Personal Translator at WMT2011

- A rule-based MTsystem with hybrid components -

Vera Aleksić
Linguatec Gmbh
Gottfried-Keller-Str. 12
Munich, Germany
v.aleksic@linguatec.de

Gregor Thurmair
Linguatec Gmbh
Gottfried-Keller-Str. 12
Munich, Germany
g.thurmair@linguatec.de

Abstract

This paper presents the Linguatec submission to the WMT 2011 sixth workshop on statistical machine translation. It describes the architecture of our machine translation system 'Personal Translator' (hereinafter also referred to as PT), developed by Linguatec, which is a rule-based translation system, enriched by statistical approaches.

We participate for the German-English translation direction. For the current submission we have chosen the latest commercial version of the system, PT14. The translation quality improvement for the submission was done mainly by lexicon tuning: detection of unknown words, extracting of possible translations, partly from the wmt11 training corpora, and enlarging the lexicon by manually coding the chosen transfer candidates.

1 Introduction

The origin of the PT technology dates back to the 80's when a translation system based on logic programming and slot grammars was developed by Michael McCord at IBM T.J. Watson Research Center. In many years of development the translation engine has been driven forward and enhanced. Most recently we have added statistical approaches for tasks such as erroneous input correction, subject area recognition and word disambiguation. Today 'Personal Translator' is one of the leading programs in the translation technology field. It is a commercial MT system whose product range includes 7 language pairs, i.e. 14 translation directions, for single users and networks. Linguatec is a leading provider of language-technology software for office use in Germany. In addition to machine translation, we develop and provide commercial products in the fields of speech recognition and speech synthesis. Linguatec is the only company to have won the European Information Technology Prize three times.

2 System fundamentals

Personal Translator is implemented as a modular system which basically consists of the following components:

- the grammar, written in Prolog, based on the concept of slot grammar
- the lexicon, administrated in the data base internally called TransLexis
- additional morphological analysers written partly in C and C++
- hybrid (rule-based and statistical) methods for word disambiguation, subject area recognition and spell-checking
- a range of pre- and post-processing components such as format converters for

Proceedings of the 6th Workshop on Statistical Machine Translation, pages 303–308,
Edinburgh, Scotland, UK, July 30–31, 2011.

html, pdf, doc, txt and rtf formats, sentence splitter, tokeniser, lemmatizer.

As Personal Translator is a commercial system, aiming at providing a complete translator work bench and creating added value for users, it integrates a wide range of advanced features such as:

- Translation memory system for management, creation, analysis and maintenance of TMs, as well as large system modules, containing tens of thousands of sentence pairs
- Translation project management tool, enabling the user to save and administer all important translation settings and project relevant options
- Text to speech functionality to support editing and learning processes such as text revision/correction in the language(s) mastered by the user, or getting a feeling for the correct pronunciation in a foreign language, to name just a few.

2.1 LMT and Slot Grammar

Personal Translator is based on the LMT (Logic programming based Machine Translation). The core of LMT uses the principles of **slot grammar**, a grammatical description system developed originally by Michael McCord[1] at IBM.

Slot grammar is based on the concept of word valence. It is dependency oriented, i.e. each phrase has a head word. Each (head) word is characterised by **slots** which represent empty places in its grammatical surroundings such as subject, object, modifier etc. which can be realised in text or not. The slots represent either **complements** of the head word which have to be defined in the lexicon or **adjuncts** which are rather associated with the part of speech and defined more generally in the grammar rules. The possible **slot fillers** are typified by their morphological, syntactic or semantic properties. The analysis of a word is finished and the phrase is considered as satisfied if the appropriate fillers are found in the text and all (obligatory) slots of the word are filled

[1] McCord (1989); McCord, Vernth (1992)

3 Advanced translation features

There are some well-known restrictions concerning the automatic translation process. One of them is the ability of most MT systems to operate on only one sentence at a time. The same is also true for the PT but only to a limited degree. PT integrates several methods for semantic and context analysis on multi-sentence level and for the identification of concepts which are repeated throughout the text. This applies in particular to the recognition of pronoun references and coreference analysis of proper names, as well as subject area recognition and neural transfer which are described further below.

