*Natural Language Engineering* **22** (4): 497–500.  
doi:10.1017/S1351324916000103

# *Preface*

After several decades of work on rule-based machine translation (MT) where linguists try to manually encode their knowledge about language, the time around 1990 brought a paradigm change towards automatic systems which try to learn how to translate by looking at large collections of high-quality sample translations as produced by professional translators. The first such attempts were called example- or analogy-based translation,[1] and somewhat later the so-called statistical approach to MT was introduced.[2] Both can be subsumed under the label *data-driven approaches to MT*. It took about 10 years until these self-learning systems became serious competitors of the traditional rule-based systems, and by now some of the most successful MT systems, such as *Google Translate* and *Moses*, are based on the statistical approach.

Their main advantage is that they can use the same core engine for all languages, and that new language pairs can be added by simply training the engine on a new parallel corpus. This is why it has been possible that within a few years statistical machine translation (SMT) systems were able to cover many more language pairs than rule-based systems over decades.

However, what rule-based and statistical systems have in common is that the translation quality which could be achieved in practice is unsatisfactory. For such reasons, in recent years there has been a strong trend towards hybrid systems which try to take the best from both worlds. Actually, this had been suggested for the example-based systems almost from the beginning. However, technology was not far enough to put this into practice at that time, which is why the less ambitious and simpler statistical systems (which neglected long distance dependencies between words) prevailed.

Despite the relative success of data-driven approaches to MT, there are two problems with them, one of them of practical, the other of theoretical nature: The first problem is that current data-driven approaches require large amounts of parallel texts, and that they can deliver optimal results only for the type of data they have been trained on. However, for most language pairs parallel corpora are a scarce resource in general, and if, moreover, we wish them to represent particular domains and genres, it becomes even more difficult to obtain them in large enough quantities.

[1] Satoshi Sato and Makoto Nagao (1990). Toward memory-based translation. *Proceedings of COLING-90*, vol. 3, Helsinki, Finland, 247–252.
Makoto Nagao (1984). A framework of a mechanical translation between Japanese and English by analogy principle. In A. Elithorn and R. Banerji (eds.), *Artificial and Human Intelligence*. Amsterdam: Elsevier Science Publishers.

[2] Peter F. Brown et al. (1990). A statistical approach to machine translation. *Computational Linguistics* 16(2): 79–85.

It should be mentioned, however, that there is some notable progress on this. For example, Google is trying to attract professional translators towards using their *Translator Toolkit*. This toolkit is essentially a translation memory system with integrated automatic pre-translation via Google Translate. The terms of use seem to imply that Google can use all human translations as training materials for Google Translate. That is, if a significant proportion of human translators would use the Google Toolkit, Google would collect exactly those combinations of language pairs, domains and genres where there is demand for high quality human translations. A similar approach is taken by the company dominating the market for translation tools, namely SDL. Whereas Google's toolkit currently only offers basic functionality, SDL combines the market leading translation memory system *SDL Trados Studio* with their statistical and hybrid MT technology which they obtained by acquiring *Language Weaver*.

More established approaches of acquiring parallel corpora involve harvesting the web or asking users of MT systems to correct unsatisfactory translations. All of these approaches have in common that only large companies such as Google, Microsoft, or SDL can fully exploit them, which is likely to lead to an oligopoly.

The second problem with MT using parallel data is that this approach cannot be taken as a model of how humans learn to translate. Obviously, human translators are not trained using parallel texts, and also human second language acquisition does not require them. This observation implies that the respective cognitive processes are based on mechanisms different from current SMT approaches, and that there must be another way of learning how to translate.

The idea underlying this special issue is that comparable corpora could be an alternative to parallel corpora, and that it might be possible to build MT systems which are based on comparable rather than on parallel corpora. In comparison to parallel corpora comparable corpora are a far more abundant resource, and systems based on them are likely to have a better potential to serve as models for the mechanisms underlying translation-related cognitive processing.

Research on comparable corpora has been going on for 20 years, but recently notable progress has been made on using them for MT. To give an overview on the current state of the art, we considered it timely to collect some of the most recent work in one publication. Our call for papers for this special issue received 15 high quality submissions. Due to space constraints we could only accommodate five submissions, which were selected on the basis of a rigorous peer reviewing process.

