Development of an open-source example-based Machine Translation system guided by Information Extraction

Bogdan Babych
<bogdan@comp.leeds.ac.uk>

Problem statement

Recent advances in Machine Translation (MT) technology made it much more useful for professional translators. MT considerably increases the productivity of translators' work by automating the search for direct translation equivalents and previously translated fragments and allowing the translators to concentrate on more creative tasks, which go beyond the direct translation strategy, e.g., on translating items that do not have available translation equivalents or cannot be decomposed into sequences of previously translated fragments. Data-driven approaches to MT, such as Statistical (SMT) and Example-Based Machine Translation (EBMT), as well as the use of large corpora for the development of Rule-Based translation systems (RBMT) allowed the developers to overcome the most serious technological bottleneck – the problem of data acquisition. Modern MT systems may rely on large collections of aligned human translations, semi-automatically constructed dictionaries, terminological databases, wide-coverage grammars, ontologies and other extensive knowledge sources, which increase their coverage and make them scalable to new subject domains.

However, these developments also reveal limitations of the current approaches to MT (limitations which were less visible before, since the data acquisition problem used to be much more serious). In the first place, the limitations now come on the processing side, hindering MT from taking full advantage of the available data sources. For example, MT systems often need to filter out redundant translation equivalents (items not intended for translation) (Babych and Hartley, 2004b: 629), which may be as beneficial for MT quality as the widely suggested use of extensive knowledge sources. Strategies of handling the databases of translation equivalents need to become more flexible, they need to be extended beyond the direct "look-up" procedures and systematically cover also oblique translation procedures used by human translators, such as "transposition" or "modulation" (Vinay and Darbelnet, 1995). To a certain extent these cases require simulation of advanced aspects of human intelligence, since many translation problems are non-trivial and require creative and flexible solutions.

It is likely that no single approach and no single system architecture will be able to solve all theoretical and technological problems in MT. The experiments that properly accommodate different "ideologies" are much more promising (e.g., Imamura et al., 2004: 99-105). The lack of a wider technological environment for translation models, where more general models from the field of Artificial Intelligence (AI) can be implemented, seriously impedes the progress in achieving better MT quality (c.f. Key, 2003: xix). Better results could be attained if MT developers concentrate on combining specific natural language processing (NLP) and AI techniques in the process of translation, which addresses a diverse variety of particular translation problems, instead of looking for a single overreaching MT model, methodology or architecture (e.g., statistical, syntax-driven, compositional, etc.). This requires a common experimental ground, where different techniques and approaches may be implemented, integrated with other modules and where their influence on the general MT performance can be evaluated, where different NLP groups can join forces in solving particular MT problems. On the other hand such common experimental ground can become a test bed for evaluating NLP modules, where the developers can test usability of their technologies for MT quality (and more generally – for natural language understanding and/or natural language generation quality), and not just claim in their papers that a particular

technology can be useful or even plays a “vital role” in MT applications (e.g., Mitkov, 2002: xii).

Open-source software systems proved to be very successful as co-operative experimental frameworks for NLP research groups. Examples of such systems that are being developed jointly by the NLP community include GATE (http://www.gate.ac.uk), which is used for Information Extraction and Text Mining tasks by academic, governmental and commercial organisations, and NLTK (http://nltk.sourceforge.net/) used for teaching Computational Linguistics. These systems adopt modular architecture and supply a framework for interaction between different modules, which can be implemented and tested independently of each other. The systems already include many modules, which are usable for MT technology, e.g., the Named Entity recognition module in GATE or the Word Sense Disambiguation module in NLTK. However, at the moment there is no transparent open-source MT system, which may implement different MT architectures, integrate a variety of potentially useful open-source NLP modules in a flexible way and measure their influence on MT quality against its baseline performance.

Goal of the project

The goal of the proposed project is to create an open-source MT development environment able to integrate independently developed NLP and AI modules and flexibly build alternative system architectures from such modules.

