Demonstration of the Dutch-to-English METIS-II MT System

Peter Dirix, Vincent Vandeghinste, and Ineke Schuurman
Centre for Computational Linguistics
Katholieke Universiteit Leuven, Belgium
{peter,vincent,ineke}@ccl.kuleuven.be

1 Introduction

The European METIS-II project[1] (Oct. 2004-Sept. 2007) combines techniques from rule-based and corpus-based MT in a hybrid approach for four language pairs (German, Dutch, Spanish, and Greek to English). We only use a dictionary, basic analytical resources and a monolingual target-language corpus in order to enable the construction of an MT system for lesser-resourced languages. Cutting up sentences in linguistically sound subunits improves the quality of the translation. Demarcating clauses, verb groups, noun phrases, and prepositional phrases restricts the number of possible translations and hence also the search space. Sentence chunks are translated using a dictionary and a limited set of mapping rules. Using bottom-up matching to match the different translated items and higher-level structures with the database information, one or more candidate translations are constructed. A search engine ranks them using occurence frequencies and match accuracy in the target-language corpus.

2 Components

The source-language analysis tools construct a source-language model. This toolset consists of a tokeniser, the TnT tagger trained on the Spoken Dutch corpus, a PoS-based lemmatiser, a chunker, and a subclause delimiter.

The translation model consists of a bilingual Dutch-English dictionary with approximately 110,000 entries and a set of tag-mapping rules between Dutch and English.

The target-language model is based on a target-language corpus, the British National Corpus (BNC). It is processed in an analogous way to the source-language input sentences. The translation engine itself is composed of an expander and a ranker. The expander inserts, deletes, moves and permutes tokens and chunks generated during dictionary look-up and the application of the tag mapping. There are currently some half a dozen rules applying. The ranker is a beam-search, bottom-up algorithm that ranks the proposed translations according to the language model. It does not alter the translations anymore. Finally, a token generator generates the correct word forms, since in all intermediate processes, only lemmas are used.

More information on the different components of the system can be found in (Dirix et al., 2005), (Dirix et al., 2006), and (Vandeghinste et al., 2006). The impact of applying hand-crafted rules is described in (Vandeghinste et al., 2007).

3 Evaluation

Our test set consists of 50 Dutch sentences, selected from newspaper texts, with three human reference translations. These sentences are selected to contain a number of classical difficult MT issues. The system generates several translation alternatives (dependent on beam size, which is 20 for all tests described in this paper), each with a weight. As our sys-

[1]Supported by the 6th European Framework Programme, FP6-IST-003768.

tem is not always capable of generating only one best translation, we present two types of results, namely the average BLEU scores of all the top-weight[2] translations generated for that test sentence ('average' score) and the highest BLEU scores of all the top-weight translations generated for that test sentence ('best' score).

Table 1: BLEU scores

	BLEU
'average'	0.3024
'best'	0.3486

A discussion of the results in Table 1 can be found in (Vandeghinste et al., 2007).

4 Current and future work

Currently, we are adding co-occurrence metrics in order to generate unique top-weight translations. These metrics are used to differentiate the weights of the different translations of a single source-language dictionary entry. It is based on the co-occurence of the differents words of the sentence in the target-language corpus. We also moved to an xml representation of our dictionary in order to better represent complex entities. We allow structural changes and discontinuous entries.

Furthermore, we are developing a post-editing interface. The corrections of human post-editors will result in an aligned corpus of machine-made and corrected translations. The corrected translations can be added to the target-language corpus and will also be used as part of the bilingual dictionary. This can be seen as a kind of supervised machine learning.

5 Related work

Related techniques are context-based machine translation (CBMT), as described in (Carbonell et al., 2006), and generation-heavy hybrid machine translation (GHMT), as described in (Habash, 2003). As in METIS, CBMT does not rely on parallel corpora, but on a large target-language corpus, an optional small source-language corpus and a bilingual dictionary. The translation and target-language generation phases do not require any linguistic knowledge, but use n-grams instead. GHMT uses about the same resources as CBMT, but involves a deep source-language analysis. Initially, the dependency structure of the source language is maintained, but at the end, a source-language-independent generation module rewrites the target language part lexically and syntactically.

[2]The *top-weight* translations are those translations that receive the highest weight.

References

Jaime Carbonell, Steve Klein, David Miller, Michael Steinbaum, Tomer Grassiany, and Jochen Frey, 2006. *Context-Based Machine Translation.* In *MTA 2006: Proceedings of the 7th Conference of the Association for Machine Translation in the Americas, "Visions for the Future of Machine Translation"*, pp. 19–28.

Peter Dirix, Vincent Vandeghinste, and Ineke Schuurman, 2005. *METIS-II: Example-based translation using monolingual corpora – System description.* In *Proceedings of MT Summit X, Workshop on EBMT*, pp. 43–50.

Peter Dirix, Vincent Vandeghinste, and Ineke Schuurman, 2006. *A new hybrid approach enabling MT for languages with little resources.* In *Proceedings of the 16th Meeting of Computational Linguistics in the Netherlands*, pp. 117–132.

Nizar Habash, 2003. *Matador: a large-scale Spanish-English GHMT system.* In *Proceedings of MT Summit IX*, pp. 149–156.

Vincent Vandeghinste, Ineke Schuurman, Michael Carl, Stella Markantonatou, and Tony Badia, 2006. *METIS-II: Machine Translation for Low Resource Languages.* In *Proceedings of the 5th International Confererence on Language Resources and Evaluation (LREC).*

Vincent Vandeghinste, Peter Dirix, and Ineke Schuurman, 2007. *The effect of a few rules on a data-driven MT system.* In *Proceedings of the METIS-II Workshop: New Approaches to Machine Translation*, pp. 27–34.