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

Kfir Bar
School of Computer Science
Tel Aviv University
Ramat Aviv, Israel
kfirbar@post.tau.ac.il

Nachum Dershowitz
School of Computer Science
Tel Aviv University
Ramat Aviv, Israel
nachumd@tau.ac.il

Abstract

An implementation of a non-structural Example-Based Machine Translation system that translates sentences from Arabic to English, using a parallel corpus aligned at the sentence level, is described. Source-language synonyms were derived automatically and used to help locate potential translation examples for fragments of a given input sentence. The smaller the parallel corpus, the greater the contribution provided by synonyms. Considering the degree of relevance of the subject matter of a potential match contributes to the quality of the final results.

1 Introduction

Ever since it was proposed by Nagao (1984), the example-based (or "memory-based") paradigm has been a fairly common method in natural language processing (NLP), especially for machine-translation applications. The main idea behind example-based machine translation (EBMT) is to translate fragments of the source-language input text based on similar translations found in a corpus of translated texts. Such a process presumably emulates the way a human translates in some cases. Since translations are based on actual manually-created samples, the results are usually more fluent than ones created artificially using other translation paradigms.

We have developed an Arabic-to-English example-based translation 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. 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. In addition, we examined the possibility of matching fragments based on source-language synonyms. For this purpose, we automatically extracted a thesaurus for Arabic, using the stem list provided by the Buckwalter (version 1.0) morphological analyzer (Buckwalter, 2002), and organized it into levels of perceived synonymy. The quality of the system's resultant translations were measured for each of the different levels.

In using synonyms for matching, we also considered the relevance of the subject matter of translation examples to the 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. In the recombination step of an example-based translation system, all the translated fragments are pasted together to form a complete target-language text, usually by preferring longer translated fragments, since the individual words appear in a larger context.

Like many other Semitic languages, Arabic is highly inflected; words are derived from a *root* and *pattern* (the stem), combined with prefixes, suffixes and circumfixes. The root consists of 3 or 4 consonants and the pattern is a sequence of consonants and variables for root letters. Using the same root with different patterns may yield words with different meanings. For instance, the combination of the root ك.ت.ب (*k.t.b*) and the pattern mXXX (here, X is a variable) results in the word مكتب (*mktb*, "office"). Combining the same root with the pattern XXAX, results in the word كتاب (*ktAb*, "book").

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.

The system described here is non-structural: it stores translation examples as textual strings, with some additional linguistic features. Currently, the system translates each fragment separately and then concatenates those translations to form an output target-language sentence. Recombining those translations into a final, coherent form is left for future work.

The following section gives a short description of some previous work. Section 3 contains a general description of our system. In Section 4, we provide some experimental results using common automatic evaluation metrics. Some conclusions are suggested in the last section.

2 Related Work

The initiator of the example-based approach applied to machine-translation is Nagao (1984), who investigated a structural Japanese-to-English example-based system. Other influential works include (Sato and Nagao, 1990; Maruyama and Watanabe, 1992; Sumita and Iida, 1995; Nirenburg et al., 1994; Brown, 1999).

Several works deal with morphologically rich languages such as Arabic. Nevertheless, we could not find any specific work that measures the effect of using synonyms in the matching step. Among relevant works there is (Stroppa et al., 2006), an example-based Basque-to-English translation system. That system focuses on extracting translation examples using the marker-based approach integrated with phrase-based statistical machine translation to translate new given inputs. As reported, that combined approach showed significant improvements over state-of-the-art phrase-based statistical translation systems.

The work by Lee (2004) is on improving a statistical Arabic-to-English translation system based on words as well as on phrases by making the parallel corpus syntactically and morphologically symmetric in a preprocessing stage. This is achieved by segmenting each Arabic word into smaller particles (prefix, stem and suffix), and then omitting some of them in order to make the parallel corpus as symmetric as possible. That method seems to increase evaluation metrics when using a small corpus. Similar conclusions were reached by Sadat and Habash (2006) in their work on improving a statistical Arabic-to-English translation system. In that research, several morphological preprocessing schemes were applied separately on different sizes of corpora.

In work on Japanese-to-English example-based machine translation (Nakazawa el al., 2006), synonyms were used in the source language for matching translation examples, similar to the idea presented in this paper. However, the effect of this idea on the final results was not measured.

