Using corpus information to improve MT quality

Gregor Thurmair

linguatec
Gottfried-Keller-Str. 12, 81254 Munich
g.thurmair@linguatec.de

Abstract

The paper presents activities to improve the quality of a rule-based MT system, using corpus information. If focuses on the area of dictionary, and shows how, and to which extent, corpus-based information can improve the system quality in the different areas of dictionary development. It deals with two main sources of errors: missing entries / translations, and wrong selections of one out of several possible translations.

1. Baseline

Quality of machine translation is still a critical topic; however, research has not really focused on this issue; instead there were many attempts to start anew, hoping that a change in technology would lead to improved system quality. However, up to now, this has not proven to be the case. A comparison of the different approaches, rule-based MT and statistical MT (henceforth SMT), found two main results (cf. Thurmair, 2005, comparing German-to-English MT):

- The overall quality of SMT is outperformed by existing rule-based MT systems.
- The overall quality of both approaches is not yet sufficient. Between 20 and 30% of the evaluated sentences were ranked as being unacceptable.

A closer evaluation of the results shows that the main sources of errors in SMT (about 60%) are related to phenomena like German split verb constructions, non-standard constituent ordering, gapping etc., all of which could be rather easily described in a rule-based context; while the main sources of errors in rule-based systems (again about 60%) consist of lexical issues, and wrong selection of lexical material, which in turn a corpus-based approach can easily avoid[1].

As a consequence, efforts seem to be adequate to merge the power of corpus-based methods with the advantages of a rule-based system architecture, starting with the dictionary, identified as one of the major weaknesses of current MT systems.

Lexical mistakes, in general, result from two sources: no translations exist in the dictionary, and too many translations exist, and a wrong one is selected.

2. Missing transfers

Missing entries damage not just the translation (as they cannot provide content fidelity), they also hamper the analysis of the rest of the text. They result from three main sources: missing general purpose words, missing special terminology, and proper name issues.

2.1. Gaps in general vocabulary

The most straightforward case is dictionary gaps. But current MT dictionaries contain several 100 K entries, and gaps are not so easy to identify. On the other hand, in most existing MT dictionaries, surprising entries can be found. Experiments have shown (Dillinger, 2001) that MT dictionaries contain a significant amount of entries on which coding effort has been spent, but are nearly never used.

Obviously, corpus-based technologies of monolingual and bilingual term extraction can be used to close dictionary gaps, based on frequency information[2]. In the context of linguatec's 'Personal Translator', missing entries with a frequency more than 5000 were identified and added to the system dictionaries.

2.2. Corpus-based terminology

Beyond general vocabulary words, there is a huge amount of terms not represented in MT dictionaries, mainly terminology for special domains. Corpus-based techniques here are to be preferred to conventional dictionary entering:

1. Studies in the automotive sector showed that even special domain dictionaries with high reputation, in a significant amount of cases, propose translations which sound plausible but are not at all used in the target language[3]. Using such translations can make the text not understandable.

2. Often it is required to meet special user terminology requirements. E.g. if users allow for cross-lingual searches on their web sites, terms must be translated in a user specific way, otherwise they lead to poor search results.

Again, corpus-based work is required to provide adequate terminology, given the fact that special terminology has multiple translations.

2.3. Proper Names

Proper names are another large source of unknown words. Although they form a considerable amount of the vocabulary, only recent research (Babych and Hartley, 2003; Jiménez, 2001) into proper names shows their potential for quality improvement.

Proper names cannot be stored in dictionaries, as there is a too large and ever growing amount of them. But users often are puzzled if proper names are treated incorrectly.

1. **Not treating** proper names at all often results in parsing errors, like with other missing lexical elements in

[1] The rest consists mainly in grammatical mistakes; wrong structures are selected in a given situation.

[2] The linguatec corpus for German and English, collected for the work presented here, consists of 700-800 million word forms each.

[3] This can easily be verified by searching for them in the Internet.

the input. In addition, the fidelity criterion is always violated if names which are used in the source text do not show up in the translation.

2. Such problems can be avoided if proper names are marked to be '**don't-translate**' words, as is possible in some systems[4]. Then the proper names undergo some default system treatment (usually: noun with some default values for gender and number). However, this can be incorrect as proper names have different syntactic properties: They inflect (like in Russian or German), they differ in number (plurale tantum like *the Hebrides*, *les Pyrénnées*), they take special prepositions, etc.; so more information is needed than just the default.

3. Therefore, a full **named entity recognition** component is required to improve the analysis, by providing information about constituency and attachment (*He robbed [the Bank of Scotland]* vs. *He robbed [the Bank] [of Scotland]*) and semantic type of proper names.

