Using unsupervised corpus-based methods to build rule-based machine translation systems

Felipe Sánchez-Martínez

PhD Thesis, May 2008

Corpus-based approaches to machine translation (MT), such as statistical MT or example-based MT, require large amounts of parallel texts to achieve a reasonable translation quality in open-domain tasks. Such amounts of parallel corpora are not always available, specially for less-resourced language pairs demanding MT services such as Occitan–Catalan, French–Catalan or English–Afrikaans. In these cases, the rule-based paradigm is the only realistic approach to tackle the MT problem. However, building rule-based MT (RBMT) systems entails a huge human effort to code all the linguistic resources needed; mainly, monolingual dictionaries, part-of-speech taggers, bilingual dictionaries, and structural transfer rules.

This dissertation focuses on using corpus-based methods to learn in an unsupervised way some of the linguistic resources required to build shallow-transfer RBMT systems. More precisely, this dissertation focuses on: (i) an unsupervised method to train hidden Markov model (HMM)-based part-of-speech taggers to be used in RBMT; (ii) the automatic inference of the set of states to be used by these HMM-based part-of-speech taggers; and, (iii) the automatic inference of shallow-transfer rules from a small amount of parallel corpora. The final goal is to reduce as much as possible the human effort needed to build a shallow-transfer RBMT system from scratch.

The first approach discussed in this dissertation shows that a statistical model of the target language can be easily used to produce, in an unsupervised way, HMM-based part-of-speech taggers specially suited for their use in RBMT. This novel approach uses information not only from the source language, as general-purpose methods do, but also from the target language and from the translation engine in which the part-of-speech tagger is to be embedded. No parallel, or comparable, corpora are required.

This new, MT-oriented training method was tested on three different language pairs —Spanish–Catalan, French–Catalan and Occitan–Catalan— and the resulting part-of-speech taggers are shown to be better suited for

RBMT, as evaluated against human-corrected translations, than those trained following the classic (unsupervised) Baum-Welch expectation-maximization algorithm. In addition, for Spanish–Catalan, the part-of-speech tagger was also compared to a Spanish HMM-based part-of-speech tagger trained in a supervised way from hand-tagged corpora. The results show that, whereas the tagger trained in a supervised way performs better in terms of part-of-speech tagging accuracy, the translation quality achieved when both taggers are embedded in an RBMT system, is almost indistinguishable.

The second approach focuses on the automatic inference of the number of states to be used by HMM-based part-of-speech taggers. The method consists of applying a clustering algorithm over the states of an initial HMM to reduce the number of states and, consequently, the number of parameters to estimate. The initial HMM has one state per fine-grained part-of-speech tag —consisting of lexical category and detailed morphological inflection information such as genre, number, mood and verb tense— and is trained via the MT-oriented method used in the first approach in the hope that, in this way, the clustering algorithm will infer a set of states better suited to the purpose of translation. This second approach was tested on the same three language pairs. The results vary depending on the language pair, whereas for Occitan–Catalan and French–Catalan translation quality is slightly worse than that of the initial HMM, in the case of Spanish–Catalan translation quality is not affected. In all cases the number of states is drastically reduced, this indicates that the size of the tagset has little impact on the translation quality achieved, and, therefore, the clustering approach described may still be useful when a small system footprint is required.

Finally, this dissertation describes a method for the automatic inference of structural transfer rules to be used in a shallow-transfer RBMT system from small parallel corpora. The structural transfer rules are based on alignment templates, like those used in statistical MT. Alignment templates are extracted from sentence-aligned parallel corpora and extended with a set of restrictions which are derived from the existing bilingual dictionary of the MT system and control their application as transfer rules. The experiments conducted using three different language pairs —Spanish–Catalan, Spanish–Galician and Spanish–Portuguese— show that translation quality is improved as compared to word-for-word translation (when no transfer rules are used), and is close to that obtained using hand-coded transfer rules.

All the software implementing the methods described in this dissertation has been released under free/open-source licenses to ease the reproducibility of these results and to allow other researchers to use and improve them.