There are also several works that use synonyms in the target language for improving example alignments. A well-known work of this nature is (Brown, 1996).

In recent work (Philips et al., 2007), an Arabic-to-English example-based system is presented. Similar to our work, they broaden the way the system performs matching. That system matches words based on their morphological information, so as to obtain more relevant chunks that could not otherwise be found, and showed some improvement over state-of-the-art example-based Arabic-to-English translation systems. This matching approach also resulted in additional irrelevant matched fragments, which had to be removed in later stages.

There are a number of works on automatic thesaurus creation. Some of them use parallel corpora for finding semantically-related source-language words based on their translations. One interesting work is (Dyvik, 2006), which uses an English-Norwegian parallel corpus for building a lattice of semantically-related English and Norwegian words. It then discovers relations like synonyms and hyponyms. Another related work (van der Plas and Tiedemann, 2006) uses a multilingual sentence-aligned parallel corpus for extraction of synonyms, antonyms and hyponyms for Dutch.

Our own work focuses on matching translation examples using various levels of morphological information plus synonyms, keeping the number of matched fragments for the transfer step as low as possible. We also measure the effect of considering the topic of the translation examples and the input sentence by allowing the system to match on the synonym level only if the candidate translation example and the input sentence are on the same topic.

3 System Description

3.1 Translation Corpus

The translation examples in our system were extracted from a collection of parallel, sentence-aligned, unvocalized Arabic-English documents, taken from a news-related corpus published by the Linguistic Data Consortium (LDC2004T18). All the Arabic translation examples were morphologically analyzed using the Buckwalter morphological analyzer, and then part-of-speech tagged using AMIRA (Diab et al., 2004) 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++ (Och and Ney, 2003) system, 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 unaligned 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 (Miller, 1995). 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.

3.2 Matching

Given a new input sentence, the system begins by searching the corpus for translation examples for which the Arabic version matches fragments of the input sentence. In the implementation we are describing, the system is restricted to fragmenting the input sentence so that a matched fragment must be a combination of one or more complete adjacent base-phrases of the input sentence. The base-phrases are initially extracted using the AMIRA tool.

The same fragment can be found in more than one translation example. Therefore, a ***match-score*** is assigned to each fragment-translation pair, signifying the quality of the matched fragment in the specific translation example.

Fragments are matched word by word, so the score for a fragment is the average of the individual word match-scores. To deal with data sparseness, we generalize the relatively small corpus by matching words on text, stem, lemma, morphological, cardinal, proper-noun, and synonym levels, with each level assigned a different score. These match-levels are defined as follows:

Text level means an exact match. It credits the words in the match with the maximum possible score.

Stem level is a match of word stems. For instance, the words الدستورية (***Aldstwryp***, "the constitutionality") and دستوريتي (***dstwryty***, "my constitutional") share the stem دستوري (***dusotuwriy***). This match-level currently credits words with somewhat less than a text-level match only because we do not have a component that can modify the translation appropriately.

Lemma level matches are words that share a lemma. For instance, the following words match in their lemmas, but not stems: مارق (***mAriq***, "apostate"); مراق (***mur~Aq***, "apostates"). The lemma of a word is found using the Buckwalter analyzer. For the same reasons as stem-level matches, an imperfect match score is assigned in this case. When dealing with unvocalized text, there are, of course, complicated situations when both words have the same unvocalized stem but different lemmas, for example, the words كتب (***katab***, "wrote") and كتب (***kutub***, "books"). Such cases are not yet handled accurately, since we are not working with a context-sensitive Arabic lemmatizer, and so cannot unambiguously determine the correct lemma of an

Arabic word. Actually, by "lemma match", we mean that words match on any one of their possible lemmas. Still, the combination of the Buckwalter morphological analyzer and the AMIRA part-of-speech tagger allows us to reduce the number of possible lemmas for every Arabic word, so as to reduce the amount of ambiguity. Further investigation, as well as working with a context-sensitive morphology analyzer (Habash and Rambow, 2005), will allow us to better handle all such situations.

Cardinal level matches apply to all numeric words. Correcting the translation of the input word is trivial.

