Svitlana Babych [1] s.babych@leeds.ac.uk
Kurt Eberle [2] k.eberle@lingenio.de
Bogdan Babych [1] b.babych@leeds.ac.uk
[1] Centre for Translation Studies, University of Leeds, Leeds, UK
[2] Lingenio GmbH, Heidelberg, Germany

Development of hybrid Machine Translation systems for under-resourced languages: Automated creation of lexical and morphological resources for MT

1 Introduction

Computer-assisted translation (CAT) tools, such as Translation Memories, Terminology Management applications, Localisation software and Machine Translation (MT) systems, are becoming an integral part of professional translators' workflow, supporting large multilingual collaborative translation projects, as well as individual work of freelance translators [1]. This uptake of CAT tools by the translation industry is motivated by the increasing demand to translate larger volumes of texts quicker and to ensure consistency of translations across documents produced at different times for the product line or the same company.

Modern MT engines, which automatically produce draft translation of previously unseen text, play increasingly important role in the CAT workflow, partly because of their tighter integration with other CAT systems, in particular – Translation Memories and terminological databases. In addition to previously successful scenarios of using MT in controlled language environment, in specialised narrow domains and between closely related languages, MT nowadays becomes useful also in the post-editing scenario between linguistically distant languages, cf. [1, p. 123]: for certain combinations of translation directions and subject domains translator's productivity can be increased by 74% on average [2].

MT systems are usually built either as Rule-Based (RBMT), relying on linguistic rules for analysing source text, bilingual transfer and generating target language output, or statistical systems (SMT), which use automatically aligned sentences and

phrases from parallel corpora (texts produced by human translators) and matches phrases on the source and target sides with statistical techniques. More recent Hybrid MT approaches combine various RBMT and SMT methods.

However, the development of all types of MT systems is dependent on the availability of wide-coverage linguistic resources for the source and target languages. For well-resourced languages, such as English, or German, MT can use large monolingual and parallel corpora, comprehensive electronic dictionaries and morphological databases. Such data is not available for under-resourced languages, so the quality of MT for them is significantly lower.

Our FP7 project HyghTra (2010-2014, funded under Marie Curie Industry-Academia Partnership and Pathways scheme: http://www.hyghtra.eu/) aims to create an open infrastructure for rapid development of MT-oriented linguistic resources for under-resourced languages, and integrating them into Lingenio's RBMT architecture [3]. The project will allow us to build high-quality MT systems between English/German and a range of new languages, such as Ukrainian, Russian, Dutch and Spanish. Unlike traditional hybrid systems, which add some aspects of linguistic representations to mainstream SMT engines, HyghTra project preserves all linguistic complexity of the high-quality RBMT engine, and uses statistical techniques for automatically creating rich wide-coverage linguistic representations from corpora.

In this paper we describe two applications of our methodology of creating lexical and morphological resources for MT: the development of the Ukrainian Part-of-Speech tagger and *de/het* noun classification for Dutch.

2 Rapid development of a Part-of-Speech tagger for Ukrainian

Part-of-Speech (PoS) taggers are automated annotation tools, that for each word in a sentence or text can determine its part-of-speech code and values of its derivational and inflectional grammatical categories, e.g., for Slavonic languages – the morphological case, number and gender for nouns and adjectives, or tense, person and number for verbs, etc. In the case of morphologically ambiguous words, e.g., the English word *changes* (Noun or Verb), PoS taggers assign a set of the most probable

values based on their immediate context. Morphological information is essential for MT in order to disambiguate source sentences and to apply correct transfer rules (in the case of RBMT and Hybrid MT), or to use appropriate lexical and morphological features in statistical models (in the case of SMT), e.g.:

The question.$_N$ changes.$_V$ every day – Питання міняється щодня

The question.$_N$ changes.$_N$ have been difficult$_A$ – Зміна питань була складною

The question.$_N$ changes.$_N$ have been agreed$_V$ – Зміну питань було погоджено

Statistical part-of-speech taggers, such as the TreeTagger [4] or TnT [5] usually separate language independent disambiguation algorithms and language-specific parameter files, which may contain morphological dictionaries for a given language and its descriptors of allowed PoS configurations with their frequencies.

We developed Ukrainian parameter files for the TnT tagger using the following method. (1) We manually derived systematic mappings between morphological features in Russian and Ukrainian sets of PoS codes (tagsets) using the description of these tagsets in MULTEXT-East project [6], [7]. A set of around 200 partially ordered rules translates morphological features specified for any Russian tag into a correctly formed Ukrainian tag. (2) The parameter file that describes frequencies of allowed Russian tag combinations (containing around 600,000 3-tag sequences with their frequencies) was mapped into the Ukrainian tag set. Normally for building such a file for a new language developers need morphologically labeled and manually disambiguated corpus of over 1 million words, which is not available for Ukrainian in the public domain. However, our mapping method is based on the assumption that closely-related languages have similar classes of morphosyntactic contexts, since they share the same principles of forming syntactic links, e.g., agreement of grammatical values in linguistic phrases. (3) The second Ukrainian parameter file – the Ukrainian lexicon with potentially ambiguous codes for 200,000 inflected forms was re-formatted from the lexicon of around 25,000 Ukrainian lemmas available from the MULTEXT-East project. This lexicon gives only 87% coverage of the text (evaluated on BBC Ukrainian Service website). However, tags for the remaining13% unknown words in corpus are often successfully reconstructed from neighboring tags.