Such environment may function as a common experimental ground, where research groups could test applicability of their technologies to MT and compare the impact of their modules on MT quality with the system’s baseline performance or with the impact of modules developed by other research groups. The environment may provide a test bed for some NLP and AI algorithms which are supposed to be language independent, so their performance on different language types and for different translation directions could be also compared.

Beneficiaries

In the first place the project will be useful for MT community. It will specifically address the need for a common framework for implementing and testing cross-lingual processing modules for MT.

A wider NLP community can also benefit from the proposed MT environment, since monolingual modules that are already developed for open NLP architectures (such as GATE or NLTK) could be integrated and evaluated from the point of view of their usefulness for MT technology.

Monolingual analysis and synthesis modules can annotate features that may become a basis for formal models of human translation. For example, such models can condition application of indirect translation procedures, which are used by human translators, on sets of the identified features. Therefore the MT development environment can help to validate theoretical models of some phenomena found in parallel texts, so it could be useful for researchers in Translation Studies.

The proposed environment will be made practical for teaching courses on MT in the curriculum of Computational Linguistics (following the examples of GATE and NLTK, which are now extensively used for teaching Information Extraction, Machine Learning, Parsing, etc.). For these purposes the environment will use maximally transparent representations and a metalanguage – on the stages of MT transfer, source text analysis and target text generation. The modules that visualise functioning of transfer algorithms will be developed specifically for teaching purposes. We will canvass the needs of other potential users of the proposed environment, e.g., the dynamic visualisations of MT algorithms can be used for teaching courses in Computer Assisted Translation for students in Applied Translation Studies.

Implementation specifications

Core EBMT engine

The core component within the MT development environment will be an MT system which will ensure the baseline functionality for a small number of translation directions, and which later may be updated with independently developed modules integrated into its processing workflow. Example-Based MT architecture is most suitable as the baseline system implementation for the proposed MT development environment. EBMT integrates both data-driven and rule-based techniques (Carl and Way, 2003: xix), it can be built around a reasonably small aligned corpus in a given subject domain and may naturally incorporate statistical techniques as well as linguistic knowledge and be transparent for the developers of the system. EBMT architecture can store examples as annotated linguistic structures (Way, 2003: 444), making use of arbitrary sophisticated linguistic representations, e.g., part-of-speech annotation, lemmatisation, automatically aligned syntactic trees (Groves et al., 2004), semantic representations, e.g., preference semantics formulas (Wilks, 1975), qualia structures (Pustejovsky, 1995), annotation of semantic classes, synsets, etc. On the other hand EBMT architecture does not necessarily require the use of such resources; it may be implemented with a resource-light approach, although availability of additional linguistic resources may substantially boost the MT quality.

Implementation of particular language directions for EBMT depends on availability of parallel texts for a given language pair. The size of the parallel corpus influences the quality of EBMT, but relatively small (preferably word aligned) corpus could be a good starting point (c.f. Lavie et al., 2004: 116). The system should also be focused on one or two reasonably limited subject domains, which nevertheless allow the use of rich and diverse language structures. For the implementation of the core system the following corpora can be used:

- DARPA 1994 MT Evaluation corpus (approx. 35000 words), which is available for English–French and English–Spanish language pairs and consists of 100 French and 100 Spanish news texts translated into English (two versions of human translation are available for each text).
- A parallel corpus of UEFA football match reports available in 7 European and 2 Asian languages (new texts are constantly added to the corpus, so the progress in MT quality as a function of the corpus size could be monitored).

For the development of the core translation modules we will focus on two pairs of languages: English-French and English-Russian in the subject domains of news articles and football match reports. For these languages parallel corpora and open-source morphological resources are available (e.g., part-of speech tagger for French: http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/, for Russian: http://www.aot.ru/).

