Translation patterns, linguistic knowledge and complexity in an approach to EBMT. In M. Carl and A. Way, editors, *Recent Advances in Example-Based Machine Translation*, pages 1–28.

A. Menezes and S. D. Richardson. 2001. A best-first alignment algorithm for automatic extraction of transfer mappings from bilingual corpora. In *Proceedings of the Workshop on Data-driven Machine Translation at ACL-01*, pages 39–46, Toulouse, France.

F. J. Och and H. Ney. 2000. Improved statistical alignment models. In *Proceedings of the ACL-00*, pages 440–447, Hong Kong, China, October.

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

F. J. Och and H. Ney. 2004. The alignment template approach to statistical machine translation. *Computational Linguistics*, 30(4):417–449.

K. Papineni, S. Roukos, T. Ward, and W. Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the ACL-02*, pages 311–318, Philadelphia, PA.

J. Pei, J. Han, B. Mortazavi-Asl, J. Wang, H. Pinto, Q. Chen, U. Dayal, and M. Hsu. 2004. Mining sequential patterns by pattern-growth: the PrefixSpan approach. *IEEE Transactions on Knowledge and Data Engineering*, 16(10):1–17.

F. Sánchez-Martínez and M. L. Forcada. 2007. Automatic induction of shallow-transfer rules for open-source machine translation. In *Proceedings of the TMI 2007*, pages 181–190.

C. Schafer and D. Yarowsky. 2002. Inducing translation lexicons via diverse similarity measures an bridge languages. In *Proceedings of CoNLL-02*, pages 1–7.

D. Wu and X. Xia. 1994. Learning an English-Chinese lexicon from parallel corpus. In *Proceedings of the AMTA-94*, pages 206–213, Columbia, MD.

Exploring Spanish-morphology effects on Chinese–Spanish SMT

Rafael E. Banchs
Barcelona Media Innovation Centre
Ocata #1, Barcelona
08003 Spain
rafael.banchs@barcelonamedia.org

Haizhou Li
Institute for Infocomm Research
Heng Mui Keng Terrace #21
119613 Singapore
hli@i2r.a-star.edu.sg

Abstract

This paper presents some statistical machine translation results among English, Spanish and Chinese, and focuses on exploring Spanish-morphology effects on the Chinese to Spanish translation task. Although not strictly comparable, it is observed that by reducing Spanish morphology the accuracy achieved in the Chinese to Spanish translation task becomes comparable to the one achieved in the Chinese to English task. Further experimentation on approaching the problem of generating Spanish morphology as a translation task by itself is also performed, and results discussed. All experiments have been carried out by using a trilingual parallel corpus extracted from the Bible.

1 Introduction

The Chinese–Spanish translation task has been recently explored by Banchs et al. (2006). As far as we know, no Chinese–Spanish parallel corpus large enough for training a statistical machine translation system is available, at least as a public resource. For this reason, in that previous work, the artificial generation of the required Chinese–Spanish parallel corpus was attempted in order to pursue machine translation experimentation for this specific language pair.

From that work it was concluded that artificial generation of the bilingual corpus did not provide better translation accuracy than cascading two independent translation systems by using English as a bridge. Even more, filtering the artificially generated corpus aiming at improving translation results did not help at all, because the negative effect of reducing corpus size was more influential than the positive effect of improving corpus quality, at least for the corpus size considered in that opportunity.

In the present work, we present some experimentation results with a small parallel corpus we have extracted from the Bible. The collected corpus includes English, as well as Spanish and Chinese. The corpus collection and preparation, as well as its statistics are presented in section 2. Then, some baseline experimentation is carried out among the three languages in order to determine the best alignment set, phrase size and language model order for each of the six possible translation tasks. These results are presented in section 3. Then, the effect of Spanish morphology is explored for the particular case of the Chinese to Spanish translation task. In this sense, Spanish morphology is reduced by using a morphological analyzer and a Chinese to Spanish-without-morphology translation system is constructed. The problem of Spanish morphology generation is also approached as a translation task, and the Chinese to Spanish translation problem is attempted in a two step procedure in order to alleviate the translation task complexity by decoupling the translation task from the morphology generation task. These results are presented and discussed in section 4. Finally, some conclusions are presented, and future research strategies in this area are depicted.

2 Corpus collection and preparation

The trilingual parallel corpus used in this work has been extracted from three versions of the Bible Chinese (ZH), English (EN) and Spanish (ES). The original documents have been obtained from the web in digital format.[1] In the case of the Chinese and English versions, the complete text was available in a single document; while in the case of the Spanish version, each of the 66 books was in a separated file. The collected corpus was preprocessed and prepared for SMT experimentation by using the procedures described below.

Alignment: alignment at the sentence level was performed. In this particular case, this step was actually a simple one since the original text included annotation marks for chapters and verses. However some manual verification and edition was required since some missing verses and annotation inconsistencies were detected among the three different versions.

Tokenization: each data file was tokenized. In the case of Spanish and English, this implies the separation between punctuation marks and words. For the case of Chinese, for which word segmentation is not obvious, automatic word segmentation was performed by using the freely available tool ICTCLAS (Zhang et al, 2003).[2]

Morphology reduction: Morphological analysis and preprocessing of Spanish data was carried out. Such a preprocessing produces a slightly different tokenization for the Spanish data mainly because some multi-word units are reduced to single lexical forms. Because of this, four different Spanish data sets are considered: the original tokenized data, a lowercased version of the original tokenized data (lwc), re-tokenized data resulting from applying morphological analysis to the lowercased data set (rtk), and a lemmatized version of the re-tokenized one (lem). The lemmatized data corresponds to a morphologically reduced corpus in which all full forms have been replaced by their corresponding lemma forms. The morphological analysis was performed by using the freely available tool FreeLing (Carreras et al, 2002).[3]

Length restriction: all sentences (in any of the three languages) containing more than 80 tokens were removed from the corpus along with their corresponding other-two-language sentences. This restriction was mainly adopted in order to avoid possible alignment problems.

Fertility filtering: all trilingual sentence sets, for which any pair of them presented a token ratio equal to or higher than 9, were removed from the corpus. This avoids symmetrization errors due to fertility filtering implemented by the word to word alignment tool used for training the models, which also considers a token ratio of 9.

Corpus segmentation: finally, the corpus was divided into three trilingual parallel data sets: training, development and test.

Table 1 presents the main corpus statistics for all data set considered in the experiments. These statistics include the total number of sentences, the total number of words, the size of the vocabulary and the average sentence length. The out-of-vocabulary rates for development and test data are, respectively, *3.7%* and *4.2%* for Chinese, *9.3%* and *8.9%* for Spanish, and *5.3%* and *4.2%* for English.

Training data set				
Language	Senten.	Tokens	Vocab.	Aver.
EN	28,887	848,776	13,216	29.38
ZH	28,887	760,451	12,670	26.33
ES	28,887	781,113	28,178	27.04
ES-lwc	28,887	781,113	26,251	27.04
ES-rtk	28,887	784,398	25,240	27.15
ES-lem	28,887	784,398	14,229	27.15
Development data set				
Language	Senten.	Tokens	Vocab.	Aver.
EN	1,033	30,199	3,234	29.23
ZH	1,033	27,235	3,404	26.37
ES	1,033	27,862	4,634	26.97
ES-lwc	1,033	27,862	4,413	26.97
ES-rtk	1,033	27,986	4,403	27.09
ES-lem	1,033	27,986	2,882	27.09
Test data set				
Language	Senten.	Tokens	Vocab.	Aver.
EN	1,035	30,008	3,158	28.99
ZH	1,035	26,794	3,396	25.89
ES	1,035	27,368	4,652	26.44
ES-lwc	1,035	27,368	4,428	26.44
ES-rtk	1,035	27,452	4,426	26.52
ES-lem	1,035	27,452	2,864	26.52

Table 1: Main corpus statistics

[1] The Spanish version was downloaded from http://es.catholic.net/biblia/, the Chinese version from http://www.o-bible.org/download/hgb.txt and the English version from http://www.o-bible.com/dlb.html

[2] Available at http://www.nlp.org.cn/project/project.php?proj_id=6

[3] Available at http://garraf.epsevg.upc.es/freeling/

3 Baseline experimentation

For all experiments presented in this work, a very basic phrase-based SMT system is used. Word to word alignments are computed for the training data sets by using GIZA++ (Och & Ney, 2003).[4] Phrases are extracted from alignments and the translation probabilities are estimated by using relative frequencies. Language models are computed by using the SRILM toolkit (Stolcke, 2002),[5] and decoding is carried out by using Pharaoh (Koehn, 2004),[6] for which only four basic feature functions are considered: the translation model, the language model, the distortion model and the word penalty factor. Model weight optimization is performed by using the standard minimum-error-training procedure (Och, 2003) which was implemented by using the Simplex algorithm for maximizing translation BLEU over the development data set.

Some baseline experimentation was carried out among the three languages in order to determine the best alignment set, phrase size and language model order for each of the six possible translation tasks. In these baseline experiments, four different alignment sets were considered for phrase extraction: source to target (sr2tg), intersection (inter), union (union) and symmetrized alignments (sym) (Matusov et al, 2004). Regarding phrase lengths and target language models, two maximum phrase lengths were considered for translation model computation: *3* and *4* tokens; and three maximum *n*–gram sizes were considered for target language model computation: *2*–, *3*– and *4*–grams.

According to results from these baseline experiments, the optimal maximum phrase length for translation model computation was consistently *4* tokens for all translation tasks; and, similarly, the optimal language model order for target language model computation was consistently *3*. However, in the case of the alignment set considered for phrase extraction interesting differences could be observed. Table 2 presents BLEU scores over the test set for all of the six possible translation tasks when extracting phrases from each of the four different alignment sets considered.[7] For all results presented in table 2, the maximum phrase lengths considered were *4* tokens, and the maximum *n*–gram sizes considered were *3*–grams.

