An Overview of the SEBAMAT Project

Reinhard Rapp
[1]ILSP / Athena R.C.,
[2]Magdeburg-Stendal University of Applied Sciences, [3]University of Mainz
reinhardrapp@gmx.de

George Tambouratzis
ILSP / Athena R.C.,
6 Artemidos & Epidavrou, Maroussi, 15125, Greece
giorg_t@ilsp.gr

Abstract

SEBAMAT (semantics-based MT) is a Marie Curie project intended to contribute to the state of the art in machine translation (MT). Current MT systems typically take the semantics of a text only in so far into account as they are implicit in the underlying text corpora or dictionaries. Occasionally it has been argued that it may be difficult to advance MT quality to the next level as long as the systems do not make more explicit use of semantic knowledge. SEBAMAT aims to evaluate three approaches incorporating such knowledge into MT.

1 The current state of the art in MT

SEBAMAT aims to show ways on how to improve the translations produced by current MT systems. For several decades the *rule-based approach* was dominant (Arnold et al., 1994) which focused on grammatical well-formedness. In the *statistical approach* (SMT, Brown et al. 1990), linguistic rules were replaced by statistical patterns as automatically extracted from large monolingual and parallel text corpora. Recently, the dominance of SMT has been contested by *neural MT* (NMT), which almost consistently generates better results. Although NMT represents the current state of the art, technical issues and problems have been raised, including NMT's inferior performance to SMT for limited training data, reduced portability across domains, and sensitivity to semantic divergence in the training data (Koehn & Knowles (2017), Carpuat et al. (2017). Sennrich & Zang (2019) improve NMT for small corpora, but marked improvements are gained using larger corpora.

However, despite considerable advances, the quality of current MT systems is still limited, and the likely reason is that the algorithms used are of a mechanical nature, employing statistical rules or deep learning architectures, without a human-like understanding of the texts to be translated. Amongst others, Kevin Knight pointed out that MT systems do not sufficiently take into account semantic considerations such as *who did what to whom, when where and why*. This is also true for NMT, where the translations are typically fluent but often semantically inadequate. SEBAMAT will suggest steps to raise MT quality by taking into account semantics more explicitly than has usually been done so far. Three main directions will be investigated.

2 Explicit semantic disambiguation

Up to now, almost all MT work involving parallel and monolingual corpora has been based on raw texts. However, recently there have been significant advances in word sense induction and disambiguation using corpus-based automatic methods. Inspired by Vintar & Fiser (2016), we suggest to pre-process parallel corpora using word sense disambiguation software, and then apply classical SMT word alignment procedures on the disambiguated rather than the original corpora. That is, word senses rather than words are aligned, and bilingual dictionaries of word senses rather than dictionaries of words are extracted. If the disambiguation can be done with sufficiently high accuracy, this may lead to an improvement in translation quality. The reason is that the average translation ambiguity of a word sense can be expected to be considerably lower than that of a word, which should make the task of finding the correct translation easier.

Of course, other MT systems also (though implicitly) try to select the target words which translate the correct senses of the given source words. However, e.g. in the case of NMT, initially only a single embedding is assigned to each

word or sub-word unit, regardless of sense, and only when building sentence representations senses come into play. This has at least the disadvantage that it is almost impossible to understand what happens internally (black box behavior), so system improvements are typically achieved by trial and error. Also, only the suggested approach delivers a new type of resource, namely a dictionary of sense translations.

3 Semantically annotated corpora

We will also research the promising approach for semantics-based translation proposed by Jones et al. (2013) which uses synchronous hyperedge replacement grammars. For training, this requires semantically annotated corpora, whereby we will focus on language-neutral annotations which can be considered as an interlingua. Our annotation scheme will probably adapt abstract meaning representations (AMR) along the lines of the AMR bank (https://amr.isi.edu/). By default these annotations are done by human experts using Ulf Hermjakob's AMR editor, but as we require large amounts of data we will also investigate how far it is possible to automate this via algorithms for semantic role labeling (Gormley et al., 2014). The next step is to train MT systems on large parallel corpora which are annotated in this way. Synchronous hyperedge replacement grammars can then be used to translate into and from such graph-shaped intermediate meaning representations. Hereby it is possible to extract synchronous grammar rules, and to also combine this with syntactic annotations. For languages with FrameNet[1] style resources, we investigate whether these are suitable for semantics-based translation.

4 Association-based MT

The novel paradigm of association-based machine translation is cognitively motivated and based on the concept of multi-stimulus association. The underlying hypothesis is that the meaning of a short sentence or phrase can be characterized by its associations. That is, the content words in this sentence/phrase are used as multi-word stimuli, and a *meaning vector* is computed. The relation between two sentences can be computed by comparing their corresponding meaning vectors using a vector similarity metric.

To translate a source language phrase, we first compute its meaning vector. Presupposing the existence of a basic dictionary, in analogy to Rapp (1999) we can translate this meaning vector into the target language. Assuming that we already know the meaning vectors of a very large number of target language phrases, we select the target language meaning vector which is most similar to the source language meaning vector. The corresponding target language phrase is considered the translation of the source phrase. This is an unsupervised vector space approach requiring meaning vectors for huge numbers of phrases, but no parallel corpora nor training.

5 Project details

SEBAMAT is a 2-year Marie Curie individual fellowship funded by the European Commission's Horizon 2020 program. It supports the first author to conduct research at Athena R.C. For details see the upcoming project website and https://cordis. europa.eu/project/id/844951.

References

Arnold, D., Balkan, L., Lee Humphreys, R., Meijer, S., Sadler, L. (1994). *Machine Translation. An Introductory Guide.* Manchester: Blackwell.

Brown, P., Cocke, J., Della Pietra, S., Della Pietra, V., Jelinek, F., Lafferty, J., Mercer, R, Rossin, P. (1990). A statistical approach to machine translation. *Computational Linguistics*, 16 (2), 79–85.

Carpuat, M., Vyas, Y., Niu, X., (2017) Detecting cross-lingual semantic divergence for neural machine translation, *Proceedings of the First Workshop on NMT*, Vancouver, Canada, 69–79.

Gormley, M.R., Mitchell, M., Van Dumme, B., Dredze, M. (2014). Low-resource semantic role labeling. Proc. 52nd ACL, Baltimore, MD, 1177–1187.

Jones, B., Andreas, J., Bauer, D., Hermann, K.M., Knight, K. (2013): Semantics-based machine translation with hyperedge replacement grammars. *Proceedings of COLING 2012*, Mumbai, 1359–1376.

Koehn, P., Knowles, R. (2017), Six challenges for neural machine translation, *Proc. of the 1st Workshop on NMT*, Vancouver, Canada, 28–39.

Rapp, R. (1999). Automatic identification of word translations from unrelated English and German corpora. *Proceedings of 37th ACL*, 1999, 519–526.

Sennrich, R., Zhang, B. (2019) Revisiting low-resource neural machine translation: a case study. *Proc. of 57th ACL 2019*, Florence, Italy, 211–221.

Vintar, S., Fiser, D. (2016). Using WordNet-based word senses to improve MT performance. In: Costa-jussa, M. et al. (eds.): *Hybrid Approaches to Machine Translation.* Springer, 191–205.

[1] https://framenet.icsi.berkeley.edu/fndrupal/