In both of the suggested subject domains – the news articles and the football match reports – a limited but sufficiently rich language is used, which makes them very suitable for an experimental EBMT system. Moreover texts in these domains have well-understood underlying ontologies and usually convey concrete factual information about some clear-cut events. This allows the developers to use them as an experimental ground for testing the usability of different NLP and AI techniques for MT (such as Word Sense Disambiguation, Anaphora Resolution, Term Extraction, Named Entity Recognition, Information Extraction, automatic paraphrasing, automatic reasoning etc. – which work best in such well-defined domains).

The translation algorithms for the core EBMT system will use relatively knowledge-light annotation of the development corpora: part-of-speech tags, lemmatisation and word alignment, which will be done semi-automatically for the chosen pairs of languages. The core system will implement an algorithm of run-time retrieval of translation examples from sentences with morphological annotation, as described in (Andriamanankasina et al., 2003). Similarly, inductive learning will be used to predict word alignment in new translation examples to be included in the corpus.

In the framework of the proposed project the core EBMT engine will be extended with MT-oriented Information Extraction module which will be based on Sheffield's NE recognisers available in GATE for English (ANNIE) and for Russian (RusIE) – (Popov et al., 2004). Further extension of the core engine beyond the scope of our project will use existing processing resources, such as open-source chunkers and parsers, which are already available for the chosen languages. Users of the system will be able to developed new modules for the existing and for new language directions. The architecture of the suggested MT development environment is shown on the following diagram:

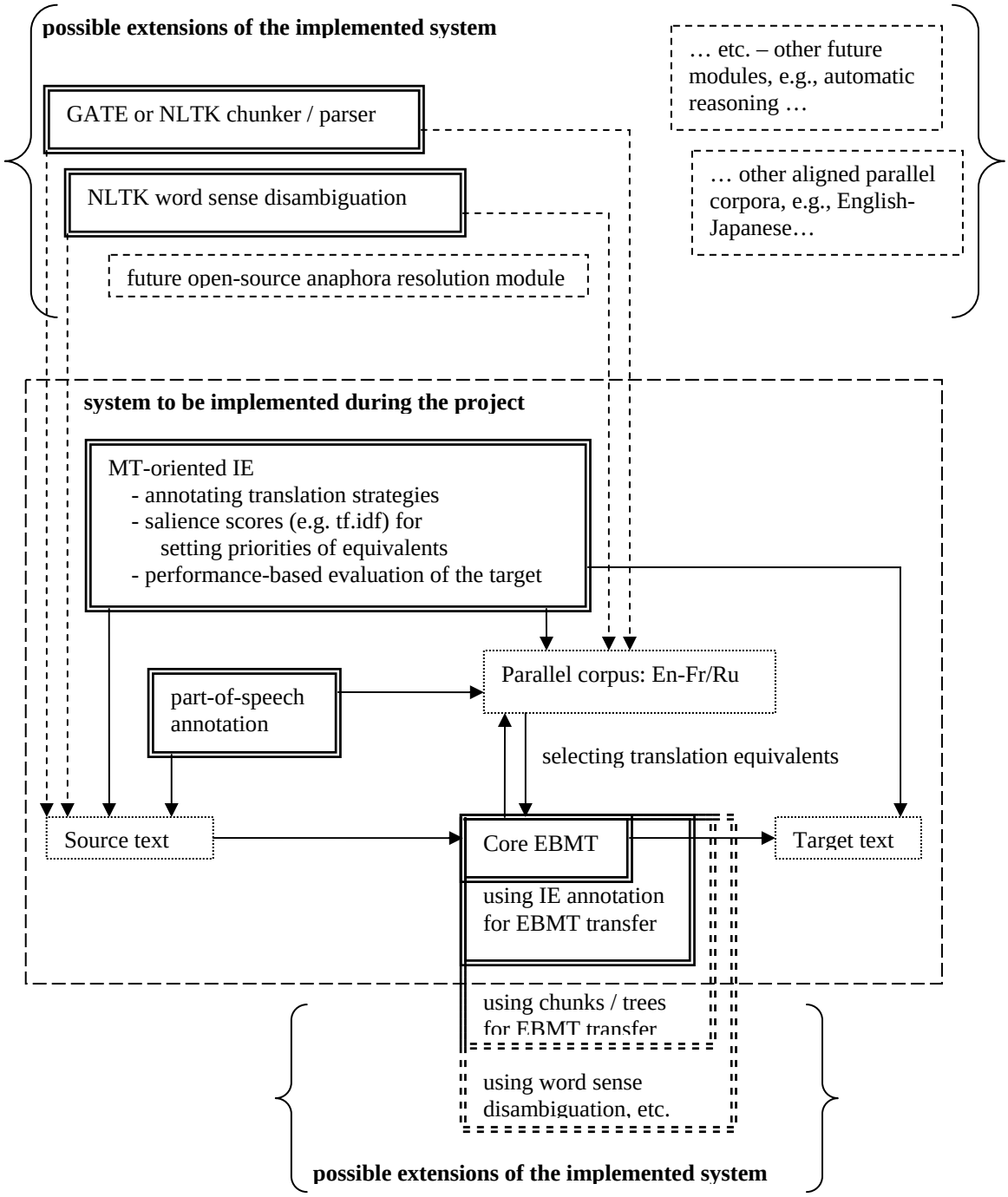

Diagram 1. Architecture of the open source MT development environment

MT-oriented Information Extraction module

MT-oriented Information Extraction (IE) module will be implemented as an example of possible extension of the core EBMT system. IE technology can target specific MT problems such as annotating correct translation strategies for proper nouns using existing Named Entity Recognition (NER) modules. Our previous experiments (Babych, Hartley 2003, 2004a) showed that significant improvement in MT quality can be achieved (for a number of commercial MT systems and language directions, both on morphosyntactic and lexical levels) using freely available open-source NER modules (such as ANNIE / GATE), even though these modules were not originally designed to address MT problems. However, our results suggest that specifically tuning NER modules for MT tasks could be even more beneficial: in this case NER may distinguish different subclasses of Named Entities that require different translation strategies for a given language pair, such as transference ("do-not-translate" or "transliterate"), literal translation, etc. In our experiments NER module interacted with MT systems via "do-not-translate" lists, but proper integration of the modules into MT architecture will give further improvement, since cases of ambiguity between proper vs. common nouns within the same text could be properly addressed in order to avoid potential over-generation for the ambiguous items.

ANNIE and RusIE will be used as a basis for developing MT-oriented NER modules for the proposed system. Name Entity annotation will be refined to explicitly annotate translation strategies for strings of proper nouns. The core MT system will be tuned to use directly the annotation in the text.

IE module may be specifically tuned to annotate translation strategies for other types of lexical items and constructions in the source text (not only the Named Entities). Strings that require oblique translation may be annotated in advance and sent to specific MT modules which will apply appropriate translation transformations.

Other aspects of IE technology may be also useful for MT. Properly defined IE templates may be used as a kind of "Interlingua", which gives proper structure of recognised events and annotates roles of event participants. Such annotation will guide the core MT system in specifying the event structures in the target language, which will allow it to avoid typical mistranslations caused by the wrong event structure, especially when the events in the source and target texts are seen from different perspectives.

IE may provide annotation of relative salience of the lexical items in the source text, identifying lexical items and constructions, which are central for the meaning of the text, using some standard scores, such as tf.idf. This annotation will allow the core MT system to prioritise the lookup of translation equivalents in relation to the salience of terms. Salience of lexical items can be also used in other external modules, e.g., as a feature for Word Sense Disambiguation algorithms.

If the core EBMT system comes up with several translation variants for a source segment, IE modules could be run on these alternatives, doing performance-based evaluation via IE (Babych and Hartley, 2004c), e.g., they may attempt to identify Named Entities or event participants. The preferred candidate would have the closest number of identified items of a particular type to the number identified by IE in the corresponding source segment, or more generally – the structure of IE templates generated on the preferred target would be maximally isomorphic to the structure of templates filled from the source segment.