Note from table 2 that, although in many cases performances are relatively similar, in the cases were Spanish is the target language the intersection clearly offers the best performance. In all other cases, with the exception of the English to Chinese task for which the source to target seems to be performing better, the symmetrized set of alignments performs slightly better.

Task	Sr2tg	Inter	Union	Sym
ZH-ES	14.1	**14.3**	12.9	13.8
ES-ZH	16.7	17.4	15.2	**17.5**
ZH-EN	18.6	19.2	17.2	**19.6**
EN-ZH	**20.2**	20.1	19.3	19.7
EN-ES	30.6	**31.5**	30.6	30.5
ES-EN	34.4	34.2	34.3	**34.5**

Table 2: Translation BLEU over the test set for all six tasks and the four alignment sets considered.

Note also from table 2, how the lowest translation qualities are obtained for the Chinese–Spanish language pair, and the highest qualities are obtained for the English–Spanish language pair. Moreover, if we take a closer look at the table, these results suggest that having Spanish as the target language seems to add a significant degree of complexity to the translation task, and the most suspicious element for explaining this behavior is, for sure, its high morphological variations.

4 Effects of Spanish morphology

Previous works have shown how morphological information can be used to improve statistical machine translation results, especially when a limited amount of training data is available (Nießen and Ney, 2004; Popovic and Ney, 2004). In this section we explore the effects of reducing the Spanish morphology on the Chinese–Spanish translation tasks. For all experiments presented here, phrases extracted from the intersection set of alignments were used, and the four different Spanish data sets described in section 2 were considered.

Table 3 presents BLEU scores for both Chinese to Spanish and Spanish to Chinese translation tasks when using the four different Spanish data sets.

From table 3 it can be seen that reducing Spanish morphology by using lemmas instead of full

[4] Available at http://www.fjoch.com/GIZA++.html

[5] Available at http://www.speech.sri.com/projects/srilm/

[6] Available at http://www.isi.edu/publications/licensed-sw/pharaoh/

[7] Note that in these experiments only one translation reference is available for computing BLEU scores, in both the optimization procedure and the evaluation procedure.

forms definitively improves the translation system performance; and, as it would be logically expected, the greater impact occurs when Spanish is the target language. In this case an absolute improvement of more than four BLEU points was achieved. In this sense, note from tables 2 and 3 that, although not strictly comparable, translation quality achieved for the Chinese to lemmatized-Spanish task seems to be similar to the quality achieved for the Chinese to English translation task. On the other hand, for the Spanish to Chinese translation task, the improvement obtained by reducing the Spanish morphology was only a little bit more than a half BLEU point.

Spanish set	ES to ZH	ZH to ES
Baseline	17.4	14.3
Lowercased	17.3	16.1
Re-tokenized	17.6	15.5
Lemmatized	17.9	18.9

Table 3: Translation BLEU over the test set for Chinese–Spanish tasks and the four Spanish sets.

It can also be observed from table 3 that the effects of lowercasing and the re-tokenization generated by the analyzer seem to have opposite effects in both translation tasks. While lowercasing helps the Chinese to Spanish task, this is not the case for opposite direction; and re-tokenization seems to be producing an opposite effect.

Additionally, the problem of Spanish morphology generation was also approached as a translation task. In this sense, a translation system was implemented by using the lemmatized Spanish data set as source language and the original Spanish data set as the target language. By training and optimizing such a system, a BLEU score of *67.4* was measured over the corresponding test set, which happens to be a very high BLEU score due to the fact that a single translation reference was used. Then, by cascading the two systems: Chinese to lemmatized-Spanish and the lemmatized-Spanish to full-Spanish, a BLEU score of *14.3* was measured over the test set. Note that this result is basically the same as the one reported in table 2 for the direct Chinese to full-Spanish translation system. At a first glance, this seems to suggest that translating from Chinese into a lemmatized version of Spanish and the subsequent generation of the Spanish final forms are independent components of the Chinese to Spanish translation task, because training and optimizing a direct system provides exactly the same translation accuracy that training and optimizing both components separately.[8]

In order to explore in more detail the possible independence of the lemma translation and the final form generation processes, we decided to perform a simultaneous optimization of both systems in the cascade. In this sense we optimized the model weights of both components in the cascade (Chinese to lemmatized-Spanish and lemmatized-Spanish to full-Spanish) with respect to the BLEU score of the overall output of the cascade. In the case both components were indeed independent, we would expect exactly the same translation accuracy that was obtained when optimizing each component independently from the other. But this was not the case because a small, but statistically significant, improvement of more than a half BLEU point was achieved when performing the simultaneous optimization (a score of 14.9 was measured over the test set). This reveals that some interactions exist among the models in both components of the cascade system. Further study is necessary in order to better understand such interactions.

5 Conclusions and future research

This paper presented some statistical machine translation results among English, Spanish and Chinese, focusing on the exploration of Spanish-morphology effects on Chinese to Spanish translation tasks. In this sense, the reduction of Spanish morphology produced an absolute improvement of more than four BLEU points in the Chinese to Spanish direction; and only produced an improvement of a half BLEU point for the opposite translation direction. Although not strictly comparable, it was also observed that the accuracy achieved in the Chinese to Spanish translation task becomes comparable to the one achieved in the Chinese to English task when Spanish morphology is dropped.

Further experimentation on approaching the problem of generating Spanish morphology as a

[8] Another interesting observation is the fact that the cascade system is actually behaving in an analog manner to a series connection of two conductances: the cascade connection will perform poorer than the poorer of the two components. As an interesting fact, the reader can verify that the series combination of BLEUs holds approximately: *67.4 x18.9 / (67.4+18.9) = 14.7 ≈ 14.3*.

translation task by itself was also performed, and a small improvement over the direct Chinese to Spanish task was achieved by jointly optimizing a cascade system of two SMT components: the first one dealing with the problem of Chinese to lemmatized-Spanish translation, and the second one dealing with Spanish-morphology generation.

According to this, further research in Chinese–Spanish SMT must consider as important issues the design and evaluation of strategies for handling Spanish morphology in the particular case of Chinese to Spanish translation tasks. In this sense, better understanding of model interactions and their implications in the translation task should be performed. We will continue exploring new strategies in the direction presented in section 4. Additionally, alternative means for Spanish morphology generation which are independent from the translation task should be considered and studied.

Nevertheless, the actual drawback of the Chinese–Spanish translation task is the lack of a parallel corpus large enough for training a state-of-the-art SMT system. Most of the problems identified in this work, which are related to the richness of Spanish morphology, can be counteracted by means of a larger data set. In this sense, the development of bilingual Chinese–Spanish resources is also another important issue to deal with. In order to pursue research in this direction, the development of Chinese–Spanish translation models by combining translation models that involve intermediate languages should be explored (Wu & Wang, 2007; Cohn & Lapata, 2007). Additionally, methods for extracting parallel corpus from comparable corpora could also be an option for the automatic generation of parallel data sets for SMT purposes (Munteanu & Marcu, 2005).

In the next future, we intend to explore in more detail some of these options in the specific context of Chinese–Spanish statistical machine translation tasks.

Acknowledgements

This work has been made possible by a grant from AGAUR (*Agència de Gestió d'Ajuts Universitaris i de Recerca*) of the Catalonian Department of Innovation, Universities and Enterprises; and by the collaboration of the Human Language Technology Department of the Institute for Infocomm Research in Singapore.

References

R.E. Banchs, J.M Crego, P. Lambert, J.B. Mariño, 2006, "A feasibility study for Chinese-Spanish statistical machine translation", in *Proc. of the 5th Int. Sym. on Chinese Spoken Language Processing*.

X. Carreras, I. Chao, L. Padró, M. Padró, 2002, "FreeLing: an open-source suite of language analyzers", in *Proc. of the 3rd Int. Conf. on Language Resources and Evaluation*.

T. Cohn, M. Lapata, 2007, "Machine translation by triangulation: making effective use of multi-parallel corpora", in *Proc. of the 45th Ann. Meeting of the Association of Computational Linguistics*, pp. 728-735.

P. Koehn, 2004, "Pharaoh: a beam search decoder for phrase-based statistical machine translation alignment models", in *Proc. of AMTA*.

E. Matusov, R. Zens, H. Ney, 2004, "Symmetric word alignments for statistical machine translation", in *Proc. of the 20th Int. Conf. on Computational Linguistics*.

D.S. Munteanu, D. Marcu, 2005, "Improving machine translation performance by exploiting non-parallel corpora", *Computational Linguistics*, vol 31, no 4, pp. 477-504.

S. Nießen, H. Ney, 2004, "Statistical machine translation with scarce resources using morpho-syntactic information", *Computational Linguistics*, vol 30, no 2, pp. 181-204.

F.J. Och, 2003, "Minimum error rate training in statistical machine translation", in *Proc. of ACL*.

F.J. Och, H. Ney, 2003, "A systematic comparison of various statistical alignment models", *Computational Linguistics*, vol 29, no 1, pp. 19-51.

M. Popovic, H. Ney, 2004, "Towards the use of word stems and suffixes for statistical machine translation", in *Proc. of the 4th Int. Conf. on Language Resources and Evaluation*, pp. 1585-1588.

A. Stolcke, 2002, "SRILM - an extensible language modeling toolkit", in *Proceedings of the International Conference on Spoken Language Processing*.

H. Wu, H. Wang, 2007, "Pivot language approach for phrase-based statistical machine translation", in *Proc. of the 45th Ann. Meeting of the Association of Computational Linguistics*, pp. 856-863.