3.1 Recognition of pronoun reference

Pronouns can refer to other words (their antecedents) which had occurred in the previous text. When translating from German into English and vice versa the fact that e.g. the English personal pronouns *he*/*she* apply only to humans and *it* to all other things, whereas in German *er*/*sie/es* can refer to any noun, has to be considered when searching for appropriate translation:

This is a desk. It is new.
Dies ist ein Schreibtisch. Er ist neu.

versus:

This is a bag. It is new.
Dies ist eine Tasche. Sie ist neu.

The user can either select the translation option „Automatic recognition of pronoun reference“, when translating a continuous text, or deselect it in case of translating lists of independent sentences (as we did for the current submission). If this option is deselected, the PT output for the sentences above reads as follows:

Dies ist ein Schreibtisch. Es ist neu.
Dies ist eine Tasche. Es ist neu.

Also the translation of other words in the context can benefit from correct pronoun reference recognition:

The dogs found biscuits. They ate them.
Die Hunde fanden Kekse. Sie fraßen sie.

versus:

The children found biscuits. They ate them.
Die Kinder fanden Kekse. Sie aßen sie.

The last example demonstrates an improvement in the translation of the verb *eat* which is to be translated into German with *fressen* if its subject is

an animal or with *essen* if the subject is a human. The pronoun *they* in the first sentence refers to dogs (animals), in the second to children (humans) respectively.

3.2 Named entity recognition

The treatment of proper names is a real challenge for machine translation. There is a huge number of proper names, even growing constantly if e.g. the companies and product names are considered. Furthermore, person names are constantly changing in their degree of topicality, so it is not of much use to have Kohl and Fischer in the lexicon when the texts to be translated speak about Merkel and Rösler. As such, the proper names are unsuitable to be primarily stored in the lexicon. The second problem is homography: If a proper name is spelled in the same way as a common word, it is very likely to be translated by an MT system (Brown => Braun; Metzger => Butcher).

Personal Translator integrates a named entity recognition component which runs both:

- as a pre-processing tool: It puts mark-ups on the proper names to exclude them of other pre-processing components such as e.g. spell checker
- as part of the translation process, integrated into the lexicon and the complete analysis-transfer-generation process: Morphological and syntactic analysis/generation bases among other things on semantic roles (person, place…), as the proper names have special inflection patterns and specific syntactic behaviour (preposition slots, appositions etc.).

By this, we could achieve an increase in translation quality of about 30% for sentences containing proper names.[2]

3.3 Word sense disambiguation

Another important issue is the treatment of ambiguous words. Most glossaries contain several million translations, among them large amounts of words with multiple meanings. Traditionally, 'Personal Translator' uses several ways to disambiguate ambiguous words and select the most proper translation:

- Interpretion of gender/number and other morphosyntactic information:
  der Kiefer (m) = jaw
  die Kiefer (f) = pine
  minute (sg) = Minute
  minutes (pl) = Protokoll
- Analysis of slot fillers:
  anmachen (Licht) = turn on (light)
  anmachen (Salat) = prepare (salad)
  anmachen (jmd.) = chat (s.o.) up
  bestehen (auf) = insist (on)
  bestehen (aus) = be made (of)
- Use of orthographic information:
  fest (lower case) = stable, firm
  Fest (capitalised) = celebration
- Definition of different subject area codes for the translations:
  die Mutter (general) = mother
  die Mutter (techn.) = nut

4 Hybrid technology

All these disambiguation methods are labour-intensive in terms of manual coding efforts, and they require, to a certain extent, user interaction (e.g. selecting appropriate options such as subject area) that in turn needs reliable knowledge of the contents to be translated which is often not the case. And not at least, manual setting of the disambiguation contexts is not only inefficient but also prone to errors.

For these reasons Linguatec continually tests new, innovative solutions to reduce manual coding efforts and increase translation quality. Therefore it seemed obvious to try to draw statistical significant, reliable, and empirically-sound information from the immense Linguatec corpus and enrich the RMT with this knowledge. Thus an innovative hybrid component, which has been filed as patent[3], has been developed.

