Using Semantic Equivalents for Arabic-to-English Example-Based Translation

Kfir Bar and Nachum Dershowitz

School of Computer Science, Tel Aviv University, Ramat Aviv, Israel

Abstract

We explore the effect of using Arabic semantic equivalents in an example-based Arabic-English translation system. We describe two experiments using single-word equivalents in translation as test cases for broadening the level of similarity and using multi-word Arabic paraphrases in the future. In the first experiment, we use synonymous Arabic nouns, derived from a lexicon, to help locate potential translation examples for fragments of a given input sentence. Not surprisingly, 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. In the second experiment, we used automatically extracted single-word verb paraphrases, derived from a corpus of comparable documents. The experiments were performed within an implementation of a non-structural example-based translation system, using a parallel corpus aligned at the sentence level. The methods developed here should apply to other morphologically-rich languages.

1. Introduction

Corpus-based translation systems use existing parallel texts to guide the translation process. One of the main problems, when using a corpus-based system for translation, is the relatively small quantity of data that the

system may have available for the purpose. Callison-Burch, in his thesis (Callison-Burch, 2007), measured the effect of the size of the parallel-corpus size used in a statistical Spanish-to-English translation system on the amount of covered *n*-grams in the resultant translations. His results show a clear increase in coverage as the corpus grows larger, but even when the corpus contains relatively many words, many three- and four-word sequences remain uncovered. We repeated his experiment, but using an example-based Arabic-to-English translation system instead. The results, which are displayed in Figure 1, are more or less similar to those obtained for Spanish. The horizontal axis represents the number of Arabic words in the parallel-corpus used by the system and the vertical axis is the number of translated test-set unique *n*-grams.

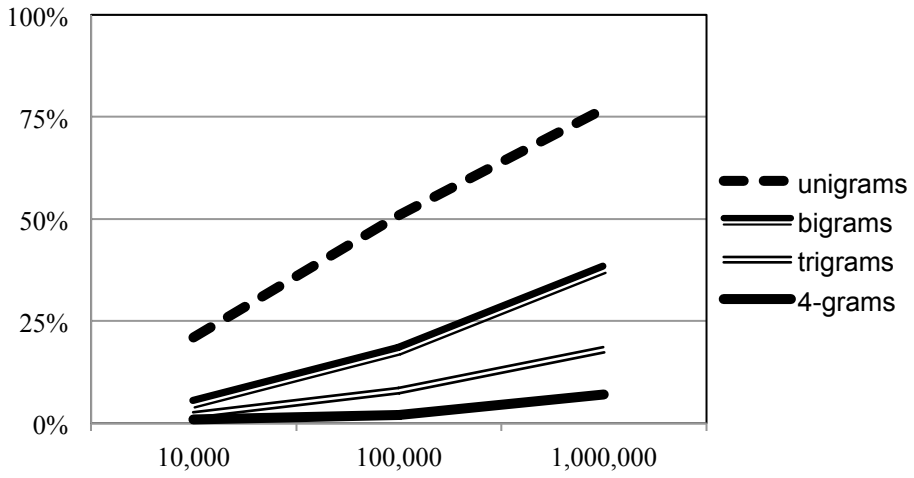

Figure 1 - Translation coverage of unique phrases from a test set

One can see that even when the system uses a corpus containing a million tokens, the total number of untranslated n-grams remain high.

When translating from 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 interested in studying the use of relatively small corpora for translation, matching phrases to examples in the corpus should be done on a spectrum of linguistic levels, so that not only exact phrases are discovered, but also related ones. In this chapter, we explore the effect of matching texts based on synonyms and single-word paraphrases.

We describe two experiments in which synonyms were considered in the matching process. In both experiments we used single-word "contextual" synonyms as a case study, with the intention of including longer semantically-replaceable phrases in the future. *Contextual synonyms* are words that are similar in meaning only within particular contexts. Figure 2 shows an example of contextual synonymous verbs in Arabic. The verbs *bv* and *nql* are semantic equivalents. One can see that the two sentences are almost the same; they differ only in the verbs. On the right side, the common meaning of verb *bv*, which is the stem of the word *bvt*, is "to broadcast", and on the left side, the common meaning of the verb *nql* (stem for *nqlt*) is "to convey". However, both sentences need to be translated in the same way – "the radio broadcast these songs". That means that the two verbs are exchangeable in this context (radio broadcasts) and are, therefore, semantic equivalents.