H. Zhang, H. Yu, D. Xiong, Q. Liu, 2003, "HHMM-based Chinese lexical analyzer ICTCLAS", in *Proc. of the 2nd SIGHAN Workshop on Chinese Language Processing*, pp. 184-187.

Linguistic Categorisation in Machine Translation using Stochastic Finite State Transducers[1]

Jorge González and Francisco Casacuberta
Departamento de Sistemas Informáticos y Computación
Instituto Tecnológico de Informática
Universidad Politécnica de Valencia
{jgonzalez, fcn}@dsic.upv.es

Abstract

In the last years, statistical machine translation has already demonstrated its usefulness within a wide variety of translation applications. In particular, finite state models are always an interesting framework because there are well-known efficient algorithms for their representation and manipulation. Nevertheless, statistical approaches have rarely been performed taking into account the linguistic nature of the translation problem. This document describes some methodological aspects of building category-based finite state transducers that are able to consider a set of linguistic features in order to produce the most linguistically appropriate hypotheses.

1 Introduction

Machine Translation (MT) is a consolidated area of research in computational linguistics which investigates the use of computer software to translate text or speech from one natural language to another. The goal of MT is very ambitious because it would involve a reduction of the linguistic barriers in human communication.

Despite their initial relative success, rule-based systems were quickly challenged by their rival inductive approaches, which adopt some pattern recognition techniques to learn the models. *Statistical* machine translation represents an interesting framework because the translation software is language-independent, that is, different MT systems are built if different parallel corpora are supplied.

Given a source sentence $\mathbf{s}_1^J = \mathbf{s}_1 \ldots \mathbf{s}_J$, the goal of statistical machine translation is to find a target sentence $\hat{\mathbf{t}}_1^I = \mathbf{t}_1 \ldots \mathbf{t}_I$, among all the possible target strings $\mathbf{t}_1^I$, that maximises the posterior probability of $\mathbf{t}_1^I$ given $\mathbf{s}_1^J$:

$$\hat{\mathbf{t}}_1^I = \underset{\mathrm{t}_1^I}{\operatorname{argmax}} \Pr(\mathbf{t}_1^I | \mathbf{s}_1^J) \qquad (1)$$

Since $\Pr(\mathbf{s}_1^J)$ is independent of $\mathbf{t}_1^I$, the equation (1) can be rewritten to (2), using a joint probability distribution that is modelled by means of stochastic finite state transducers:

$$\hat{\mathbf{t}}_1^I = \underset{\mathrm{t}_1^I}{\operatorname{argmax}} \Pr(\mathbf{s}_1^J, \mathbf{t}_1^I) \qquad (2)$$

Despite the linguistic nature of languages has been traditionally ignored in statistical machine translation, there is some recent related work that tries to incorporate some linguistic knowledge into a statistical framework (Niessen, 2004; Gispert, 2006; Koehn, 2006).

The organization of this paper is as follows: next section presents the statistical framework; section 3 describes the methodological aspects of building a category-based system, where training and decoding steps are

[1] This work was supported by the EC (FEDER) and the Spanish projects TIN2006-15694-C02-01 and the Consolider Ingenio 2010 CSD2007-00018

explained in depth; the experimental setup and results are shown in section 4; finally, conclusions are briefly summed up at section 5.

2 Statistical framework

Machine translation can be seen as a process of pattern recognition, where objects to be tested are sentences from a source language. These sentences should be coded in a process of feature extraction in order to be classified or described by a previously estimated model.

On the one hand, geometric feature extraction defines a real object $\mathbf{s}$ as a feature vector where every observed feature is measured on $\mathbf{s}$ and then annotated to the right position. On the other hand, syntactic feature extraction establishes a structural description of $\mathbf{s}$, according to some structure-based instructions.

Given that a text sentence $\mathbf{s}$ represents a structural description, i.e. a string of symbols, these word sequences have been traditionally employed in the field of computational linguistics as a result of a feature extraction process.

However, nobody ignores that the linguistic nature of languages could be statistically exploited in order to obtain some better models. In such a line, every word in a sentence is expanded into a tuple of three different pieces of information: on the one hand, the written word itself, also known as surface form; on the other hand, its base form, also referred in the literature as lemma; finally, a linguistic feature vector reports information about its lexical category together with a set of linguistic properties, such as gender, number, etc. In this way, a traditional definition of $\mathbf{s} = \mathbf{s}_1 \ldots \mathbf{s}_J$ would be replaced by an extended string $\mathbf{s} = (\mathbf{s}_1, m_1, u_1) \ldots (\mathbf{s}_J, m_J, u_J)$, where m_j stands for the lemma of word $\mathbf{s}_j$, and u_j stands for its linguistic feature vector.

Given that a lemma can be seen as a linguistic cluster, where words sharing the same lemma are classified into the same cluster, the vocabulary can be significantly reduced by changing the words to their lemmas during the estimation of the joint probability model.

Let $\mathbf{s} = (\mathbf{s}_1^J, m_1^J, u_1^J)$ and $\mathbf{t} = (\mathbf{t}_1^I, n_1^I, v_1^I)$ be a source and a target sentence respectively, equation 2 can be tackled through a categorisation scheme:

$$\begin{aligned} \Pr(\mathbf{s}_1^J, \mathbf{t}_1^I) \;=\; & \Pr(m_1^J, n_1^I) \cdot \Pr(u_1^J | m_1^J, n_1^I) \cdot \\ & \Pr(v_1^I | m_1^J, n_1^I, u_1^J) \cdot \\ & \Pr(\mathbf{s}_1^J | m_1^J, n_1^I, u_1^J, v_1^I) \cdot \\ & \Pr(\mathbf{t}_1^I | m_1^J, n_1^I, u_1^J, v_1^I, \mathbf{s}_1^J) \end{aligned}$$

which, under certain assumptions, turns to:

$$\begin{aligned} \Pr(\mathbf{s}_1^J, \mathbf{t}_1^I) \;\approx\; & \Pr(m_1^J, n_1^I) \cdot \Pr(u_1^J | m_1^J) \cdot \\ & \Pr(v_1^I | n_1^I) \cdot \Pr(\mathbf{s}_1^J | m_1^J, u_1^J) \cdot \\ & \Pr(\mathbf{t}_1^I | n_1^I, u_1^J, v_1^I) \end{aligned}$$

Lemma-based joint probability distributions $\Pr(m_1^J, n_1^I)$ can be modelled by stochastic finite state transducers, whereas specialised stochastic dictionaries can be estimated to model uncategorising lemma-to-word transformations $n_1^I \rightarrow \mathbf{t}_1^I$, according to a given source feature vector u_1^J, assuming that $\Pr(\mathbf{t}_1^I | n_1^I, u_1^J, v_1^I)$ is also independent of v_1^I. This behaviour is based on a Spanish↔Catalan machine translation system (González, 2006) which assumes that linguistic information is transferred from input to output, remaining unaltered in most cases.

The equation (2) will then be expressed as:

$$\begin{aligned} \hat{n}_1^I &= \operatorname*{argmax}_{n_1^I} \Pr(m_1^J, n_1^I) \\ \hat{\mathbf{t}}_1^I &= \operatorname*{argmax}_{\mathbf{t}_1^I} \Pr(\mathbf{t}_1^I | \hat{n}_1^I, u_1^J) \end{aligned} \tag{3}$$

The search must be constrained in order to perform first a lemma transduction operation, that is, translating from source to target lemmas, then turning lemmas into words, through their corresponding feature vectors.

Specialised stochastic dictionaries can be estimated following the maximum likelihood approach in order to compute $\Pr(\mathbf{t}_1^I | \hat{n}_1^I, u_1^J)$. The specialisation criteria can be seen from two equivalent points of view: on the one hand, a stochastic dictionary can be trained for every different target lemma, thus every entry informs about how a feature vector can be translated into a target word; or, maybe more intuitively, training a lemma-to-word stochastic dictionary per each feature vector. The calculation of $\Pr(\mathbf{t}_1^I|\hat{n}_1^I, u_1^J)$ is carried out by means of the contribution of all the individual translation probabilities, that is:

$$\Pr(\mathbf{t}_1^I|\hat{n}_1^I, u_1^J) \approx \prod_{i=1}^{I} \Pr(\mathbf{t}_i|\hat{n}_i, u_{\alpha_i})$$

Formally, an alignment function α is a mapping $\alpha : i \rightarrow j$ that assigns a source position j to a target position i, $\alpha_i = j$. Alignments are used as hidden variables in statistical machine translation models such as IBM models (Brown, 1990) or hidden Markov models (Zens, 2002). Therefore, target lemmas being generated are able to know which source position was responsible for their occurrence.

3 Probabilistic models

A weighted finite-state automaton is a tuple $\mathcal{A} = (\Gamma, Q, i, f, P)$, where Γ is an alphabet of symbols, Q is a finite set of states, functions $i : Q \rightarrow \mathbb{R}$ and $f : Q \rightarrow \mathbb{R}$ give a weight to the possibility of each state to be initial or final, respectively, and partial function $P : Q \times \{\Gamma \cup \{\lambda\}\} \times Q \rightarrow \mathbb{R}$ defines a set of transitions between pairs of states in such a way that each transition is labelled with a symbol from Γ or the empty string λ, and is assigned a weight.

A weighted finite-state transducer (Mohri, 2002; Kumar, 2006) is defined similarly to a weighted finite-state automaton, with the difference that transitions between states are labelled with pairs of symbols that belong to the cartesian product of two different (input and output) alphabets, $\{\Sigma \cup \{\lambda\}\} \times \{\Delta \cup \{\lambda\}\}$.