4.1 Neural transfer

We as humans rarely have problems to distinguish between two or more different meanings of a word. The decision happens automatically, supported by accessing the world knowledge stored in our brains. Many efforts have been made to artificially imitate these processes. In linguistics, traditionally ontologies have been created which aim at

[2] cf. Thurmair (2005)

[3] cf. Linguatec Patent „Hybrid transfer selection in Machine Translation" US: 11/885.688, EPA: Nr. 05715789.3

reflecting the relations and the hierarchy in the nature. In information technology, artificial neural networks try to approximate the operation of the human brain. Linguatec's hybrid disambiguation model tries to single out the best translation for a word by identifying its semantic network. We call it 'neural transfer'.

The disambiguation model for the neural transfer has been trained on a significant amount of different contexts for each lexicon entry with multiple translations, where this method could be considered as appropriate. Clusters of different meanings of words were built manually and statistical methods were applied on them in order to identify the most distinctive terms in their surroundings and represent the results in neural networks. The neural transfer technology has been integrated into the PT by modifying the affected lexicon entries, and by adding a pre-processing component which assigns a semantic net to the affected text passage.

The neural transfer enables the PT to 'understand' the context beyond sentence boundaries. Thus it is possible to deliver two different translations for the word *Gericht* (court, dish) in absolutely identical sentences, depending on the textual context:

*Ich kann mich noch an dieses **Gericht** erinnern. Es hat die Klage meiner Firma auf Entschädigung abgewiesen.*
*I can still remember this **court**. It has rejected the complaint of my company on reimbursement.*

versus:

*Ich kann mich noch an dieses **Gericht** erinnern. Es war eines dieser Gerichte aus der Küche der Balkanländer, mit Gemüse und Knoblauch.*
*I can still remember this **dish**. It was one of these dishes from the kitchen of the Balkan States with vegetables and garlic.*

The test results showed an improvement of the translation quality by about 40% for texts containing the affected concepts.

4.2 Automatic subject area recognition

In order to overcome the problems mentioned above (manual coding effort, required user interaction), a component for automatic topic identification has been developed and integrated into the PT. Its principle works in a similar way to neural transfer. The most important difference is that the automatic topic identifier assigns the recognised subject area to the whole text to be translated, whereas the neural transfer can operate on the single paragraph level.

4.3 SmartCorrect

Regarding the enormous amount of texts to be translated, most of which are from internet or other unscanned sources, it is not reasonable to expect from MT users to keep control of correct spelling. Nevertheless, a MT system is only able to translate correctly spelled words. For these reasons most MT systems, as well as text processing programmes, include a spellchecker. The problem is that they mostly just identify the typos/spelling errors and leave it up to the user to choose the correct form from a list of suggestions. This is process which requires intensive user interaction and experience has taught us, that users are not always ready to invest their time. In addition, this can only be expected if the text to be corrected belongs to the language mastered by the user.
Therefore Linguatec developed SmartCorrect which not only recognises spelling errors in the text but also corrects them automatically. Trained on very large corpora, the model is likely to detect the best variant in nearly all cases. Clever enough, it cooperates with the named entities recogniser and thus does not identify unknown proper names as spelling errors. Entries from the user lexicons are also save from SmartCorrect intervention.
However, a major part of the misspelling corrections is already performed in a pre-processing step, which adopts some proven methods[4] to identify and correct frequent errors, such as letter deletion, insertion, substitution, inversion and duplication.

5 WMT2011 Submission

We participate for the German-English translation direction. Linguatec has not used the training corpus because we wanted to submit the results of our general purpose MT system.

The only system tuning consisted of lexicon coding. Unknown words were detected automatically by analysing the test set. Appropriate translations were found, some of them from the training corpus. About 200 terms were manually coded or imported into the PT lexicon.