Proper-noun level matches are words that are both tagged as proper nouns by the part-of-speech tagger. In most cases the words are interchangeable and, consequently, the translation can be easily fixed in the transfer step.

Morphological level matches are words that match based only on their morphological features. For example, two nouns that have the definite-article prefix ال (*Al*, "the") at the beginning constitute a morphological match. This is a very weak level, since it basically allows a match of two different words with totally different meanings. In the transfer step, some of the necessary corrections are done, so this level appears, all the same, to be useful when using a large number of translation examples.

Synonym level matches, the additional feature investigated in the current work, are words that are deemed to be synonyms, according to our automatically extracted thesaurus. Since synonyms are considered interchangeable in many cases, this level credits the words with 0.95, which is almost the maximum possible. Using a score of 1.0 reduces translation results because sometime synonym based fragments hide other text based fragments, and the latter are usually more accurate.

At this point in our experiments, we are using ad-hoc match-level scores, with the goal of a qualitative evaluation of the effect of including the synonym level for matching. Exact-text matches and cardinal matches receive full weight (100%); synonyms, just a tad bit less, namely 95%; stems and proper nouns, 90%; lemmas and stems are scored at 80%; morphological matches receive only 40%.

Fragments are stored in a structure comprising the following: (1) source pattern – the fragment's Arabic text, taken from the input sentence; (2) example pattern – the fragment's Arabic text, taken from the matched translation example; (3) example – the English translation of the example pattern; (4) match score – the score computed for the fragment and its example translation. Fragments with a score below some predefined threshold are discarded, since passing low-score fragments to the next step would dramatically increase the total running time and sometimes make it unfeasible to process all fragments.

3.3 Thesaurus Creation

Since Arabic WordNet is still under development, we have developed an automatic technique for creating a thesaurus, using the Buckwalter gloss information, extended with English WordNet relations.

Currently, the thesaurus we built contains only nouns. Synonyms for other word types, such as verbs, are planned. Dealing with verbs seems to be more difficult than nouns, since the meaning of an Arabic verb usually changes when used with a different preposition.

Every noun stem in the Buckwalter list was compared to all the other stems when looking for synonym relations. Each Buckwalter stem entry provides one or more translations. Sharing an English translation, however, is insufficient for determining that two stems are synonymous, because of polysemy; we do not know which of a translation's possible senses was intended for any particular stem. Therefore, we need to attempt to determine stem senses automatically. We ask the English WordNet for all (noun) synsets (sets of synonyms) of every English translation of a stem. A synset containing two or more of the Buckwalter 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 probably be interpreted as their common meaning. We also consider the hyponym-hypernym relation between the translations' senses and understand a stem to have the sense of the shared hyponym in this case.

Based on the above information, we define five levels of synonymy for Arabic stems: Level 1 – two stems have more than one translation in common. Level 2 – two stems have more than one sense in common, or they have just one sense in common but this sense is shared by all the translations. Level 3 – each stem has one and the same translation. Level 4 – each stem has exactly one translation and the two translations are English synonyms. Level 5 – the stems have one translation in common. Every stem pair is assigned the highest possible level of synonymy, or none when none of the above levels applies. The resultant thesaurus contains 22,621 nouns, 20,512 level-1 relations, 1479 relations on level 2, 17,166 on level 3, 38,754 on level 4, and 137,240 on level 5.

The quality of the translation system was tested for each level of synonymy, individually, starting with level 1, then adding level 2 and so forth. Figure 1 shows an example of a relation between two Arabic stems. In this example, the stem اعادة (*AEAdp*, “return”) is matched to the stem كرور (*krwr*, “return”) on level 2 because the first stem is translated as both “repetition” and “return”, which share the same synset. The second stem is translated as “return” and “recurrence”, which also share the same synset as the first stem. Therefore level 2 is the highest appropriate one. Table 1 shows some extracted synonyms and their levels.