When weights are probabilities, and under certain conditions, a weighted finite-state model can define a distribution of probabilities on the free monoid. In that case it is called a stochastic finite-state model. Then, given some input/output strings $\mathbf{s}_1^J$ and $\mathbf{t}_1^I$, a stochastic finite-state transducer is able to associate a probability $\Pr(\mathbf{s}_1^J, \mathbf{t}_1^I)$ to them.

3.1 Inference of stochastic transducers

The GIATI paradigm (Casacuberta, 2005) has been revealed as an interesting approach to infer stochastic finite-state transducers through the modelling of languages. Rather than learning translations, GIATI first converts every pair of parallel sentences from the training corpus into only one string to, after all is done, infer a language model from.

More concretely, given a parallel corpus consisting of a finite sample C of string pairs: first, each training pair $(\bar{x}, \bar{y}) \in \Sigma^\star \times \Delta^\star$ is transformed into a string $\bar{z} \in \Gamma^\star$ from an extended alphabet, yielding a string corpus S; then, a stochastic finite-state automaton $\mathcal{A}$ is inferred from S; finally, transition labels in $\mathcal{A}$ are turned back into pairs of strings of source/target symbols in $\Sigma^\star \times \Delta^\star$, thus converting the automaton $\mathcal{A}$ into a transducer $\mathcal{T}$.

The first transformation is modelled by some labelling function $\mathcal{L} : \Sigma^\star \times \Delta^\star \rightarrow \Gamma^\star$, whereas the last transformation is defined by an inverse labelling function $\Lambda(\cdot)$, such that $\Lambda(\mathcal{L}(C)) = C$. Building a corpus of extended symbols from the original bilingual corpus allows for the use of many useful algorithms for learning stochastic finite-state automata (or equivalent models) that have been proposed in the literature about grammatical inference.

Every extended symbol from Γ has to condense somehow the meaningful relationship that exists between the words in the input and output sentences. Discovering these relations is a problem that has been throughly studied in statistical machine translation and has well-established techniques for dealing with it. The concept of statistical alignment formalises this problem. Whether this function is constrained to a one-to-one, a one-to-many or a many-to-many correspondence depends on the particular assumptions that we make. Constraining the alignment function simplifies the learning procedure but reduces the expressiveness of the model. The available algorithms try to find a trade-off between complexity and expressiveness.

One-to-one and one-to-many alignment functions would enable models to adopt the categorisation scheme presented here because they allow for alignments where one target position is aligned to only one source position.

One-to-one models do not seem a very appropriate approach provided that they would require that source-target aligned sentences had exactly the same number of words. Nevertheless, one-to-many alignment models are a current reference in machine translation research community by means of their well-known IBM models (Brown, 1990).

A smoothed n-gram model may be inferred from the string corpus previously generated. Such a model can be expressed in terms of a weighted finite-state automaton. Since every transition consumes only one symbol, and given that all those extended symbols are composed of exactly one source element, the inverse labelling function can be straightforwardly applied. This way, transition labels are turned back into pairs of source and target items, thus becoming a stochastic transducer.

3.1.1 Alignment models

The conversion of every pair of parallel sequences into an extended symbols string follows this algorithm: for each target item from left to right, merge it with its corresponding source element iff the alignment does not cross over any other alignment, in which case it is delayed and attached to the last implied source item. Spurious source and target elements are placed at their right position, given that a monotonous order is always demanded. This procedure ensures that every extended symbol is composed of one and only one source symbol, optionally followed by an arbitrary number of target symbols. For a more detailed description about the labelling function, see (Casacuberta, 2005).

The implementation of the categorisation scheme will require increasing the information to be included in every compound symbol. More concretely, all the target lemmas being produced by the model need to report which relative source position they are coming from.

Figure 1 displays the two situations which the labelling function may be involved with.

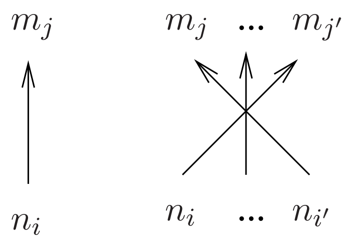

Figure 1: Two types of alignments

Whereas the first example (namely, the relation $n_i \rightarrow m_j$) is undoubtedly easy to solve, the second one implies a little more of work. One-to-one relationships clearly establish that n_i is aligned to the current source symbol being analysed m_j. This is denoted as a relative movement of 0, as it can be seen in figure 2.

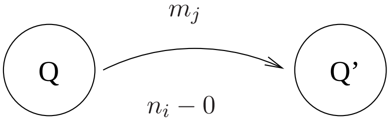

Figure 2: One-to-one compound symbols

On the other hand, crossing alignments would imply delaying the output of $\{n_i \ldots\}$ until $m_{j'}$ is being parsed, then producing the full target segment $n_i \ldots n_{i'}$. Therefore, every lemma being generated may not be aligned with its corresponding input symbol as before, but with some previously parsed one instead.

As a consequence, target lemmas are annotated together with their relative distance to the source lemma which they were aligned to. Spurious elements do not need such annotation because of their own spontaneous generation, which is independent of any particular source item. In figure 1, n_i is aligned to the current source element $m_{j'}$, thus indicated as a 0 relative movement. However, the emission of $n_{i'}$ will be delayed, then moving it further away from its aligned input item m_j. This relative distance is then annotated next to the output symbol $n_{i'}$ as a reminder to allow for a posterior backtracking performance. The result of such a labelling algorithm can be seen over the final transducer, as figure 3 shows.

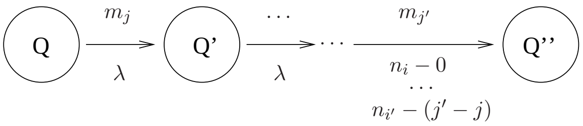

Figure 3: One-to-many compound symbols

Note that the relative distance for $n_{i'}$ is computed as the subtraction of the input position aligned to, j, from the current one, j'.

3.2 The search problem

Word-to-word translation as in equation (2), or lemma-to-lemma translation as in the first equation of (3), are expressions of the MT problem in terms of a finite state model that is able to compute a joint probability. Given that only the input sentence is known, the model has to be parsed, taking into account all the outputs that are compatible with the input. The best target hypothesis would be that one which corresponds to a path through the transduction model that, with the highest probability, accepts the input sequence as part of the input language of the transducer.

Although the navigation through the model is constrained by the input sequence, the search space can be extremely large. As a consequence, only those partial hypotheses with the highest scores are being considered as possible candidates to become the solution. This search process is very efficiently carried out by the well known Viterbi algorithm.

3.3 Stochastic dictionaries

A weighted dictionary is a table $(a, b, W(a, b))$ containing a set of translation pairs together with a numerical indicator for their reliability. If $W(a, b) = \Pr(a|b)$ and $\forall y \sum_x \Pr(x|y) = 1$, then it can be called a stochastic dictionary.

Once a lemmatised source sentence has been analysed by the transduction model, output is expressed as a sequence of target lemmas. They can be turned into their corresponding surface forms by means of specialised stochastic dictionaries that take into account the linguistic information of the source elements which they are attached to.

Following the maximum likelihood approach, a stochastic dictionary can be estimated by counting the absolute frequencies of the observed events, properly normalised:

$$\Pr(t_i|n_i) = \frac{F(t_i, n_i)}{\sum_x F(x, n_i)}$$

These dictionaries can be learnt by means of two different estimation methods: one considers only a monolingual target corpus, thus learning conversions through their own target linguistic information; and another one that takes into account the statistical alignments over a bilingual corpus in order to train lemma-word transformations according to their corresponding source feature vectors. In this case, the alignments that are needed for learning the stochastic lemma-based transducers are also adequate for the extraction of the lemma-to-word relative frequencies. An outline of this method is depicted in figure 4.

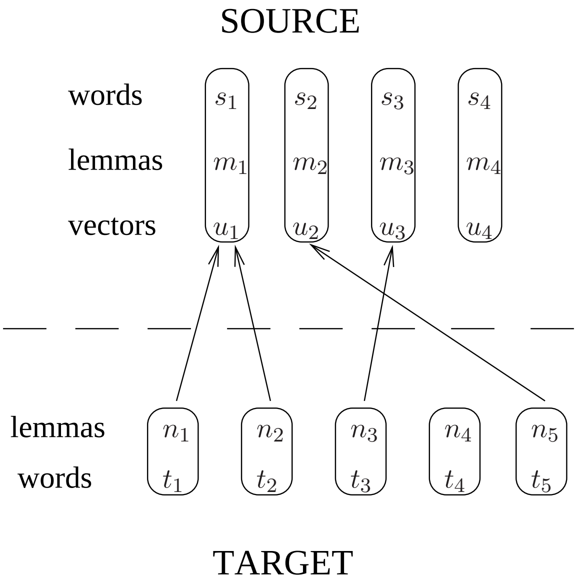

Figure 4: Using the source vectors for a bilingual estimation of lemma-to-word dictionaries

3.4 On-the-fly integrated architecture

The equations in (3) represent the search strategy in order to translate a test sentence from a source language to a target language. According to these equations, translation is carried out in two separate steps: first, source lemmas are transformed into target lemmas through a finite state approach, then lemmas are turned into words by means of specialised, linguistic-based stochastic dictionaries.

However, this two step procedure can be integrated into only one process, thus merging the lemma-word conversions into the parsing algorithm of the lemmatised input sentences.

Let j be a current analysis position of the input sequence m_1^J, and let n_i be a target lemma being produced during the parsing of m_j. Given that a lemma-word translation probability $\Pr(\mathbf{t}_i|n_i)$ has been assumed to also (and only) depend on the source feature vector u_{α_i} which n_i has been aligned to, and since α_i is always guaranteed to be a position $0 \leq \alpha_i \leq j$ that has already been analysed, then $\Pr(\mathbf{t}_i|n_i, u_{\alpha_i})$ can be applied in order to turn a target lemma n_i into a target word $\mathbf{t}_i$.