The suggested MT-oriented IE module will test the assumption that IE technology may provide the necessary flexibility for core MT systems, properly addressing the problem of interaction between language, knowledge and the structure of the subject domain.

Extendibility of the core EBMT system

The proposed open-source MT development environment will specify interfaces between independently developed NLP and AI modules which their developers deem to be useful for MT tasks. The core MT system will be extendible with a variety of modules and will have a

flexible architecture, so the modules could be inserted and removed from the processing pipelines. Obvious extensions of the core EBMT system would be chunking and parsing modules (which are already available for English in GATE 3.0 and NLTK, and may be developed for other languages). Transfer algorithms will also require modification to take advantage of the richer linguistic annotation. Annotation and sub-phrasal alignment of the syntactic structures in the development corpus will bring the core EBMT system into the framework of Data-Oriented Translation.

Interestingly, the proposed environment will allow the developers to extend the system in a principled way, providing a general structure for the extendibility of the system and ensure that transfer algorithms take full advantage of monolingual processing of the source and target texts. Richer linguistic annotations and new modules, such as Anaphora Resolution, Word Sense Disambiguation, etc. will be run on the aligned development corpus and will provide new features which will accommodate translation shifts and transformations done by human translators. E.g., availability of Word Net hierarchies for source and target languages will allow the system to annotate the cases of using hyponyms and hyperonyms as aligned translation equivalents (Shveytser, 1988: 131), to generalise such cases (e.g., with supervised Machine Learning techniques) and to apply dynamically these translation transformations to new sentences. Similarly, an AI module that implements an automated reasoning system in a given subject domain may introduce appropriate annotation for the changes of the point of view on certain events, so the "modulation" translation procedures can be learnt, e.g.: "En.: *it is not difficult to show*" – Fr.: "*il est facile de démontrer*" [lit.: it is easy to show] (Munday, 2001: 57). Annotation of information structure will accommodate the procedures where professional translators change sentence's syntactic perspective in order to convey an appropriate order of presenting given and new information, e.g.: Rus.: *Иную позицию заняли Франция и Германия* [*lit.: "Different*$_{case.acc}$ *position*$_{case.acc}$ *took France*$_{case.nom}$ *and Germany*$_{case.nom}$] – Eng: *A different stand was taken by France and Germany* (An active sentence with the inverse word order was translated by a passive sentence, in order to preserve the information structure of the original) – (Breus, 2003: 23).

Richer monolingual annotations and more sophisticated monolingual processing modules will give deeper insights into human approaches to translation, letting the system to simulate more complex indirect translation strategies.

Evaluation framework for MT engines

It is essential to monitor the progress in MT quality, which is achieved by introducing new NLP / AI modules and by changing the system's architecture. A framework for an automated MT evaluation will be developed for these purposes. The framework will allow the developers to obtain the figures of MT performance obtained by introducing new modules and new system architectures (i.e., new arrangements of modules and processing pipelines) and compare them with the baseline performance of the core EBMT system and / or the performance of the system with previously developed modules. The framework will also allow them to do qualitative error analysis and annotate the cases of improvement and deterioration, caused by introduction of the new modules. Thus it will be possible to make a direct comparison of competing approaches to different NLP problems from the point of view of their usability for MT and will provide a test bed for evaluating viability of different MT architectures for different language pairs given the linguistic resources available for those languages.

The framework will consist of an evaluation corpus (approx. 40 texts, 15000 words). Automated evaluation scores, such as BLEU, NIST, WNM (Papineni et al., 2002; Babych and Hartley, 2004b), will be generated for the baseline EBMT system and for the system which uses an integrated IE module. (The automated scores were found to closely correlate with human intuitive judgements about different aspects of MT quality – adequacy and fluency).

The scores for each evaluated segment, average scores for each text and tools for qualitative analysis of translation differences will be made available as well.

The evaluation framework will be also extendable, so new MT evaluation methods and tools may be implemented and tested in this framework.