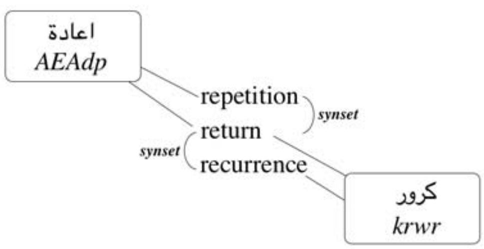

Figure 1. Synonym relation level-2 example

Synonyms	Level
n$yj / *dmE* (“crying”)	4
sTH / *sqf* / (“ceiling”)	5
zlEwm / *Hlqwm* (“throat”)	1
njdp / *AEAnp* (“help;support”)	2
AbtdA' / *ftH* (“beginning”)	5
AxtrAE / *AbtkAr* (“invention”)	3

Table 1. Examples of extracted synonyms

3.4 Matching Synonyms

The extracted thesaurus was used for matching source-language fragments based on synonyms. Finding a synonym for a given word is not a simple task, considering that input sentences are not given with word senses. Matching input words based on synonymy without knowing their true senses is error-prone, because one might match two synonym words based on a specific sense that is not the one used by the author. One way to handle this issue would be to use a word-sense-disambiguation tool for Arabic to uncover the intended sense of each input sentence word. Although there has been some research in this area, we could not find any available tool that produces reasonable results. Even were we to find one, it would probably use English WordNet senses, since Arabic WordNet is not ready yet.

Another option for matching synonyms is to use the immediate context of a candidate word for matching. Given a pair of words, a window of several words appearing around each may be compared on several WordNet levels and a final score can be computed on that basis. Candidate pairs crossing a predefined threshold can be considered as having the same sense. This direction was left for future investigation.

In this work, we decided to experiment with a different route. We classify each input sentence by topic, as well as all the corpus translation examples. For each translation example, we consider synonyms only if its topic-set intersects with that of the input sentence. The classification was done using the manually-tagged Reuters-21578 corpus for English, since we could not find a similar corpus for Arabic. First, we trained a simple classifier on the training-set given by Reuters, building statistical model for every topic of the predefined Reuter topic list. We used the support-vector-machine (Joachims, 2002) model for this classification task, it having proved to be one of the most appropriate one for classification for this corpus. Feature-vectors consisted of *tf-idf* values for English stems, extracted from English WordNet by a morphological analyzer, ignoring stems of stop words. The classifier was tested on 1219 documents from the test-set provided by Reuters, producing accurate results in the 94% range in most cases.

In the next step, we used this classifier to classify the English half of all the translation examples in our parallel corpus, allowing for more then one topic per document. In addition, the Arabic part of those translation examples was used as a training-set for training another classifier for the same topic list for Arabic. Like its English equivalent, it uses stems as features, ignores stem of stop words, and creates feature-vectors using the *tf-idf* function. Stems were extracted using the Buckwalter morphological analyzer. The accuracy of this classifier was not measured due to the lack of any manually tagged test-set.

Back to the translation process. Given a new sentence from an input document, the system begins by classifying the entire input document using the Arabic classifier and determining its topic-set, which is assigned to all sentences within that document. Finally, during the matching step, we allow the system to consider synonyms only in the case of a non-empty intersection of topic-sets of the input sentence and the examined translation example. The efficacy of this classification feature was examined and results show a slight improvement in final translations compared to the same conditions running without classification. We elaborate further on this in the results section.

3.5 Transfer

The input to the transfer step consists of all the collected fragments found in the matching step, and the output is a set of translations for those fragments. Translating a fragment is done in two main steps: (1) extracting the translation of the example pattern from the English version of the translation example; (2) fixing the extracted translation to form a translation of the corresponding input fragment.

First Step – Translation Extraction

The first step is to extract the translation of a fragment's example pattern from the English version of the translation example. Here we use the prepared alignment table for every translation example within our corpus. For every Arabic word in the pattern, we look up its English equivalents in the table and mark them in the English version of the translation example. Recall that the English equivalent may be composed of more than one token. Next, we extract the shortest English segment that contains the maximum number of corresponding parts. Sometimes a word in an Arabic example pattern has several English equivalents, which makes the translation extraction process complicated and error prone. For this reason, we also restrict the ratio between the number of Arabic words in the example pattern and the number of English words in the extracted translation, bounding them by a function of the ratio between the total number of words in the Arabic and English versions of the translation example.

For example, take the following translation example:

A: الخدمات الاستشارية والتعاون التقني في ميدان حقوق الإنسان

E: "Advisory services and technical cooperation in the field of human rights."