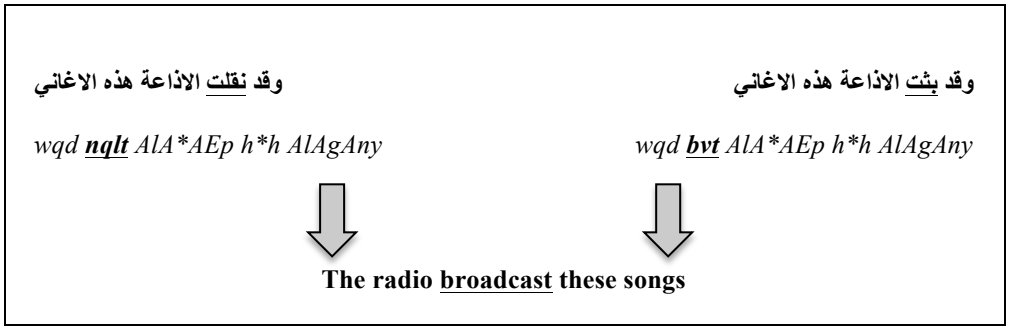

Figure 2 – Example of two semantically equivalent words

Hereinafter, whenever we use the word "synonym", we will mean such a contextual synonym. We will consider different techniques for the extraction of noun and verb synonyms, and will report on the results of both experiments.

Noun synonyms were automatically created for the first experiment using the stem list provided by the Buckwalter (version 1.0) morphological analyzer (Buckwalter, 2002). The word-pairs were organized in various levels of perceived synonymy. The quality of the system's resultant translations was 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.

Experiments were conducted on two English-Arabic corpora, one about twice as large as the other. The system was tested on all levels of synonymy and the effect of using the classification feature at each level was examined. The results, reported in (Bar and Dershowitz, 2010), show that, in general, the system performs slightly better when using synonyms.

In the second experiment, we automatically created a list of synonymous verbs and then used them to further the translation process. This time we used a corpus of Arabic "comparable" documents for learning synonym pairs. *Comparable documents* are texts dealing with the same event, but which are not necessarily translations of the same source. The basic learning technique we describe here was implemented merely on verbs, as motivation for future work seeking alternative approaches for learning large sets of (multi-word) Arabic paraphrases and using them to extend the coverage of a corpus-based translation system.

For the purposes of this research, 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 imitate 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.

Ever since first proposed by Nagao (1984), the example-based paradigm has been a fairly common technique for natural language processing (NLP), and 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 known example translations. Such a process presumably emulates the way a human translator operates in some cases. Since it uses real human-translated data, the resultant translations are usually more fluent than ones created artificially using other translation paradigms.

In general, our system performs three main steps in order to translate a given input sentence: matching, transfer and recombination.

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. A *match-score* is created for every matched fragment, which is based on the individual words matching levels. We elaborate more on this score in Section 3. In the *transfer step*, those matched phrases are translated using the target-language (English, in our case) version of a sentence-aligned parallel corpus. For each translated fragment, we calculate a *translation-*

score, which is the ratio between the number of translated words and the total number of words in the Arabic part of the fragment. 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.

In the *recombination step,* all the translated fragments are pasted together to form a complete target-language text, usually by preferring larger translated fragments, because they use more context. The recombination process is implemented similar to the way the decoding process is implemented within a statistical translation system. The difference is that recombination is based on the total-score of every fragment, which captures the various aspects of relevancy of a given fragment to the final translation.

We decided to work with an example-based system, rather than a statistical one, mainly because it is the more natural platform for using paraphrases during the translation process. Example-based systems store translations as pairs of strings augmented with some morpho-syntactic information. The source-language part of the examples is indexed on several morphological levels to expedite the on-line matching process. Once a fragment is found, the system calculates its translation score, which is based on the matching levels of the individual words and the number of covered source-language words by the extracted translations. However, it is

not based on statistics. Paraphrases can be simply indexed with their translations, extracted using their corresponding source-language corpus fragments. On the other hand, in a statistical translation system, one needs to calculate the translation probability of every paraphrase added to the system's phrase-table. Since the new, added paraphrases are not actually found within the loaded parallel-corpus, these probabilities are usually estimated using various techniques.

The following section summarizes some previous related work. Section 3 contains a general description of our example-based system. In Sections 4 and 5, we describe our experiments with noun and verb synonyms, respectively. In both sections, we provide some experimental results using common automatic evaluation metrics. Some conclusions are suggested in the last section.