Let us give a brief overview on the five selected articles: In their paper *End-to-End Statistical Machine Translation with Zero or Small Parallel Texts* Anne Irvine and Chris Callison-Burch describe a statistical system which relies solely on pairs of monolingual texts and without any use of parallel data. It uses a bilingual lexicon induction technique based on a discriminative model which produces higher accuracy translations than previous induction methods. The system can also be used to enhance conventional low resource statistical MT systems.

The title of the second paper is *Extracting Parallel Phrases from Comparable Data for Machine Translation*. Here, the authors Sanjika Hewavitharana and Stephan

Vogel address the problem of data sparseness in standard SMT systems by mining parallel phrase pairs embedded in comparable corpora. For this purpose, they present a novel phrase alignment approach which is designed to only align parallel sections and to bypass non-parallel sections of a sentence. Their method outperforms two alternative algorithms, to which it is compared, and the effectiveness of the system is shown by adding the extracted phrase pairs to an SMT system, whereby an improvement of the BLEU score is achieved.

In the third paper, which is entitled *Exploiting Unbalanced Specialized Comparable Corpora for Bilingual Lexicon Extraction* Emmanuel Morin and Amir Hazem describe a new method for dictionary extraction from comparable corpora. Whereas previous such methods typically assumed that two large corpora of the respective languages are used and that these corpora are typically of similar size, here a context-based projection method is applied which is insensitive to corpus size. It thus can be applied to unbalanced specialized comparable corpora, which is shown to lead to a significant improvement in the quality of the extracted lexicons.

The title of the fourth paper is *Building and Using Multimodal Comparable Corpora for Machine Translation*. Here, the authors Haithem Afli, Loic Barrault, and Holger Schwenk exploit a multimodal comparable corpus, which consists of audio in the source language and text documents in the target language. The audio part is converted to text using a speech recognition system, and then translated to the target language using a standard SMT system. Subsequently, information retrieval techniques are used to extract parallel sentences and phrases. It is shown that adding this data to an English to French translation system leads to significant improvements in translation quality.

The last paper, *Building a Multi-Domain Comparable Corpus Using a Learning to Rank Method* is authored by Razieh Rahimi, Azadeh Shakery, Javid Dadashkarimi, Mozhdeh Ariannezhad, Mostafa Dehghani, and Hossein Nasr Esfahani. Given that the previous papers showed the usefulness of comparable corpora for MT, this work concentrates on how to optimize the construction of comparable corpora. Their approach is that, given two sets of documents in the source and in the target language, the task is to find for each source language document the best matching (i.e. most comparable) target language document. They do this in a sophisticated way by employing a learning-to-rank method. Hereby, each evidence indicating document similarity is considered as a feature, and the feature weights are learned automatically. It is shown that learning the feature weights significantly improves the quality of the document alignments, and the practicality of the method is shown by building a multi-domain English–Persian comparable corpus.

To provide an overview on important other work, these five papers are preceded by an article highlighting some of the recent progress in the field.

In the compilation of this special issue the help of numerous colleagues was indispensable. Let us first acknowledge the excellent cooperation with the JNLE editorial team and the publisher, in particular Natalia Konstantinova and Ruslan Mitkov. Our special thanks also go to the members of our guest editorial board who in a two-step reviewing process kindly volunteered to provide detailed comments on the papers and on the revised versions, thereby supporting us in the selection

of the best papers and providing valuable feedback to the authors for further improvements. These are the board members in alphabetical order:

Akiko Aizawa, Ahmet Aker, Marianna Apidianaki, Nuria Bel, Helena Blancafort, Dhouha Bouamor, Chenhui Chu, Kenneth W. Church, Beatrice Daille, Estelle Delpech, Silvia Hansen-Schirra, Amir Hazem, Diana Inkpen, Kyo Kageura, Kevin Knight, Philipp Koehn, Bo Li, Bin Lu, Belinda Maia, Maria Teresa Martin-Valdivia, Tomas Mikolov, Emmanuel Morin, Santanu Pal, Uwe Quasthoff, Iñaki San Vicente Roncal, Xabier Saralegi, Carolina Scarton, Inguna Skadina, Kamel Smaili, Marko Tadic, George Tambouratzis, Benjamin Tsou, Stephan Vogel, Yorick Wilks, Geoffrey Williams, Krzysztof Wolk, Chengzhi Zhang.

Last but not least, we would like to thank all authors without whom this special issue would not have been possible.

Reinhard Rapp, Serge Sharoff and Pierre Zweigenbaum