Named Entity recognition often uses statistical or shallow parsing technology, and there are two options of integration into an MT system: running as some pre-processor, or being integrated into the full syntactic analysis. Full integration tends to be less robust (in case of parsing errors), but can better deal with homographs (de *Peter Maurer war Maurer* -> en *Peter Maurer was a bricklayer*) or gender issues (en *Anna Frank was a teacher* -> *Anna Frank war Lehrerin*)[5]. In addition, there is another feature of Named Entity recognisers, namely coreference analysis, which affects conventional MT system structure: Coreference is a feature which is text based, and MT systems which are sentence-based cannot really cope with it. In the following example, while the first occurrence of *Schneider* is recognised by contextual analysis, sentence-based MT systems fail to identify it in the third sentence, and therefore incorrectly translate the name there:

Das FDP-Mitglied Dr. Schneider lebt in München. Dort ist es heiß. Schneider ist der erste ausländische Politiker.

The FDP member Dr. Schneider lives in Munich. It is hot there. Tailor is the first foreign politician. (instead of: *Schneider is the first foreign politician*).

4. A special challenge consists in the **translation** of proper names. This is where MT systems need to extend standard NE recognisers, which only identify their entities. While it is a common mistake of MT systems to translate proper nouns (en *Mrs. Rice* -> de **Fr. Reis*, de *Hr. Fischer* -> en **Mr. Fisherman*), it is only true for person names that they must not be translated[6]. Dates usually must be translated to accommodate to the target language's conventions. Places behave differently: some are translated (en *Ivory Coast* -> fr *Côte d'Ivoire* -> de *Elfenbeinküste*), others are not (e.g. *Santiago de Compostela*). Often such place names are put into the dictionary.

Proper names can also have different linguistic properties in source and target language, which is relevant for generation: The *Désert du Thar* is masculine in French but *Thar Wüste* is feminine in German, and so is *Rhône* where even the lemma is identical in both languages. *Balkan* is singular in English but plural in Russian (*Балканы*). For product names, the gender seems to be dependent on the ‚base type': cars like *Renault* default to be masculine in German (derived from *der Wagen*) but feminine in French (derived from *la voiture*); determiner placement is language specific as well:

fr *L'Italie* -> de *Italien* but
fr *La Suisse* -> de *die Schweiz*.

While some of these cases can be handled by default assumptions, others are idiosyncratic, need to be detected by corpus work (cf. Jiménez 2001) and require a special resource to describe them.

5. The **evaluation** of integrating a named entity component into an MT system (the linguatec 'Personal Translator') was done as follows: A total of 1500 sentences from the news domain was selected in three language directions, 15% of which contained proper names. They were analysed with and without the proper name recogniser, and the results were compared.

The evaluation showed an increase in translation quality for sentences containing proper names by about 30% on average. The main improvements were:

- no erroneous translations of person names, esp. in coreference positions
- better contextual adaptations (correct preposition and determiner selection; and correct pronominalisation)
- better parses in some cases (e.g. segmentation of dates containing periods).

Of course the overall quality gain for a given corpus depends on the number of sentences containing proper names, and will be higher in news text translation than e.g. in computer manuals.

3. Wrong translation selection

While the problem of missing dictionary entries seems to be reducible to a tolerable size, the opposite problem is much more difficult to solve. It consists in an improper selection of a target term from a number of candidate translations. This problem aggravates with growing numbers of dictionary entries and increased system intelligence. And this is what articles like 'Have fun with MT' refer to:

Wortebene is *word level* and not *word plane*, and *Stromunternehmen* is not a *river expedition* but an *electric power producer*.

The challenge consists in the selection of the proper translation in a given context. It should be noted that dictionaries for humans contain much more translation variants than even large MT dictionaries, which increases the relevance of the problem.

3.1. Current disambiguation means

State-of-the-art systems offer two possibilities to select translation alternatives:

1. **Global settings** by users. Systems provide options for subject area settings, for customer settings (to cover customer-specific terminology), for locales (to select for *truck*$_{US}$ or *lorry*$_{UK}$), for conservative vs. progressive

[4] Babyh & Hartley (2003) tested a recogniser for named entities, and marked all of them as don't-translate words.

[5] Frank et al., 2004

[6] Albeit transliterated, which opens a problem when translating cyrillic or arabic scripts, cf. (Virga and Khudanpur, 2003).

spelling (to select for German *Gemse* vs. *Gämse*), and several other options.