Table 2 is the corresponding alignment table.

English	Arabic	
Services	*AlxdmAt*	الخدمات
Advisory	*AlAst$Aryp*	الاستشارية
Cooperation	*wAltEAwn*	والتعاون
Technical	*Altqny*	التقني
In	*fy*	في
Field	*mydAn*	ميدان
Rights	*Hqwq*	حقوق
Human	*AlAnsAn*	الإنسان

Table 2. Alignment table

Now, suppose the example pattern is ميدان حقوق الإنسان (*mydAn Hqwq Al<nsAn*, "the field of human rights"), and we want to extract its translation from the English version of the example. Using the extracted look-up, we mark the English equivalents of the pattern words in the translation example, "Advisory services and technical cooperation in the field of human rights", and then we extract the shortest English segment that contains the maximum number of corresponding words, viz. "field of human rights".

This is, of course, a simple instance. More complicated ones would have more than one equivalent per Arabic word.

Second Step – Fixing the Translation

Recall that the match of a corpus fragment to the input fragment can be inexact, since words may be matched at several levels. Exactly matched words or synonyms may be assumed to possess the same translation, whereas stem- or lemma-matched words may require modifications of the extracted translation (mostly inflection and preposition issues). These "massaging" issues are left for a future enhancement.

Words matched on the morphological level, however, require a complete change of meaning. For example, take the input fragment مجلس الامن

(*mjls AlAmn*, "the Security Council") matched to the fragment مسؤولية الامن (*ms&wlyp AlAmn*, "the security responsibility") in some translation example. The words مجلس (*mjls*, "council") and مسؤولية (*ms&wlya*, "responsibility") match only on the morphological level (both are nouns). Assume that the extracted translation from the translation example is "the security responsibility", which is actually a translation of مسؤولية الامن (*ms&wlyp AlAmn*), not the translation of the input pattern at all. But, by replacing the word "responsibility" from the translation example with the translation of مجلس (*mjls*, "council") from the lexicon, we get the correct phrase, namely, "the Security Council". Our lexicon is constructed using glossaries extracted from the Buckwalter morphological analyzer and expanded with WordNet synonyms, as explained above.

For each final translated fragment, we calculate a ***translation-score***, which is the ratio between the number of covered words and the total number of words in the Arabic pattern. The ***total-score*** of a fragment is the average of the match-score and the translation-score multiplied by the ratio between the number of input tokens covered by the fragment and the total amount of the input sentence tokens. This formula is the result of several adaptations, based on experiments, and resulted in the best performance.

3.6 Recombination

In the recombination step, we paste together the extracted translations to form a complete translation of the input sentence. This is generally composed of two subtasks. The first is finding the best recombination of the extracted translations that covers the entire input sentence, and the second is smoothing out the recombined translations to make a fully grammatical English sentence. Currently, we handle only the first subtask, which chooses the recombination obtaining the best cover of the given input source-language sentence. This is obtained by preferring long translated fragments to short ones, as well as preferring covers composed of fewer fragments. Finding the best cover is performed in a dynamic-programming fashion. By multiplying the total scores of the comprised fragments, we calculate a final translation-score for each generated recombination.

4 Experimental Results

Experiments were conducted on two corpora. The first contains 29,992 (1,247,468 Arabic words) translation examples and the second one contains 58,115 (1,997,434 Arabic words). The system was tested on all levels of synonyms relations and the effect of using the classification feature on every level was examined.

The following results are based on a test set of 586 sentences from 68 documents (17370 words) taken from the 2009 NIST MT Evaluation data and compared to four reference translations. Despite the fact that our system still does not perform the last, smoothing stage of the translation process, we evaluated results under some of the common automatic criteria for machine-translation evaluation: BLEU (Papineni, 2002) and METEOR (Banerjee and Lavie, 2005). Table 3 shows some experimental results, presented as BLEU and METEOR score.