Thanks to including the alignment information in between the output symbols, it is possible to know for each lemma being generated which input position it has been connected to.

As a result, every target lemma being produced as part of a partial output hypothesis may be converted and stored as a target word, without the need for waiting for the best output hypothesis $\hat{n}_1^I$ to be completely generated.

Once the input sequence m_1^J has been fully parsed through the finite state model, a final surface form $\hat{\mathbf{t}}_1^I$ has been produced on the fly.

4 Experiments

A set of preliminary experiments were carried out in order to test the viability of our integrated category-based translation approach.

Two tasks of very different difficulty degrees were employed for the design of the experimental setup. The EuTrans task is defined on the restricted domain of sentences that a tourist traveller would say at a hotel's desk. It is artificially generated from a set of schemas of sentences. The characteristics of the EuTrans corpus can be seen in table 1. Spanish to English translation was carried out over this low-perplexity task.

On the other hand, this approach has been also applied to a Portuguese–Spanish section of the EuroParl corpus. The EuroParl corpus is built on the proceedings of the European Parliament, which are published on its web and are freely available. Because of its nature, this corpus has a large variability and complexity, since the translations into the different official languages are performed by groups of human translators. The fact that not all translators agree in their translating criteria implies that a given source sentence can be translated in various different ways throughout the corpus. Since the proceedings are not available in every language as a whole, a different subset of the corpus is extracted for every different language pair, thus evolving into somewhat different corpora for each pair. The corpus characteristics can be seen in table 2.

Table 1: EuTrans corpus characteristics

EuTrans		Spanish	English
Training	Sentences	10.000	
	Run. words	97.1K	99.3K
	Vocabulary	686	513
Closed test	Sentences	2.996	
	Perplexity	4.9	3.6
Open test	Sentences	3.000	
	Perplexity	4.9	3.6

Table 2: Characteristics of pt–es EuroParl

EuroParl		Portuguese	Spanish
Training	Sentences	915.570	
	Run. words	23.76M	23.95M
	Vocabulary	141.6K	140.4K
Sub-train	Sentences	50.000	
	Run. words	1.3M	1.3M
	Vocabulary	37.3K	37.6K
Test	Sentences	1.000	
	Train pp.	71.9	66.2
	Sub-train pp.	121.3	103.5

EuTrans lemmatisation and linguistic labelling were carried out through the FreeLing toolkit (Carreras, 2004), whereas SisHiTra (González, 2006) was employed to analyse the Spanish sentences from EuroParl. Portuguese lemmas and feature vectors were provided by the Spoken Language Systems Laboratory from the Instituto de Engenharia de Sistemas e Computadores I+D in Lisbon. Both EuTrans and EuroParl corpora were aligned at word level by means of the toolkit GIZA++.

Several tokenisation options were tested to establish a starting point where the categorisation scheme proposed here could be applied.

4.1 Evaluation metrics

The results were obtained by using the following evaluation measures:

BLEU *(Bilingual Evaluation Understudy) score*: This indicator computes the precision of unigrams, bigrams, trigrams, and tetragrams with respect to a set of reference translations, with a penalty for too short sentences. BLEU measures accuracy, not error rate.

WER *(Word Error Rate)*: The WER criterion calculates the minimum number of editions (substitutions, insertions or deletions) needed to convert the system hypothesis into the sentence considered ground truth. Because of its nature, this measure is a pessimistic one.

4.2 Translation results

EuTrans is a very artificial translation task which is frequently used for debugging purposes. New approaches to statistical machine translation are first tested on such a toy task in order to establish some behaviour criteria. The EuTrans results are reported in table 3.

Table 3: EuTrans results

	Vocab.		Pp.		Metrics	
EuTrans	In	Out	In	Out	WER	BLEU
Baseline	686	513	4.9	3.6	8.3	88.0
Tokenisation	624	513	5.2	3.6	**8.1**	88.0
Categorisation	476	503	4.6	3.6	11.8	82.0
Monolingual					22.0	64.3
Bilingual					13.1	78.9

As it can be seen, our linguistic categorisation approach is not worth the trouble for EuTrans. Tokenisation techniques do perform a slight improvement on word error rate, but lemmatisation make results get worse. Whereas the results from "Categorisation" lines represent a comparison with a predefined lemmatised reference, thus evaluating somehow the effect of the lemma transduction model, "Monolingual" and "Bilingual" lines refer to the overall process of translation, according to the way specialised stochastic lemma-to-word dictionaries were learnt. Therefore, the "Categorisation" error rates are always a lower limit of the overall system. It can also be appreciated that there is a significative difference between using a monolingual or a bilingual lemma-to-word approach.

On the other hand, EuroParl is a more complex task which is reflected through its vocabulary and perplexity figures (see table 2). Due to technical issues, experiments were carried out by using only a subset of the training corpus, which is composed of 50.000 sentences. Lemmatisation can reduce vocabularies about 50%, thus causing perplexities to significatively fall as well, as table 4 shows.

Table 4: EuroParl vocabulary and perplexity

	Vocab.		Pp.	
EuroParl	In	Out	In	Out
Baseline	37.3K	37.6K	121.3	103.5
Tokenisation	37.3K	37.5K	121.3	120.9
Categorisation	18.3K	19.3K	91.1	91.1

The EuroParl results are reported in table 5.

Table 5: EuroParl results

	Metrics		Model size	
EuroParl	WER	BLEU	States	Arcs
Baseline	67.8	19.8	205K	1.06M
Tokenisation	**65.7**	**20.0**	200K	1.04M
Categorisation	**61.3**	**23.0**	166K	925K
Monolingual	81.0	3.0	38K	
Bilingual	**63.2**	**21.4**	94K	

In this case, using morphologically annotated corpora helps to the translation process. As well as tokenisation, categorisation also allows for a better modelling of transference relations between source and target languages. The sizes of the models are also significatively reduced, which means not only a memory saving, but also accelerating the decoding time.

Globally, if a bilingual approach is followed to estimate the lemma-word dictionaries, thus using the *source* linguistic feature vectors to specialise them, then the methodology presented here outperforms the baseline system.

Again, monolingual estimation of dictionaries does not perform well and table 6 can show the reasons for such a so different behaviour.

Table 6: Analysis of lemma-word conversions. An impact is defined as a successful search over the lemma-word dictionaries. If the search fails, then lemmas are left unchanged.

Training		EuTrans	EuroParl
	Spurious	3.6%	8.1%
Monolingual	Impacts	11.3%	0%
	Fails	85.1%	91.9%
Bilingual	Impacts	93.1%	88.5%
	Fails	3.3%	3.4%

From table 6, it seems quite clear why monolingual training is doing worse. Impacts and fails are oppositely distributed with respect to the ones from a bilingual training. Whereas a bilingual training reflects an approximate 90% of impacts, a monolingual training associates this percentage to fails. If most lemmas remain unchangeable, then the evaluation results from tables 3 and 5 can be explained, since the lemma-based hypotheses are being compared to word-based references.

Massive fails for a monolingual training are caused by a mismatch between source and target feature vectors. This could be perfectly understood on the EuroParl task, as two language-dependent linguistic tools were employed for labelling. However, the FreeLing toolkit was used on EuTrans task for both languages, thus resulting quite disappointing that labels are not consistent inter languages.

5 Conclusions

This paper has presented a category-based approach to statistical machine translation, which is based on linguistic information. An integrated architecture, combining finite state transducers and stochastic dictionaries has been proposed. Some preliminary results are rather limited but also encouraging enough.

Acknowledgements

We would like to thank the people of the Spoken Language Systems Laboratory from the Instituto de Engenharia de Sistemas e Computadores I+D in Lisbon, for providing the linguistic analysis of the Portuguese Europarl.

References

A. de Gispert and J. B. Mariño 2006. *Linguistic knowledge in statistical phrase-based word alignment.* Natural Language Engineering (Vol 12, Issue 01, Pgs 91-108).

F. Casacuberta, E. Vidal and D. Picó. 2005. *Inference of finite-state transducers from regular languages.* Pattern Recognition (Vol 38, Num 9, Pgs 1431–1443).

J. González, A. L. Lagarda, J. R. Navarro, L. Eliodoro, A. Giménez, F. Casacuberta, J. M de Val, and F. Fabregat. 2006. *SisHiTra: a Spanish-to-Catalan hybrid machine translation system.* 5th SALTMIL Workshop on Minority Languages, Pgs 69-73, Genoa.

Koehn, P., Federico, M., Shen, W., Bertoldi, N., Hoang, H., Callison-Burch, C., Cowan, B., Zens, R., Dyer, C., Bojar, O., Moran, C., Constantin, A., and Herbst, E. 2006. *Open source toolkit for statistical machine translation: Factored translation models and confusion network decoding.* Technical report, John Hopkins University Summer Workshop.

M. Mohri, F. Pereira and M. Riley. 2002. *Weighted Finite-State Transducers in Speech Recognition.* Computer Speech and Language (Vol. 16, Num. 1, Pgs 69–88).

P. F. Brown, J. Cocke, S. A. Della Pietra, V. J. Della Pietra, F. Jelinek, J. D. Lafferty, R. L. Mercer, and P. S. Roosin. 1990. *A statistical approach to machine translation.* Computational Linguistics (Vol 16, Num. 2, Pgs. 79–85).