These settings require user interaction, and a level of user skills which often is not available. Also, MT systems linked to search engines do not even ask users for subject area settings.

2. **Linguistic context** description. Such descriptions are coded in the dictionaries as transfer tests; they describe linguistic contexts which trigger the transfer selections:

See (gender = <feminine>) -> sea
See (gender = <masculine>) -> lake
ausführen (dir. object = <person> ->take out
ausführen (dir. object=<program>) -> execute

Such tests can be described as configurations of feature settings of underspecified tree structures[7]. Translation candidates are compared, in a specific order, to the input trees, and if their test configuration matches the input tree configuration then this translation is picked.

Such a technique has two problems to solve:

- In case of parse failures, the structures with which the transfer candidates are compared are erroneous, so the comparison may fail, and a poorer translation is selected
- There are many cases of underspecification, i.e. the information which would trigger a transfer selection is not present: In cases where
  de *Bank* (plural *Bänke*) -> en *bench / benches*
  de *Bank* (plural *Banken*) -> en *bank / bank*
  but the sentence contains only a singular (*er steht vor der Bank*), then the system cannot apply the test, and randomly has to pick a translation, which can be wrong.

Both options, parameter setting and linguistic tests, obviously need improvements in translation selection. For the parameter settings, an obvious solution is to set such parameters automatically.

3.2. Automatic subject area selection

To overcome the problem that not even the options which can be provided by the system (especially subject area selection) are used, a topic identification component has been added to the MT system, to compute to what subject area a text would have to be assigned.

1. There are two main lines of **technology** to build topic identification, or text classification, systems (Jackson and Moulinier, 2002): Selecting classification features (usually words) from an example corpus by machine learning techniques, or using manually selected key words describing the respective topic. While the former crucially depends on the similarity of test corpus and runtime text material, and therefore is less robust, the later depends on a careful selection of key words and tends to have a too small keyword basis. An e.g. in context where an MT system must translate internet material, the selection of a corpus which would be sufficiently similar to the texts to be translated at runtime is a very challenging task, so the second option seems to be preferable.

2. In an MT environment, the most plausible option seems to use the system **dictionary** as a resource for text classification. But although dictionaries are sensitive for subject area selection, they follow a different purpose:

- They use subject area tags only in cases where disambiguation is needed; and for 1:1 translations such a tag assignment often does not need to be assigned, as the respective translation is selected anyway. For a classification tool, however, this is a drawback.
- Also, there are subject areas containing only very few terms (again only the ones which need to be disambiguated), which is not suitable for good classification either.

So, although MT dictionaries can be a good starting point, more intelligence is required.

3. Therefore, a different approach was taken: A large text corpus was searched, starting with some **seed terms** (like '*sports football hockey racing*'); the system returned the highest correlated terms (both single and multiwords) to the seed words, using standard retrieval technique. From the resulting terms, the experts selected the ones which they believed to describe the topic best, and repeated this procedure. For each of the about 40 topics, between 400 and 1500 terms per language were collected to describe it.

These terms were processed with statistical classification tools to compute their relative importance related to the topic in question.

The classification is implemented in such a way that it gives the best (or the several best) subject areas if they match a given threshold, and gives no indication if it is not sure, and leave it to the users to decide; we felt that false assignments would do more harm than no assignment.

4. The **evaluation** of the component shows ambivalent results.

a. For a test corpus of several hundred documents in two languages, the correct subject area was identified in over 80% of the cases, and no false positives were returned. This is quite acceptable.

b. However, correct subject area recognition is just a prerequisite for proper selection of translation alternatives by the MT system. It depends on the organisation of the dictionaries what use of this information the system can make, and how sensitive it is to subject area coding. The result here was that the improvement was not really overwhelming, even if the classifier works fine[8].

During the evaluation, it also turned out that a subject area code rather means that a given translation alternative is unlikely outside of a certain subject area, but it does not mean that within a subject area this translation is always correct. Many general vocabulary terms occur in specific domains both with their special and their general meaning, like (in the automotive domain):

en *project* -> de *Restaurierungsobjekt* vs. *Projekt*
de *Übersetzung* -> en *gear ratio* vs. *translation*

As a result, a subject area test, even if the subject area is recognised correctly, is not the most helpful information for transfer selection; additional means need to be used.

3.3. Neural transfer

[7] An attempt to define a kind-of-standard representation for this has been made in OLIF, cf. (McCormick, 2001)

[8] This, of course, depends on the organisation of the MT dictionary, and may be different in the different systems.

Beyond improving global settings, the linguistic criteria for transfer selection should also be extended.