(4) The TnT tagger with Ukrainian parameter files was used to tag a Ukrainian news corpus, 250 million words (http://smlc09.leeds.ac.uk/internet2.html) The Ukrainian parameter files are freely available for research purposes from our website: http://smlc09.leeds.ac.uk/svitlana/tnt/ua/. We use the Ukrainian PoS tagger in our project to build Ukrainian-English and Ukrainian-German SMT systems.

3 Automatically deriving *de/het* classifications for Dutch nouns

Nouns in Dutch belong to one of the two gender classes which define the choice of determiners: neuter nouns take determiners *het, dat, dit, ons*, and nouns with the common gender take *de, die, deze, onze*. Nouns can only be disambiguated when used as singular and take a definite determiner, so not all contexts in corpus can be useful for disambiguation. For MT task this information needs to be supplied by the target language generation rules, since it is normally not present in the source text, and cannot be derived from application of transfer rules. In our experiment TiMBL / Frog tagger and lemmatiser (Van den Bosch et al., 2007), was used to automatically annotate a 60-million-word section of the balanced Dutch SoNaR corpus (Oostdijk et al., 2008).

Prediction of the *de/het* classes was performed by a set of rules, which cover most typical contexts, where these determiners are distinguished. If both determiners were found in the context, then the class that has the majority of contexts was assigned. Regular expressions covered simple contexts like *Det (Adj)? Noun (*e.g., *de nieuwe geschischrijving)*, but not more complex ambiguous contexts, e.g., sequences of nominal compounds *("waar is de apparaat-code van mijn kamera?")*: we assume that such contexts are less frequent and error rate will be limited. Since TiMBL/Frog also provides independent lexical information about de/het classes, we were able to evaluate the performance of our method (Table1):

	Nouns in corpus	Disambiguated	Percent
Total	157,066	74,505	45.9% (Recall)
Correct		72,088	96.8% (Precision)

Table 1 Performance of de/het disambiguation from corpora.

The results indicate that our method is highly successful if it has disambiguation context for a noun (i.e., a noun was used in singular with a definite determiner), but such contexts are available only for 45.9% of nouns found in corpus.

Future work in HyghTra project will include automatic acquisition of inflectional paradigms for lexical items, attachment preference detection, automatic acquisition of lexical functions, subcategorisation frames and word order, which will allow a much quicker development of Hybrid MT systems for new translation directions which include under-resourced languages.

References

[1] Hartley, A. (2008). 'Technology and translation'. In: Munday, J (ed.) (2008). *The Routledge Companion to Translation Studies*. Routledge. 106-127.

[2] Plitt, M. & Masselot, F. (2010). A Productivity Test of Statistical Machine Translation: Post-Editing in a Typical Localisation Context'. In: *The Prague Bulletin of Mathematical Linguistics*, 93. 7–16.

[3] Babych, B., Eberle, K., Geiß, J., Ginesti-Rosell, M., Hartley, A., Rapp, R., Sharoff, S., and Thomas, M. (2012) Design of a hybrid high quality machine translation system. In: *Proceedings of Joint Workshop on Exploiting Synergies between Information Retrieval and Machine Translation (ESIRMT) and Hybrid Approaches to Machine Translation (HyTra)* at EACL-2012. Pp. 101-112.

[4] Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. In: *Proceedings of International Conference on New Methods in Language Processing*, Manchester, UK.

[5] Brants, T. (2004). TnT: a statistical part-of-speech tagger, Proceedings of the sixth conference on Applied natural language processing, p.224-231, April 29-May 04, 2000, Seattle, Washington

[6] Erjavec, T. (2012): MULTEXT-East: Morphosyntactic Resources for Central and Eastern European Languages. Language Resources and Evaluation, 46/1, pp. 131-142.

[7] Kotsyba, N., Shevchenko, I., Derzhanski, I. and Mykulyak, A. (2010) MULTEXT-East Morphosyntactic Specifications, Version 4. 3.11. Ukrainian Specifications: URL: http://nl.ijs.si/ME/V4/msd/html/msd-uk.html. Accessed on 10/02/2013

[8] Van den Bosch, A., Busser, G.J., Daelemans, W., and Canisius, S. (2007). An efficient memory-based morphosyntactic tagger and parser for Dutch, In F. van Eynde, P. Dirix, I. Schuurman, and V. Vandeghinste (Eds.), Selected Papers of the 17th Computational Linguistics in the Netherlands Meeting, Leuven, Belgium, pp. 99-114.

[9] Oostdijk, N., M. Reynaert, P. Monachesi, G. van Noord, R. Ordelman, I. Schuurman, V. Vandeghinste. From D-Coi to SoNaR: A reference corpus for Dutch. In: LREC 2008.