[4] cf. Habash (2008)

Furthermore, we have observed that the test set contained some spelling errors which have been corrected by SmartCorrect (ca. 150 misspelling corrections were done), for example:

offiziel	=>	*offiziell*
Sympatie	=>	*Sympathie*
enhüllten	=>	*enthüllten*
bessseren	=>	*besseren*
unbwohnbar	=>	*unbewohnbar*
zwiwchen	=>	*zwischen*

Thus, for comparison purposes we translated the test set three times:

- Out-of-the-box PT, without SmartCorrect
- Out-of-the-box PT, with SmartCorrect
- Out-of-the-box PT, with SmartCorrect plus lexicon adaptation

The BLEU score in the first run was 17,0. Interestingly, the BLEU score of the second run did not reflect any improvements caused by correction of typos; on the contrary, it declined by 0,2 from 17,0 to 16,8. However, by manual evaluation of sample sentences we gained a more positive impression of the results. With the third run, after the lexicon coding, a BLEU of 17,1, i.e. a minimal increase compared with the firs run, was achieved. Here again, the manual inspection of random sentences, containing the coded terms, left an impression of some more significant improvements than measured by BLEU.

5.1 Conclusion

Automatic metrics have shown a minimal improvement of translation quality. However, the manual inspection suggested much more significant influences of spelling correction and lexicon coding on the translation adequacy and sentence structure and consequently on the readability of the output than the BLEU score did.

5.2 Combined system submission by DFKI

At WMT 2011 our PT will also participate in the combined translation task in a combination of rule-based and SMT systems submitted by the DFKI[5].

[5] Xu et al.(2011)

6 Outlook

Simultaneously with the current submission a 'hybrid experiment' was performed: An attempt at using SMT methods to improve the transfer selection for coding new entries in PT.

An existing (crawled) parallel corpus in the automotive domain was cleaned, segmented by Liguatec sentence splitter, sentence-aligned by Hunalign (supported by using the Linguatec dictionary), word-aligned by GIZA++ and finally phrase tables were produced by using Moses. The objective was to extract meaningful phrases and their translations which are particularly suitable for import into the PT lexicon and thus generate a glossary.

First a phrase table filter, based on frequency, was applied. Then part of speech information was added to both source and target entries as a basis for filtering linguistically motivated phrases. A glossary was generated. For testing purposes a very small set of about 250 terms, namely those which were unknown in the PT lexicon, was chosen to be imported. On a test corpus of about 320 sentences from the automotive domain the translation quality improvement, measured by BLEU, turned out to be about 3.1% (before coding: 14.87, after coding: 17.97).

We will continue researching in that field.

References

Bogdan Babych, Anthony Hartley. 2003. Improving Machine Translation Quality with Automatic Named Entity Recognition. Proc. EACL-EAMT, Budapest.

Arendse Bernth. 1992. The LMT Book. IBM Deutschland Informationssysteme GmbH Scientific Center Institute for Logic and Linguistics.

Nazar Habash. 2008. Four techniqes for Online Handling of Out-of-Vocabulary Words in Arabic-English Statistical Machine Translation. Proceedings of ACL-08: HLT, Short Papers (Companion Volume), pages 57-60, Columbus, Ohio, USA.

Roland Kaplan and Joan Bresnan. 1982. Lexical functional grammar: A formal system for grammatical representation. In Joan Bresnan (Ed.) The mental representation of Grammatical Relations. MIT Press.

Michael McCord. 1989. A new version of slot grammar. Research report RC 14506, IBM research division, Yorktown Heights.

Michael McCord and Arendse Bernth. 1992. Using Slot Grammar. IBM Deutschland Informationssysteme GmbH Scientific Center Institute for Logic and Linguistics.

Gregor Thurmair. 2004. Using corpus information to improve MT quality. In Yuste Rodrigo, Elia (ed) Paris: ELRA (European Language Resources Association): Proceedings of the Third International Workshop on Language Resources for Translation Work, Reseach & Training (LR4Trans-III)

Gregor Thurmair. 2005. Improving MT Quality: Towards a Hybrid MT Architecture in the Linguatec 'Personal Translator'. International MT Summit X, Phuket. Invited paper.

Gregor Thurmair. 2009. Comparing different architectures of hybrid Machine Translation systems. Proceedings of the Twelft Machine Translation Summit. Ottawa, Canada. p.340-348

Jia Xu, Xiaojun Zhang, David Vilar, Casey Kennington and Hans Uszkoreit. 2011. The DFKI Hybrid Machine Translation System for WMT 2011 - On the Integration of SMT and RBMT. Submission paper for WMT 2011 sixth workshop on statistical machine translation. Edinburgh.