From these results, one can observe that, in general, the system performs slightly better when using synonyms. The most prominent improvement in the BLEU score was achieved when using all levels, 1 through 5, on the small corpus. However, the same experiments using the large corpus did not show significant improvements. This was expected: the larger corpus has more translation examples that might match more fragments exactly. Using synonyms at level 5 caused reductions in all scores in the large corpora. This is probably because level 5 gives synonyms of low confidence, thereby introducing errors in matching corpus fragments, which may hide better fragments that could participate in the output translation. On the other hand, when using level 5 synonyms on the small corpus, the system performed even better than when not using them. That can be explained by the fact that the small corpus probably produces fewer fragments, and the ones based on synonyms can cover ranges of the input sentence, which were not covered by other fragments. However, when using the classification feature over the large corpus, the system was able to remove some of the problematic fragments, resulting in better scores.

In general, when synonyms are used and contribute significantly, this classification feature did show some improvement. This strengthens our intuition that real synonyms are more likely to be found in documents dealing with similar subject matters. We expect that taking the words' local context into consideration, as mentioned above, would result in even better performance.

In addition to the traditional automatic evaluation for the resulted translations, we have measured the effect of using synonyms on the corpus coverage. Table 4 summarizes the number of uncovered 1-4 grams when using synonyms vs. without using synonyms on the small corpus. The results show that when using synonyms the system was able to find an additional 252 bigrams; however, on longer N-grams the system did not show significant improvement. As expected, increasing the size of the corpus reduced the positive effect on N-gram coverage.

5 Conclusions

The system we are working on has demonstrated the potential for using synonyms in an example-based approach to machine translation, for Arabic, in particular. 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 a significant improvement of the final results, as it did for a smaller corpus. This suggests that synonyms can contribute to EBMT for language pairs lacking large parallel corpora, by enabling the system to better exploit the small number of examples in the given corpus.

More work is still needed for better aligning the translation examples. Sometime, even if the system succeeds in matching examples based on synonyms, the final translation was wrong due to a sparse alignment table for the retrieved translation example. Trying to use a word-based dictionary for Giza++ is one direction, but we intend to also explore other alignment methods.

Of course, smoothing out the output translations is an essential step toward understanding the real potential of our system. This step is currently being investigated and planned for implementation in the near future.

Though the scores achieved by our system remain low, primarily because of the above-mentioned alignment and smoothing issues, a detailed examination of numerous translations suggests that the benefits of using matches based on synonyms will carry over to more complete translation systems. What is true for our automatically-generated thesaurus, is even more likely to hold when a quality Arabic thesaurus will become available for mechanical use. In the meanwhile, we are working of different methods for automatic extraction of thesaurus for Arabic. We have begun to investigate the potential of also using verb synonyms for Arabic. We have already realized that the prepositions used by the verbs should also be taken into account, as they might change sense, when trying to find synonyms. That could be difficult, since we have not found any freely available thesaurus for Arabic containing this information on verbs. Considering semantically-related expressions (paraphrases) in example-base machine translation is another direction we intend to explore.

In general, we believe that the example-based method is an interesting way to find realistic translations for parts of the given input. Small corpora should be better exploited, especially when dealing with languages with few available large parallel corpora.

Test	Small Corpus				Large Corpus			
	w/ classification		w/o classification		w/ classification		w/o classification	
	BLEU	**MTOR**	**BLEU**	**MTOR**	**BLEU**	**MTOR**	**BLEU**	**MTOR**
Level 1	0.1186	0.4748	0.1176	0.4756	0.1515	0.5183	0.1506	0.5185
Levels 1 – 2	0.1176	0.4769	0.1173	0.4748	0.1515	0.5183	0.1505	0.5186
Levels 1 – 3	0.1186	0.4762	0.1176	0.4770	0.1520	0.5186	0.1510	0.5189
Levels 1 – 4			0.1179	0.4756	0.1519	0.5184	0.1509	0.5188
Levels 1 – 5	**0.1192 (+9%)**	0.4746	0.1177	0.4751	0.1500	0.5181	0.1484	0.5170
No synonym			0.1084	0.4460			0.1485	0.5194

Table 3. Experimental results – BLEU and METEOR (MTOR) scores

	w/ synonyms	w/o synonyms
Unigrams	733	738
Bigrams	7612 **(+3.2%)**	7864
Trigrams	11554	11632
4-grams	11224	11243

Table 4. Experimental results – Uncovered N-grams in the small corpus

References

Banerjee, Satanjeev and Alon Lavie. 2005. Meteor: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of the 43rd Annual ACL Meeting. Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization*, Ann Arbor, MI, 65-72.