R. Zens, F. J. Och and H. Ney. 2002. *Phrase-based statistical machine translation.* `http://citeseer.nj.nec.com/zens02phrasebased.html`

S. Kumar, Y. Deng and W. Byrne. 2006. *A weighted finite state transducer translation template model for statistical machine translation.* Natural Language Engineering (Vol 12, Num 1, Pgs 35–75).

S. Niessen and H. Ney 2004. *Statistical Machine Translation with Scarce Resources Using Morpho-syntactic Information.* Computational Linguistics (Vol 30, Num 2, Pgs. 181-204).

X. Carreras, I. Chao, L. Padró and M. Padró 2004. *FreeLing: An Open-Source Suite of Language Analyzers.* Proceedings of the 4th International Conference on Language Resources and Evaluation, Lisbon.

Vocabulary Extension via PoS Information for SMT

Germán Sanchis, Joan Andreu Sánchez
Instituto Tecnológico de Informática
Universidad Politécnica de Valencia
Camino de Vera, s/n. 46022 Valencia, Spain
{gsanchis,jandreu}@dsic.upv.es

Abstract

One of the weaknesses of the so-called phrase based translation models is that they carry out a blind extraction of the phrase translation table, i.e., they do not take into account the linguistic information which is inherent to every language. On the other hand, Part of Speech (PoS) tagging is a problem that, nowadays, presents a pretty mature state of the art, obtaining error rates of almost 2%. Because of this, the use of automatically PoS-tagged corpora in Statistical Machine Translation (SMT) with the purpose of incorporating syntactical knowledge and enhancing the results obtained by state of the art SMT systems seems quite natural. In this work, we present results obtained on the EuroParl corpus by creating an extended vocabulary composed of the regular words and their PoS tags concatenated to them.

1 Introduction

Machine Translation (MT) is a research field of great importance in the European Community, where language plurality implies both a very important cultural richness and not negligible obstacle towards building a unified Europe. Because of this, a growing interest on MT has been shown both by politicians and research groups, which become more and more specialised in this field. Although the language plurality problem can be seen as a more global problem, reaching in fact world wide, in this paper we will be focusing on European languages, due to the vast amount of free data which is available for them.

Moreover, Statistical Machine Translation (SMT) systems are receiving an increasing importance in the last years. In the tasks they have been trained on, SMT systems are able to deliver similar translation quality than rule-based machine translation systems, with the benefit of requiring little human effort when adapting to new language pairs, whenever suitable corpora are available.

(Brown et al., 1993) established what is considered nowadays as the mathematical background of modern SMT, defining the machine translation problem as follows: given a sentence s from a certain source language, an adequate sentence $\hat{t}$ that maximises the posterior probability is to be found. This leads to the following formula:

$$\hat{t} = \underset{t}{\operatorname{argmax}}\, p(t|s)$$

Applying the Bayes theorem on this definition, one can easily reach the next formula

$$\hat{t} = \underset{t}{\operatorname{argmax}} \frac{p(t) \cdot p(s|t)}{p(s)}$$

and, since we are maximising over t, the denominator can be neglected, arriving to

$$\hat{t} = \underset{t}{\operatorname{argmax}}\, p(t) \cdot p(s|t)$$

where $p(t|s)$ has been decomposed into two different probabilities: the *statistical language model* of the target language $p(t)$ and *the (inverse) translation model* $p(s|t)$.

Although it might seem odd to model the probability of the source sentence given the target sentence, this decomposition has a very intuitive interpretation: the translation model $p(s|t)$ will account for the possible word relations which can be established between input and output language, whereas the language model $p(t)$ will ensure that the output sentence is a well-formed sentence belonging to the target language.

Recently, there have been several efforts, coming from various research groups, to incorporate syntactic information into SMT systems (Kirchhoff et al., 2006; Popović and Ney, 2006a). More specifically, Part of Speech (PoS) tags have been used with the purpose of reordering the input or output sentence and obtaining a monotonous translation (Popović and Ney, 2006b).

In this context, we will be exploring the usefulness of including PoS information within the surface form (i.e. words) in each language, with the purpose of performing a sort of disambiguation over words which cannot be differentiated otherwise. We will be applying this extension on Moses (Koehn et al., 2007b), a phrase based (PB) SMT system.

Similar work was performed throughout the JHU Summer Workshop 2006 (Koehn et al., 2007a), when Moses was first built. In this work, however, factored translation models were used, and results on only a fraction of EuroParl were reported.

The rest of this work is structured as follows: first, in section 2, we will make a brief overview of the state of the art in PoS tagging. In section 3, we will review briefly phrase based SMT systems. In the next section, the experimental setup we carried out is detailed, and the translation results obtained are presented in section 5. Lastly, section 6 presents the conclusions we arrived to.

2 Part-Of-Speech Tagging

As is usual in many fields of Pattern Recognition and Language Modelling, the first PoS taggers were rule based systems (Greene and Rubin, 1971; Brill, 1992). However, the tasks where these systems could be applied belonged to a very restricted field, although their use was enough general to enable them to build tagged corpora, which were later on revised by human experts. These corpora were the key towards developing new, more efficient taggers.

More recently, the statistical framework gained a lot of importance, mainly because of the easiness with which the statistical models could be applied to new tasks. In fact, state-of-the-art PoS tagging is still driven by Hidden Markov Models (HMM), which were first applied by (Church, 1988), and later on by (Brants, 2000), who developed a tagger which still now belongs to the state of the art.

Within this framework, the hidden states represent the tags, whereas the observables are the words in the original corpus. Hence, transition probabilities depend of the origin and target states, i.e., tag pairs. On the other hand, observables only depend on the PoS tag assigned in the emission state. Formally, as defined by (Brants, 2000):

$$\underset{l_1 \dots l_T}{\operatorname{argmax}} \left[\prod_{i=1}^{T} p(l_i|l_{i-1}, l_{i-2}) \cdot p(w_i|l_i) \right] \cdot p(l_{T+1}|l_T)$$

where $w_1 \dots w_T$ is a sequence of words for length T and $l_1 \dots l_T$ are elements of the set of PoS tags. The tags l_{-1}, l_0, l_{T+1} are tags indicating the beginning and the end of the sequence and are added to the set of tags for coherence purposes, but also because their inclusion implies a slight performance increase.

Moreover, Brants introduced a smoothing technique based on unigram, bigram and trigram interpolation, obtaining the following formula for the probability of a trigram:

$$\begin{aligned} p(l_3|l_1, l_2) \;=\; & \lambda_1 \hat{p}(l_3) + \lambda_2 \hat{p}(l_3|l_2) + \\ & \lambda_3 \hat{p}(l_3|l_1, l_2) \end{aligned} \qquad (1)$$

where $\hat{p}$ are the maximum likelihood estimations of the probabilities, and λ_n represents the weights of each one of the n-grams, obeying the restriction $\lambda_1 + \lambda_2 + \lambda_3 = 1$, so that p will remain a probability distribution.

In this work, we will be using the TnT Tagger (Brants, 2000) for tagging the German–English corpus and the FreeLing (Asterias et al., 2006) for tagging the Spanish–English corpus. Both these taggers are HMM based. Although PoS tagging is a monolingual problem and the English side of both parallel corpora could be tagged with the same toolkit, we did not do so because we took advantage of data we already had available, and both taggers present similar precision rates of over 97% (Brants, 2000; Asterias et al., 2006).

3 Phrase-based models

In the last years, phrase based (PB) models (Tomas and Casacuberta, 2001; Marcu and Wong, 2002; Zens et al., 2002; Zens and Ney, 2004) have proved to provide a very efficient framework for MT. Computing the translation probability of a given *phrase*, i.e. a sequence of words, and hence introducing information about context, these SMT systems seem to have mostly outperformed single-word models, quickly evolving into the predominant technology in the state of the art (Koehn and Monz, 2006a).

3.1 The model

The derivation of PB models stems from the concept of bilingual segmentation, i.e. sequences of source words and sequences of target words. It is assumed that only segments of contiguous words are considered, the number of source segments being equal to the number of target segments (say K) and each source segment being aligned with only one target segment and vice versa.

Let I and J be the lengths of t and s respectively[1]. Then, the bilingual segmentation is formalised through two segmentation functions: μ for the target segmentation ($\mu_1^K : \mu_k \in \{1, 2, \ldots, I\}$, $0 < \mu_1 \leq \mu_2 \leq \ldots \leq \mu_k = I$) and γ for the source segmentation ($\gamma_1^K : \gamma_k \in \{1, 2, \ldots, J\}$, $0 < \gamma_1 \leq \gamma_2 \leq \ldots \leq \gamma_k = J$). The alignment between segments is introduced through the alignment function α ($\alpha_1^K : \alpha_k \in \{1, 2, \ldots, K\}, \alpha(k) = \alpha(k')$ iff $k = k'$).

By assuming that all possible segmentations of s in K phrases and all possible segmentations of t in K phrases have the same probability independent of K, then $p(s|t)$ can be written as:

$$p(s|t) \propto \sum_{K} \sum_{\mu_1^K} \sum_{\gamma_1^K} \sum_{\alpha_1^K} \prod_{k=1}^{K} p(\alpha_k|\ \alpha_{k-1}) \cdot p(s_{\gamma_{\alpha_k - 1}+1}^{\gamma_{\alpha_k}} | t_{\mu_{k-1}+1}^{\mu_k}) \qquad (2)$$

where the distortion model $p(\alpha_k|\ \alpha_{k-1})$ (the probability that the target segment k is aligned with the source segment α_k) is usually assumed to depend only on the previous alignment α_{k-1} (first order model).

3.2 Learning phrase-based models

Ultimately, when learning a PB model, the purpose is to compute a *phrase translation table*, in the form

$$\{(s_j \ldots s_{j'}), (t_i \ldots t_{i'}), p(s_j \ldots s_{j'} | t_i \ldots t_{i'})\}$$

where the first term represents the input (source) phrase, the second term represents the output (target) phrase and the last term is the probability assigned by the model to the given phrase pair.