Bibliography

Andriamanankasina, T., K.Araki and K.Tochinai. 2003. EBMT of POS-Tagged Sentences via Inductive Learning. In: In: *M.Carl and A.Way (eds.) Recent Advances in Example-Based Machine Translation*. pp. 225-252.

Babych B. and A Hartley. 2003. Improving Machine Translation quality with automatic Named Entity recognition. In: *EACL 2003, 10th Conference of the European Chapter. Proceedings of the 7th International EAMT workshop on MT and other language technology tools. Improving MT through other language technology tools. Resources and tools for building MT*. April 13th 2003, Budapest, Hungary. Pp. 1-8.

Babych, B. and A.Hartley. 2004a. Selecting Translation Strategies in MT using Automatic Named Entity Recognition. In: *Proceedings of the EAMT 2004 Workshop*, Malta, 26-27 April 2004. pp. 18-25.

Babych, B. and A.Hartley. 2004b Extending the BLEU MT Evaluation Method with Frequency Weightings. In: *Proceedings of ACL 2004*: pp. 621-628.

Babych, B. and A.Hartley. 2004c. Comparative Evaluation of Automatic Named Entity Recognition from Machine Translation Output. In: *Proceedings of the IJCNLP Workshop on Named Entity Recognition for Natural Language Processing Applications*. 26 March 2004. Sanya, Hainan Island, China. pp. 41-48.

Breus, E.V. 2002. Osnovy teorii i praktiki perevoda s russkogo jazyka na anglijskij. URAO, Moskow. 208 pp. ("Foundations of the theory practice of translation from Russian into English", in Russian).

Groves, D., M.Hearne and A.Way. 2004. Robust Sub-Sentential Alignment of Phrase-Structure Trees. In.: *Proceedings of COLING* 2004: 1072-1078.

Imamura, K., H.Okuma, T.Watanabe and E.Sumita Example-based Machine Translation Based on Syntactic Transfer with Statistical Models. 2004. In: *Proceedings of COLING 2004*: pp. 99-105.

Kay, M. 2003. Introduction. In: *The Oxford Handbook of Computational Linguistics*. Ed. By R.Mitkov. Oxford University Press. Pp. xvii-xx.

Lavie, A., K.Probst, E.Peterson, S.Vogel, L.Levin, A.Font-Llitjos, J.Carbonell. 2004. A Trainable Transfer-Based Machine Translation Approach for Languages with Limited Resources. In: *Proceedings of the Ninth EAMT Workshop, 26-27 April 2004,* Valetta, Malta: pp. 116-123.

Mitkov, R. 2002. Anaphora Resolution. Pearson Education. 220 pp.

Munday, J. 2001. Introducing Translation Studies: Theories and Applications. Routledge. 222 pp.

Papineni, K., S. Roukos, T. Ward, W.-J. Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. Proceedings of the 40th Annual Meeting of the Association for the Computational Linguistics (ACL), Philadelphia, July 2002, pp. 311-318.

Popov, B., A.Kirilov, D.Maynard, D.Manov. 2004. Creation of reusable components and language resources for Named Entity Recognition in Russian. In: *Proceedings of LREC 2004*: 309-312.

Pustejovsky, J. 1995. The Generative Lexicon. MIT Press, Cambridge, 312 pp.

Shveytser, A.D. 1988. Teorija perevoda: status, problemy, aspekty. Nauka, Moskow. 216 pp. ("Theory of translation: the status, problems, aspects", in Russian).

Vinay, J.P. and J.Darbelnet. 1995. Comparative stylistics of French and English : a methodology for translation / translated and edited by Juan C. Sager, M.-J. Hamel. J. Benjamins Pub., Amsterdam, Philadelphia. 358 pp.

Way, A. 2003. Translating with Examples: the LFG-DOT Models of Translation. In: *M.Carl and A.Way (eds.) Recent Advances in Example-Based Machine Translation*. pp. 443-472.

Wilks, Y.A. 1975. A Preferential Pattern-Seeking Semantics for Natural Language Inference. *Artificial Intelligence* 6: 53-74.