Semantic role labeling for neural machine translation

Reinhard Rapp
Athena R.C.; Hochschule Magdeburg-Stendal; Universität Mainz
reinhardrapp@gmx.de

Neural machine translation (NMT) systems are usually trained with parallel corpora of plain text, and the systems discover all information required for translation by themselves. In contrast, the idea pursued here is to annotate the source language side of a parallel training corpus with semantic roles, thus providing explicit semantic information to the NMT system. The hope is that this may lead to improvements in translation quality as a specialized system for semantic role labeling might do better in taking semantics into account in comparison to what NMT is doing implicitly.

We use the state-of-the-art neural system for semantic role labeling (SRL) provided by the Allen Institute for Artificial Intelligence to annotate the English (en) parts of the French (fr), German (de), Greek (el), and Spanish (es) sections of the Europarl corpus. To give an example, the sentence "John gives the flowers to Mary" is annotated as "[ARG0: John] [V: gives] [ARG1: the flowers] [ARG2: to Mary]". Subsequently, four NMT systems based on the Marian NMT toolkit and using Google's transformer architecture (for details see Rapp, 2021) were trained using the semantically annotated English Europarl portions on the source language side and the unannotated German, Greek, French, and Spanish translations on the target language side. We evaluated the translation quality by applying the BLEU metric on a test set of 2000 randomly held out sentence pairs and obtained the following BLEU scores: en→de: 31.6; en→el: 37.0; en→es: 43.9; en→fr: 39.0.

To have a baseline for comparison, beforehand we had trained systems for the same language pairs using unannotated portions of the Europarl corpus not only on the target but also on the source language side. This led to the following BLEU scores: en→de: 30.2; en→el: 36.5; en→es: 43.3; en→fr: 38.8. Although the improvements are only small, for all four language pairs the evaluation scores when training with SRL-annotated source language text were better than when using unannotated text.

References: Rapp, Reinhard (2021). Similar language translation for Catalan, Portuguese and Spanish using Marian NMT. *Proceedings of the 6th Conference on Machine Translation*.