Using Synonyms for Arabic-to-English Example-Based Translation

Kfir Bar and Nachum Dershowitz
School of Computer Science, Tel Aviv University, Ramat Aviv, Israel
{kfirbar,nachumd}@post.tau.ac.il

We have developed an experimental Arabic-to-English example-based machine translation (EBMT) system, which exploits a bilingual corpus to find examples that match fragments of the input source-language text Modern Standard Arabic (MSA), in our case—and imitates its translations. Translation examples were extracted from a collection of parallel, sentence-aligned, unvocalized Arabic-English documents, taken from several corpora published by the Linguistic Data Consortium. The system is non-structural: translation examples are stored as textual strings, with some additional inferred linguistic features.

In working with a highly inflected language, finding an exact match for an input phrase with reasonable precision presumably requires a very large parallel corpus. Since we are interesting in studying the use of relatively small corpora for translation, matching phrases to the corpus is done on a spectrum of linguistic levels, so that not only exact phrases are discovered but also related ones. In this work, we looked in particular at the effect of matching synonymous words.

To explore the possibility of matching fragments based on source-language synonyms, we created a thesaurus for Arabic, organized into levels of perceived synonymy. Since an Arabic WordNet is still under development, we developed an automatic technique for creating a rough thesaurus, based on English glosses provided with the Arabic stem list of the Buckwalter morphological analyzer. To create a thesaurus of nouns, we looked at the English WordNet synsets of every English translation of a stem in the Buckwalter list. A synset containing two or more of the translations is taken to be a possible sense for the given stem. This assumption is based on the idea that if a stem has two or more different translations that semantically intersect, it should likely be interpreted as their common meaning. We also considered WordNets hyponym-hypernym relations between the translations senses, and take a stem to have the sense of the shared hyponym. Different strengths of synonymy were defined according to the closeness and uniqueness of these relations. The quality of the systems resultant translations were measured for each of the different levels of synonymy.

In the matching step, the system uses various levels of morphological information to broaden the quantity of matched translation examples and to generate new translations based on morphologically similar fragments. All the Arabic translation examples were morphologically analyzed using the Buckwalter morphological analyzer, and then part-of-speech tagged using AMIRA, in such a way that, for each word, we consider only the relevant morphological analyses with the corresponding part-of-speech tag. For each Arabic word in the translation example, we look up its English equivalents in a lexicon created from the Buckwalter glossaries, and also expand those English words with synonyms. Then we search the English version of the translation example for all instances of these words at the lemma level, creating an alignment table containing one-to-one alignment entries. In addition, several special alignment cases are handled. For instance, an English noun-phrase that contains unaligned words is usually combined with its aligned words, if any, creating a one-to-many entry in the alignment table. In this way, most of the prepositions, definite articles and indefinite articles are covered. Another special case is connecting the immediate noun of an aligned verb to its equivalent.

Demarcating noun-phrase boundaries and obtaining part-of-speech information for the English part is accomplished using Brills part-of-speech tagger and the BaseNP chunker, respectively.

The Arabic version of the corpus was indexed on the word, stem and lemma levels (stem and lemma, as defined by the Buckwalter analyzer). So, for each given Arabic word, we are able to retrieve all translation examples that contain that word on any of those three levels.

In using synonyms for matching, we also considered the relevance of the subject matter of translation examples to any given input sentence. Topics were determined using a classifier that was first trained on the English Reuters training corpus and then used for classifying the English part of the translation examples in our parallel corpus. With this classification of the samples in hand, we trained an Arabic-language classifier on the Arabic version of the parallel corpus, which was then used to classify new Arabic input documents.

During the transfer step, matched fragments are translated using the English version of the parallel corpus. Currently, the system translates each fragment separately and then concatenates those translations to form an output target-language sentence, preferring longer translated fragments, since the individual words appear in a larger context. Recombining those translations into a final, coherent form is left for future work.

We found that synonyms benefit from being matched carefully by considering the context in which they appear. Comparing other ways of using context to properly match the true senses of ambiguous synonyms is definitely a direction for future investigation.

Another interesting observation is the fact that using synonyms on a large corpus did not result in any improvement of the final results, as it did for the smaller corpus. This suggests that synonyms can contribute to EBMT for resource-poor languages other than Arabic, by enabling the system to better exploit the small number of examples in the given corpus.