In the last years, a wide variety of techniques to produce PB models have been researched and implemented (Koehn et al., 2003). Firstly, a direct learning of the parameters of the equation $p(s_j^{j'}|t_i^{i'})$ was proposed (Tomas and Casacuberta, 2001; Marcu and Wong, 2002). At the same time, heuristics for extracting all possible segmentations coherent with a word-aligned corpus (Zens et al., 2002), where the alignments were learnt by means of the GIZA++ toolkit (Och and

[1] Following a notation used in (Brown et al., 1993), a sequence of the form $z_i, \ldots, z_j$ is denoted as z_i^j. For some positive integers N and M, the image of a function $f : \{1, 2, \ldots, N\} \rightarrow \{1, 2, \ldots, M\}$ for n is denoted as f_n, and all the possible values of the function as f_1^N

Ney, 2003), were also proposed. Other approaches have been suggested, exploring more linguistically motivated techniques (Sánchez and Benedí, 2006; Watanabe et al., 2003). In this paper, we report experiments using the heuristic, (word) alignment-based phrase extraction algorithm.

3.3 Decoding in phrase-based models

Once a SMT system has been trained, a decoding algorithm is needed. Different search strategies have been suggested to define the way in which the search space is organised. Some authors (Ortiz et al., 2003; Germann et al., 2001) have proposed the use of an $A^\star$ algorithm, which adopts a *best-first* strategy that uses a stack (priority-queue) in order to organise the search space. On the other hand, a *depth-first* strategy was also suggested in (Berger et al., 1996), using a set of stacks to perform the search.

4 Experimental setup

In this section we will be describing the Europarl corpus (Koehn, 2005) on which we performed our experiments, how it is structured, how we added PoS information to the system built, and how this information affects the language model and vocabulary sizes.

4.1 The Europarl corpus

The Europarl corpus (Koehn, 2005) is built from the proceedings of the European Parliament, which are published on the web, and was acquired in 11 different languages. However, in this work we will only focus on the German–English and Spanish–English corpus, due to the fact that it is much easier to find good PoS taggers for these languages.

For our experiments, we used the second version of this corpus, which is the one described in (Koehn, 2005) and the one that was used in the 2006 Workshop on Machine Translation of the NAACL (Koehn and Monz, 2006b). This corpus is divided into four separate sets: one for training, one for development, one for test and another test set which was the one used in the workshop for the final evaluation. This test set will be referred to as "Test", whereas the test set provided for evaluation purposes outside the final evaluation will be referred to as "Devtest". It must be noted that the Test set included a surprise out-of-domain subset, and hence the translation quality on this set will be significantly lower.

Since the original corpus is not sentence-aligned, and not every English sentence has its corresponding translation in German and Spanish (or vice-versa), two different corpora are obtained while constructing the German–English and Spanish–English parallel bilingual corpora. The characteristics of these corpora can be seen in Table 1.

It seems important to point that the average length of the sentences in German is always shorter than the average mean of the sentences in English, and the sentences in English are as well longer than the ones in Spanish. Moreover, the vocabulary size in German is more than 2,5 times bigger than the English vocabulary. This is due to the agglutinative nature of German, that has the ability of building compound words from simple words. For example, "Nachttisch" comes from the words "Nacht" and "Tisch" and means, literally "nighttable". This grants German an enormous lexical richness, but hinders the training of MT systems that involve German, either as source or target language. In addition, the fact that the average sentence length in the training subsets is much shorter than in the other sets is because in the cited workshop the training set was restricted to sentences with a maximum length of 40 words, whereas the other three subsets did not have this restriction.

Since the translations in the corpus have been written by a big number of different human translators, a same sentence may be translated in several different ways, all of them correct. This fact increases the difficulty of the corpus, and can be seen in the number of different pairs that constitute the training set, which is very similar to the total number of pairs. An example is the English sentence "*We shall now proceed to vote.*". It appears translated both as "*Se procede a la*

Table 1: Characteristics of the German–English and Spanish–English Europarl corpus

		German	English	Spanish	English
Training	Sentences	751088		730740	
	Different pairs	735792		715615	
	Running words	15257871	16052702	15725136	15222505
	Vocabulary size	195291	65889	102886	64123
	Average length	20.3	21.4	21.5	20.8
Development	Sentences	2000		2000	
	Running words	55147	58655	60628	58655
	Average length	27.6	29.3	30.3	29.3
	Out of vocabulary	432	125	208	127
Devtest	Sentences	2000	2000	2000	2000
	Running words	54260	57951	60332	57951
	Average length	27.1	29.0	30.2	29.0
	Out of vocabulary	377	127	207	125
Test	Sentences	3064	3064	3064	3064
	Running words	82477	85232	91730	85232
	Average length	26.9	27.8	29.9	27.8
	Out of vocabulary	1020	488	470	502

Table 2: Perplexity of the various corpus subsets with 3-grams and 5-grams.

		3-gram	5-gram
Dev	German	127.6	148.6
	English	74.6	89.9
	Spanish	74.2	89.0
Devtest	German	128.8	149.8
	English	73.7	88.9
	Spanish	75.3	90.6
Test	German	199.7	221.1
	English	118.5	134.5
	Spanish	103.2	117.9

votación.", which is quite a faithful translation, and "*El debate queda cerrado.*", which means "the debate is now closed". Although these two Spanish sentences are clearly different, one can clearly imagine a scenario where both translations would fit.

In the shared task of the NAACL06 Workshop on Statistical Machine Translation, the baseline system used 3-grams as language model, whereas in the shared task of the ACL07 Workshop, which used a newer and somewhat bigger version of the Europarl corpus, the baseline system was constructed with a language model consisting on 5-grams. Since we will be performing experiments both with 5-grams and with 3-grams, the perplexity of the various subsets of the corpus are shown in Table 2. These language models were computed with the SRILM (Stolcke, 2002) toolkit, applying interpolation with the Kneser-Ney discount.

4.2 Preparing the system

Before training the translation models, we PoS tagged all the subsets of the two corpora, obtaining a tagged bilingual corpus. Then, we concatenated the PoS tag to each one of the words, obtaining an extended vocabulary and producing two new different *"languages"*. Although the PoS taggers used have very high success rates, the fact of learning a translation model that involving PoS tags introduces noise in the system, and the error rates of the PoS tagger must affect the final translation quality. Nevertheless, we expect that the benefit obtained will be higher than the error introduced.

Given that for translating we will also need a target language model, we trained three new language models, one for each of the new *"languages"* that was produced by adding the PoS

Table 3: Perplexity of the various corpus subsets with concatenated PoS tags.

		3-gram	5-gram
Dev	German^PoS	129.9	151.1
	English^PoS	77.0	89.9
	Spanish^PoS	74.0	89.0
Devtest	German^PoS	130.9	152.0
	English^PoS	76.1	88.9
	Spanish^PoS	75.1	90.4
Test	German^PoS	202.7	223.7
	English^PoS	124.5	134.5
	Spanish^PoS	102.9	117.7

tags. Their with respect to the different subsets of the corpus is shown in table 3. It can be seen that the perplexity does not suffer an important variation by introducing the PoS tags. This seems encouraging, since it implies that adding the PoS information does not necessarily mean that the language model will be worse. However, it must also be taken into account that the vocabulary sizes do increase significantly: in the case of German, the size increases from 195291 to 212929, in the case of Spanish from 102886 to 109634 and in the case of English from 65889 to 81436, in the German–English subcorpus, and from 64123 to 79229 in the Spanish–English subcorpus. This means an increment of about 10% for German, 5% for Spanish and 22% for English. The fact that it is in English where the vocabulary size is most increased can be explained because of the relatively small vocabulary size that the original English corpus has: since there are fewer words, each word is bound to have, in average, a higher number of different syntactic functions, and hence will be assigned to a wider range of different PoS tags.

5 Translation Experiments

For our translation experiments we used the Moses toolkit (Koehn et al., 2007b). This toolkit involves the estimation of four different translation models, which are in turn combined in a log-linear fashion by adjusting a weight for each of them by means of the MERT (Och, 2003) procedure. For this purpose, a held-out corpus was used, namely the "Development" subset described in section 4.1.

Following previous works in SMT, and for comparability purposes, we will be evaluating our system with BLEU (Papineni et al., 2001) and WER. BLEU measures the precision of unigrams, bigrams, trigrams and 4-grams with respect to a set of reference translations, with a penalty for too short sentences. The WER criterion computes the minimum number of editions (substitutions, insertions and deletions) needed to convert the translated sentence into the sentence considered as ground truth. WER is a pessimistic measure when applied to MT.

Once the different corpus subsets had been tagged, we trained three different translation models.

The first one, which we used as baseline, was trained by applying the Moses toolkit directly. The second one was trained with the extended vocabulary corpus, using the extended words throughout the whole training and translation (decoding) process. Finally, a third translation model was learnt by using the extended vocabulary only to obtain the word alignments, necessary for the phrase-extraction algorithm to obtain phrases. The results can be seen in table 4. In all cases, we used a 5-gram language model, which is the one used as baseline for the 2007 Workshop in Machine Translation of the ACL.

In this table, the column "word^PoS" shows the results for the second experimental setup described above. The last column presents the results obtained by only using the extended vocabulary for alignment purposes.