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.

Brown, Ralf D. 1996. Example-Based Machine Translation in the Pangloss System. In: *Proceedings of the 16th International Conference on Computational Linguistics (COLING-96)*, Copenhagen, Denemark, Vol. 1:22-32.

Brown, Ralf D. 1999. Adding Linguistic Knowledge to a Lexical Example-Based Translation System. In: *Proceedings of TMI*, 22-32.

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.

Dyvik Helge. 2004. Translations as Semantic Mirrors: From Parallel Corpus to WordNet. *Language and Computers*, *Rodopi*, Vol. 49, 1:311-326.

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

Fung, Pascale and Kathleen McKeown. 1994. Aligning Noisy Corpora Across Language Groups: Word Pair Feature Matching by Dynamic Time Warping. In: *Proceedings of the 1st Conference of the Association for Machine Translation in the Americas (AMTA-94)*, Columbia, MD, USA, 81-88.

Joachims, Thorsten. 2002. Learning to Classify Text Using Support Vector Machines. *Kluwer*.

Habash, Nizar and Owen Rambow. 2005. Arabic Tokenization, Morphological Analysis, and Part-of-Speech Tagging in One Fell Swoop. In: *Proceedings of the Conference of American Association for Computational Linguistics*, Ann Arbor, MI, 578-580.

Lee, Young-Suk. 2004. Morphological Analysis for Statistical Machine Translation. In: *Proceedings of HLT-NAACL-04,* Lisbon, Portugal, 57–60.

Lonneke van der Plas and Jörg Tiedemann. 2006. Finding Synonyms Using Automatic Word Alignment and Measures of Distributional Similarity. In: *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, 866-873.

Maruyama, Hiroshi and Hideo Watanabe. 1992. Tree Cover Search Algorithm for Example-Based Translation. In: *Proceedings of TMI*, 173-184.

Miller, George A. 1995. WordNet: A Lexical Database for English. *Communications of the ACM*. Vol. 38, 11:39-41.

Nagao, Makoto. 1984. A Framework of Mechanical Translation between Japanese and English by Analogy Principle. In: A.Elithorn and R.Banerji, eds., *Artificial and Human Intelligence*. North-Holland.

Nakazawa, Toshiaki, Kun Yu, Daisuke Kawahara, and Sadao Kurohashi. 2006. Example-Based Machine Translation Based on Deeper NLP. In: *Proceedings of International Workshop on Spoken Language Translation (IWSLT'06)*, Kyoto, Japan, 64-70.

Nirenburg, Sergei, Stephen Beale and Constantine Domashnev. 1994. A Full-Text Experiment in Example-Based Machine Translation. In: *International Conference on New Methods in Language Processing (NeMLaP)*, Manchester, UK, 78-87.

Papineni, Kishore, Salim Roukos, Todd Ward and Wei-Jing Zhu. 2002. Bleu: A Method for Automatic Evaluation of Machine Translation. In: *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL)*, Philadelphia, PA, 311-318.

Phillips, Aaron B., Cavalli-Sforza Violetta and Ralf D. Brown. 2007. Improving Example-Based Machine Translation through Morphological Generalization and Adaptation. In: *Proceedings of Machine Translation Summit XI*, Copenhagen, Denmark, 369-375.

Sadat, Fatiha and Nizar Habash. 2006. Arabic Preprocessing Schemes for Statistical Machine Translation. In: *Proceedings of Human Language Technology Conference of the NAACL*.

Sato, Satoshi and Makoto Nagao. 1990. Toward Memory-Based Translation. In*: Proceedings of the 13th conference on Computational linguistics*, Helsinki, Finland, Vol. 3:247-252.

Stroppa, Nicolas, Declan Groves, Kepa Sarasola and Andy Way. 2006. Example-Based Machine Translation of the Basque Language. *7th Conference of the Association for Machine Translation in the Americas (AMTA-06)*, 232 – 241.

Sumita, Eiichiro and Hitoshi Iida. 1995. Heterogeneous Computing for Example-Based Translation of Spoken Language. In: *Proceedings of TMI*, 273-286.