# Using Verb Paraphrases 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 investigate particularly the effect of matching synonymous verbs.

To explore the possibility of matching fragments based on source-language synonyms, we created a thesaurus for Arabic verbs, based on Arabic comparable corpus. A comparable corpus contains groups of articles dealing with the same topic, which were extracted from the Arabic Gigaword. In order to find synonymous verbs within the given corpora, we began with a small list of synonymous verbs, which was extracted using the English glosses provided with the Arabic stem list of the Buckwalter morphological analyzer. To create this list, we looked at the English WordNet synsets of every English translation of a verb 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 WordNet's hyponym-hypernym relations between the translation senses, and take a stem to have the sense of the shared hyponym.

For extracting additional synonymous verbs, we marked every two matched documents in the comparable corpus the synonymous verbs based on the initial list and started to look for those contexts in which synonymous verbs exist. A context is defined as a list of features extracted from the local environment of the verb and mainly contains morpho-syntactic information. Following the inspiring work by R. Barzilay and K. McKeon for finding paraphrases in comparable corpus for English, we identify the best contexts using the strength and frequency of each context, where the strength of a positive context is defined as $p/N$ and the strength of a negative context is defined as $n/N$, where $p$ is the number of times the context appears in a positive example (synonymous verbs), $n$ is the number of times it appears in a negative example, and $N$ is simply the frequency of the context in the entire corpus. We then select the most frequent $k$ contexts ($k$, a parameter) that their strength is higher than a predefined threshold and use them for extracting additional synonymous verbs.

This is done by finding all the instances of each selected context in every document that was identified as dealing with the same topic. The verbs that are surrounded by those contexts are deemed synonymous.

The effect of the extracted synonyms is examined by our translation system. 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. Each translation example was aligned on the word level, using the Giza++ system (Och and Ney, 2003), which is an implementation of the IBM word alignment models (Brown et al., 1993). Although we did not provide the Giza++ algorithm with a word-based dictionary file, for each un-aligned Arabic word in the translation example, we look up its English equivalents in a lexicon, created using the Buckwalter glossaries, and then expand those English words with synonyms from the English WordNet. Then we search the English version of the translation example for all instances of these words at the lemma level, augmenting the *alignment table* with additional one-to-one entries.

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. 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. Quantitative results will be available before the meeting.

Comparing other ways of using context to properly match the true senses of ambiguous synonyms is a promising direction for future investigation.

We also plan to broaden the level of similarity and use longer paraphrases in the matching step.

## References


Brown, Peter F., Stephen A. Della-Pietra, Vincent J. Della-Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, Vol. 19, 1:263–311.

Buckwalter, Tim. 2002. Buckwalter Arabic Morphological Analyzer Version 1.0. *Linguistic Data Consortium*, Philadelphia, PA.

Diab, Mona, Kadri Hacioglu and Daniel Jurafsky. 2004. Automatic Tagging of Arabic Text: From Raw Text to Base Phrase Chunks. *The National Science Foundation*, Washington, DC.

Och, Franz Josef and Ney, Hermann. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, Vol. 29, 1:19-51.

Regina Barzilay and Kathleen McKeown. 2001. Extracting paraphrases from a parallel corpus. In: *Proceedings of ACL.*