Unfortunately, in the case of "word^PoS" almost all the results obtained are slightly (although not significantly) worse than those obtained with the baseline system. In the case of "pos-align", most of the results obtained improved by some tenths the baseline, except for the case of English→Spanish. On the other hand, adding PoS information seems to perform slightly better on the *test* set, where out-of-domain sentences were added. However,

Table 4: Translation scores when extending the vocabulary with the PoS tags.

		baseline		wordˆPoS		pos-align	
pair	subset	WER	BLEU	WER	BLEU	WER	BLEU
Es-En	Devtest	57.7	31.6	57.8	31.5	57.5	31.7
	Test	57.8	30.6	58.1	30.3	57.5	30.8
En-Es	Devtest	58.4	31.3	58.7	31.1	58.6	31.0
	Test	57.5	30.3	57.7	30.2	57.6	30.1
De-En	Devtest	65.5	26.2	65.5	26.2	65.0	26.3
	Test	68.1	23.7	68.7	23.7	67.5	24.1
En-De	Devtest	71.6	18.8	71.3	18.9	71.3	18.9
	Test	72.5	16.4	72.6	16.4	72.5	16.5

these slight improvements are not statistically significant.

Only as a small experiment, we checked what would the situation be if the language model used was a 3-gram instead of a 5-gram. In this case, and for the pair German→English, the score was boosted by 1.4 BLEU points on the *devtest* subset, from a 24.55 baseline score to a 25.95 obtained in the “wordˆPoS” setting. Quite interestingly, the score obtained in this setting is almost the same (just two tenths less) than the one obtained with a 5-gram. Hence, PoS information might be more useful in a task where the amount of data available is lower.

6 Conclusions

The results shown in this paper are discouraging in the sense that they seem to imply that adding PoS-tag information does not yield significant improvements on the quality of the final translation produced.

However, this might be so in the case of the EuroParl corpus, where a fairly big amount of data is available. Nevertheless, the use of PoS-tag information could be explored in tasks where the amount of training data is sparser. As future work, we plan to investigate this.

Acknowledgements

This work has been partially supported by the EC (FEDER) and the Spanish MEC under grant TIN2006-15694-CO2-01, by the MEC scholarship AP2005-4023 and by the Universidad Politécnica de Valencia with the ILELA project.

References

J. Asterias, B. Casas, E. Comelles, M. González, L. Padró, and M. Padró. 2006. Freeling 1.3: Syntactic and semantic services in an open-source nlp library. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy.

A.L. Berger, P.F. Brown, S.A. Della Pietra, V.J. Della Pietra, J.R. Gillet, A.S. Kehler, and R.L. Mercer. 1996. Language translation apparatus and method of using context-based translation models. In *United States Patent 5510981.*

T. Brants. 2000. TnT - a statistical part-of-speech tagger. In *Proceedings of the Sixth ANLP*, Seattle, WA.

Eric Brill. 1992. A simple rule-based part of speech tagger. In *Proceedings of the Third ANLP*, Trento, Italy.

P. F. Brown, S. A. Della Pietra, V. J. Della Pietra, and R. L. Mercer. 1993. The mathematics of machine translation. In *Computational Linguistics*, volume 19, pages 263–311, June.

K.W. Church. 1988. A stochastic parts program and noun phrase parser for unrestricted text. In *Proceedings of the 1st ANLP*, pages 136–143, ACL.

U. Germann, M. Jahr, K. Knight, D. Marcu, and K. Yamada. 2001. Fast decoding and optimal decoding for machine translation. In *Proceeding of the 39th. Annual Meeting of the ACL*, pages 228–235, Toulouse, France.

B.B. Greene and G.M. Rubin. 1971. Automated grammatical tagging of English. In *Technical*

Report, Department of Linguistics, Brown University.

K. Kirchhoff, M. Yang, and K. Duh. 2006. Statistical machine translation of parliamentary proceedings using morpho-syntactic knowledge. In *TC-STAR Workshop on Speech-to-Speech Translation*, pages 57–62, Barcelona, Spain, June.

P. Koehn and C. Monz. 2006a. Manual and automatic evaluation of machine translation between european languages. In *Proceedings of the NAACL 2006, Workshop on SMT*, pages 102–121, New York City.

P. Koehn and C. Monz, editors. 2006b. *Proceedings on the Workshop on Statistical Machine Translation*. Association for Computational Linguistics, New York City, June.

P. Koehn, F.J. Och, and D. Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Conf. of the NAACL on Human Language Technology*, volume 1, pages 48–54, Edmonton, Canada.

P. Koehn, M. Federico, W. Shen, N. Bertoldi, O. Bojar, C. Callison-Burch, B. Cowan, C. Dyer, H. Hoang, R. Zens, A. Constantin, C. Corbett Moran, and E. Herbst. 2007a. Open source toolkit for statistical machine translation: Factored translation models and confusion network decoding. In *Final Report of the 2006 Language Engineering Workshop*, John Hopkins University, Center for Speech and Language Processing.

P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. 2007b. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the ACL 2007*, Prague, Czech Republic, June.

P. Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT Summit*.

D. Marcu and W. Wong. 2002. Joint probability model for statistical machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP02)*, Pennsylvania, Philadelphia, USA.

F.J. Och and H. Ney. 2003. A systematic comparison of various statistical alignment models. In *Computational Linguistics*, volume 29, pages 19–51.

F.J. Och. 2003. Minimum error rate training for statistical machine translation. In *ACL 2003: Proc. of the 41st Annual Meeting of the ACL*, Sapporo, Japan, July.

D. Ortiz, I. García-Varea, and F. Casacuberta. 2003. An empirical comparison of stack-based decoding algorithms for statistical machine translation. In *New Advance in Computer Vision, Springer-Verlag, Lecture Notes in Computer Science, 1st Iberian Conference on Pattern Recongnition and Image Analysis (IbPRIA2003)*, Mallorca, Spain.

Papineni, A. Kishore, S. Roukos, T. Ward, and W. Jing Zhu. 2001. Bleu: A method for automatic evaluation of machine translation. In *Technical Report RC22176 (W0109-022)*, IBM Research Division, Thomas J. Watson Research Center, Yorktown Heights, NY.

M. Popović and H. Ney. 2006a. Error analysis of verb inflections in spanish translation output. In *TC-STAR Workshop on Speech-to-Speech Translation*, pages 99–103, Barcelona, Spain, June.

M. Popović and H. Ney. 2006b. Pos-based word reorderings for statistical machine translation. In *5th International Conference on Language Resources and Evaluation (LREC)*, pages 1278–1283, Genoa, Italy, May.

J.A. Sánchez and J.M. Benedí. 2006. Stochastic inversion transduction grammars for obtaining word phrases for phrase-based statistical machine translation. In *Proceedings of the Workshop on SMT*, pages 130–133, New York City.

A. Stolcke. 2002. Srilm - an extensible language modeling toolkit. In *Proceedings of the International Conference on Spoken Language Processing*, volume 2, pages 901–904.

J. Tomas and F. Casacuberta. 2001. Monotone statistical translation using word groups. In *Proceedings of the Machine Translation Summit VIII*, pages 357–361, Santiago de Compostela, Spain.

T. Watanabe, E. Sumita, and H.G. Okuno. 2003. Chunk-based statistical translation. In *Proceedings of the 41st. Annual Meeting of the ACL*, Sapporo, Japan.

R. Zens and H. Ney. 2004. Improvements in phrase-based statistical machine translation. In *Proceedings of the Human Language Technology Conference (HLT-NAACL)*, pages 257–264, Boston, USA.

R. Zens, F.J. Och, and H. Ney. 2002. Phrase-based statistical machine translation. In *Advances in artificial intelligence. 25. Annual German Conference on AI. Lecture Notes in Computer Science*, volume 2479, pages 18–32.

Moses: Moving Open Source MT towards Linguistically Richer Models

Philipp Koehn
University of Edinburgh, UK

The Moses open source machine translation system has been developed since 2006. Not only a competitive platform for building machine translation system has been created, it was also widely adopted by the research community as the baseline or reference system.

One aspect of recent research efforts that are implemented in Moses are statistical machine translation models that leverage linguistically annotated text, and thus allow the generalization of linguistic categories as well as the enforcement of linguistic contraints.

Recent Advances in Spoken Language Translation

Marcello Federico

FBK, Trento, Italy

The talk is structured in three parts. The first part overviews problems and approaches to spoken language translation. The second part presents challenges and achievements of the European Project TC-STAR, that ended in 2007. The third part describes advances in the use of confusion networks as interface between automatic speech recognition and machine translation. In particular, I will discuss an efficient implementation of a decoding algorithm for confusion networks and describe its use to translate ASR word lattices and to enrich translations with punctuation marks. I will also report experimental results on the translation of speeches of the European Parliament, from Spanish to English and viceversa.

Combining Approaches to Machine Translation: the DCU Experience

Andy Way

DCU University

Until quite recently, having a 'hybrid' MT system meant enriching rules in a transfer-based system with statistics in order to constrain the processing of the system depending on different contexts.

We have conducted a number of novel pieces of research where this concept of 'hibridity' has been extended to allow sources of information other than just 'rules' and 'statistics' to be combined to good effect. These include:

- comparing EBMT and word-based SMT [Way & Gough, 2005]
- combining chunks from EBMT and PB-SMT [Groves & Way, 2005a/b, 2006]
- adding statistical language models to EBMT [Groves & Way, 2005b, 2006]
- (attempts at) combining chunks from two different EBMT systems augmenting PB-SMT with subtree pairs [Tinsley et al., 2007]
- incorporating supertags into PB-SMT [Hassan et al., 2006, 2007, 2008]
- adding source language context into PB-SMT [Stroppa et al., 2007]
- combining examples, statistics and rules in tree-based translation [Hearne & Way, 2003, 2006]

We will present the rationale behind these pieces of research, describe the various improvements made, and comment on other possible system combinations which might improve system performance further.