1. When observing **human behaviour** in transfer selection, it can be seen that people often refer to the conceptual context, to explain that 'even in the automotive domain, '*Übersetzung'* in the context of '*documentation'* and '*language'* and other such terms can only be '*translation'*, not '*gear ratio*''. The question is if such human behaviour can be modelled in an MT system to improve transfer selection using conceptual context.

The task is similar to word sense disambiguation, but applied not to abstract word senses (as in WordNet) but to concrete word senses as represented in different translations. It requires the identification of conceptual contexts which indicate a certain word sense, and consequently a certain translation of a term.

2. As a consequence, all **dictionary entries** with more than one translation were evaluated, and 'clear' cases like

en *teacher*$_{masculine}$ -> de *Lehrer*
en *teacher*$_{feminine}$ -> de *Lehrerin*

were eliminated. From the remaining set, several hundred candidates were selected for further analysis. Each of them was looked up in a standard dictionary to make sure that the most important readings of the term were represented.

3. For each term, a **corpus lookup** was done, using the linguatec corpus, resulting in a couple of thousand contexts per term. Each of these contexts was assigned a reading of the word in question, to enable the formation of clusters of concepts for each reading. These clusters were then statistically analysed, using a standard Bayesian classifier, to identify the most distinctive terms for a given reading, and represented as a neural network[9].

4. Examples of the effect are shown in the following texts, for different translations of *fan* and of *coach* into German *Fan* vs. *Ventilator* and *Trainer* vs. *Bus*, respectively:

(1) en *The fans make noise. The whole club was already drunk when they came to the stadium to support their soccer heroes, although their coaches had to leave.*
=> de *Die Fans machen Lärm. Der ganze Klub war schon betrunken, als sie zum Stadium kamen, um ihre Fußballhelden zu unterstützen, obwohl ihre Trainer abfahren mussten.*

(2) en *The fans make noise. Their rotor does not distribute the air evenly, and the electric motor is not in full operation. All the coaches full of tourists were disappointed.*
=> de *Die Ventilatoren machen Lärm. Ihr Rotor verteilt die Luft nicht gleichmäßig, und der elektrische Motor ist nicht in vollem Betrieb. All die Busse voll von Touristen waren enttäuscht.*

The first sentence is translated differently in the two contexts, although both times identical in the source language. Sentence-based translation is not able to grasp the difference.

5. The next task was the **integration** of the neural networks into the MT system. There are two challenges:

- Like in proper name recognition, neural transfer needs more context than just a sentence; systems with a only sentence-based architecture create artificial limitations. More context is required.
- The neural transfer must be integrated into the transfer selection architecture of the MT systems, and be related to the other transfer selection criteria.

5. The component was **evaluated** as follows: In the German-to-English system, 30 concepts were randomly selected for the tests, and texts containing these concepts were downloaded from the internet, without reading disambiguation. The texts contain 165 occurrences of the test concepts. These sentences were translated, and the result was compared.

Of those, 162 (98%) were correctly translated, using neural transfer. Without neural transfer, just 92 (56%, which is close to random) were correct, so there is an improvement in quality of more than 40%.

Of course the real quality gain depends on the frequency of such concepts in the complete corpus.

4. Conclusion

These examples show that the quality of MT systems is not yet at its limits; it also shows that it will develop in an evolutionary process rather than in a completely new technology.

The most promising approach seems to consist in hybrid system architectures, enriching rule-based approaches (which model the language competence) by corpus-based and statistical techniques (modelling the language performance aspects) as presented above.

5. Acknowledgements

The work presented here was mainly carried out by Vera Aleksić (who also composed the story), Alois Baumer and Thilo Will.

6. References

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

Dillinger. M. (2001). Dictionary Development Workflow for MT: Design and Management. Proc. MT Summit, Santiago de Compostela. Spain.

Frank, A., Hoffmann, Chr., Strobel, M., (2004). Gender Issues in Machine Translation. Univ. Bremen

Jackson, P., Moulinier, I. (2002). Natural Language Processing for Online Applications. Amsterdam (J. Benjamins).

Jiménez, M. (2001). Generation of Named Entities. Proc. MT Summit, Santiago de Compostela. Spain.

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

Thurmair, Gr. (2005). Hybrid architectures for Machine Translation Systems. Language Resources and Evaluation, 2005, 91-108.

Virga, P., Khudanpur, S. (2003). Transliteration of Proper Names in Cross-Language Applications. Proc. SIGIR Toronto.

[9] This is why we call this kind of transfer 'neural transfer'.