Comparison of two different interactive translation prediction approaches integrated in a computer-assisted translation workbench

Daniel Torregrosa dtorregrosa@dlsi.ua.es
Juan Antonio Pérez-Ortiz japerez@dlsi.ua.es
Mikel L. Forcada mlf@dlsi.ua.es
Departament de Llenguatges i Sistemes Informàtics
Universitat d'Alacant, E-03690 Sant Vicent del Raspeig, Spain

Abstract

The objective of interactive translation prediction (ITP), a variant of computer-assisted translation (CAT), is to help professional translators by offering them context-based computer-generated suggestions as they type; the professional translators evaluate then the suggestions, and either accept one of them (sometimes, a prefix can be selected instead of the full proposal) or ignore them all and keep typing the translation in their own way. Most state-of-the-art ITP approaches, such as the one described by Barrachina et al. (2009) or the Thot system[1] (Ortiz-Martínez and Casacuberta, 2014) obtain the suggestions by means of a modified (or tailor-made) statistical machine translations system (SMT) (Koehn, 2004) that is able to provide additional information (such as word alignments, alternative translations, and scores or probabilities for the translation); the suggestions range from a few words to complete translations of the sentence constrained by the already typed target-language prefix. These ITP systems are able to leverage more information from the bilingual resource than if it was used unmodified as a black-box, but doing so they inherit the common requirements of SMT, namely, the dependency on the availability of extensive parallel corpora. It is worth noting that integrating other resources of bilingual information would be almost impossible in this kind of systems, as the ITP tool needs the additional information obtained from the underlying SMT engine.

The authors of this presentation have already proposed an ITP approach (Pérez-Ortiz et al., 2014) that, unlike those previously described, is able to use any bilingual resource capable of delivering one or more translations into target language, regardless of how they are obtained and without the need to modify the resource; suggestions are created by generating all possible sub-segments of words in the source-language sentence (up to a given length) and then querying the available bilingual resources for their translations. The nature of these bilingual resources is not limited to MT systems, but they may also include translation memories, dictionaries, catalogues of bilingual phrases, or any combination of them. A neural-based machine learning algorithm trained on features extracted from the source sentence, on the current prefix of the target sentence, and on the translated sub-segments is used to rank and select which suggestions to show at each time step.[2] Not having to rely on the inner workings of each system allows us to

[1] Thot can be used both as a statistical machine translation (SMT) system and as an ITP tool. We will refer to each system as *SMT Thot* and *ITP Thot*, respectively.

[2] A paper describing this neural system will be presented at the main AMTA 2016 conference. This presentation, however, will focus on an original comparison between our approach and Thot.

seamlessly integrate new resources without modifying how the ITP system works.

In this presentation, we will compare for the first time our proposal with the state-of-the-art system Thot (Ortiz-Martínez and Casacuberta, 2014). *ITP Thot* obtains the proposals as complete translations of the sentences constrained by the typed prefix, and allows the user to accept a prefix or the whole of them. In order to allow a fair comparison of the systems, we will use *SMT Thot* as a black-box bilingual resource for our approach. We will present the results of a comparison between our approach and Thot, both with an extensive automatic evaluation and with a preliminary qualitative and quantitative human evaluation. For the preliminary human evaluation, we will evaluate the usage of both ITP approaches integrated into the widely used open-source professional CAT tool OmegaT.[3] OmegaT has a robust plug-in mechanism that allows us to integrate both approaches as part of the native interface already known by many professional translators; another plug-in[4] is used to measure and log user performance transparently.

References

Barrachina, S., Bender, O., Casacuberta, F., Civera, J., Cubel, E., Khadivi, S., Lagarda, A., Ney, H., Tomás, J., Vidal, E., and Vilar, J.-M. (2009). Statistical approaches to computer-assisted translation. *Computational Linguistics*, 35(1):3–28.

Koehn, P. (2004). Statistical significance tests for machine translation evaluation. In *Conference on Empirical Methods on Natural Language Processing*, pages 388–395.

Ortiz-Martínez, D. and Casacuberta, F. (2014). The new thot toolkit for fully automatic and interactive statistical machine translation. In *Proc. of the European Association for Computational Linguistics (EACL): System Demonstrations*, pages 45–48, Gothenburg, Sweden.

Pérez-Ortiz, J. A., Torregrosa, D., and Forcada, M. L. (2014). Black-box integration of heterogeneous bilingual resources into an interactive translation system. *EACL 2014*, page 57.

[3] `http://www.omegat.org/`

[4] `https://github.com/mespla/OmegaT-SessionLog`