2. Related Work

Nagao (1984) initiated the example-based approach to machine-translation with a structural Japanese-to-English translation system. Other influential works include (Sato and Nagao, 1990; Maruyama and Watanabe, 1992; Sumita and Iida, 1995; Nirenburg et al. 1994; Brown, 1999). This is the style of machine translation we are using in this work.

There are several works dealing with morphologically-rich languages such as Arabic. Nevertheless, we could not find any specific one 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 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. There, several morphological preprocessing schemes were applied separately on different size corpora.

In some work on Japanese-to-English example-based machine translation (Nakazawa et al., 2006), synonyms were used in the source

language for matching translation examples, similar to the idea presented here. However, the effect of this idea on the final results was not reported.

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).

Philips et al. (2007) present an Arabic-to-English example-based system. 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. It 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. Lin and Pantel (2001) extracted paraphrases from a monolingual corpus by measuring the similarity of dependency relationship. They use a syntactical parser to parse every sentence in their corpus and measure the similarity between paths in the dependency parses using mutual information. Paths with high mutual information were defined as paraphrases. Glickman and Dagan (2003) describe an algorithm for finding synonymous verbs in a monolingual corpus. This was also done using a syntax parser for building a vector containing the subject, object and other arguments for every verb

they find in their corpus. Later, they use these vectors to find similarities between verbs. Overall this technique showed competitive results with the one introduced by Lin and Pantel (2001). Nonetheless, since both techniques may perform differently on a given case, they suggested that the methods should be combined to obtain better results.

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 such as 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 about the same topic.

3. System Description

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). 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.

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 (*Aldusotuwriyħ*, "the constitutionality") and دستوريتي (*dusotuwriytiy*, "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 (*maAriq*, "apostate"); مراق (*mur~aAq*, "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 degree of ambiguity. Further investigation, as well as working with a context-sensitive morphology analyzer (Habash and Rambow, 2005), will allow us to better handle 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 95%, which is almost the maximum possible. Using a score of 100% reduces translation results because sometimes 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, because passing low-score fragments to the next step would dramatically increase the total running time and sometimes make it unfeasible to process all fragments.

4. Noun Experiment

In this experiment, we derive noun synonyms and use them in the matching step. Since Arabic WordNet is still under development, we developed an automatic technique for creating a thesaurus for Arabic nouns, using the Buckwalter gloss information, extended with English WordNet relations. Synonyms for verbs were created in a different way, since verbs seem to be more difficult to work with than nouns: the meaning of an Arabic verb usually changes when used with different prepositions. We describe our approach for automatically finding verb synonyms in the next section.

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 glosses. 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 considered 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 are co-translations, that is, they 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 3 shows an example of a relation between two Arabic stems. The stem اعادة (*AçAdħ*, "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.

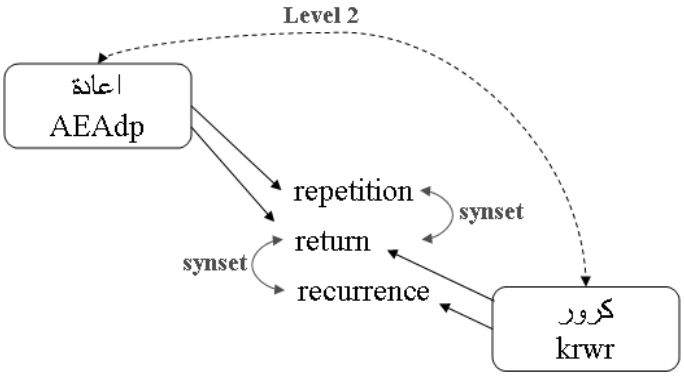

Figure **3**. Synonym relation, level-2 example

Synonyms	Level
nšyj / dmç ("crying")	4
sTH / sqf / ("ceiling")	5
zlçwm / Hlqwm ("throat")	1

njdħ / AçAnħ (“help”; “support”)	2
AbtdA' / ftH (“beginning”)	5
AxtrAç / AbtkAr (“invention”)	3

Table 1. Examples of extracted synonyms

Using Noun Synonyms for Translation

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.

Returning 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.

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. We evaluated the results under some of the common automatic criteria for machine-translation evaluation: BLEU (Papineni, 2002) and METEOR (Banerjee and Lavie, 2005). Table 2 shows some experimental results, presented as BLEU and METEOR score.

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	0.1187	0.1179	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 2. Experimental Results – BLEU and METEOR (MTOR) Scores

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. We can also see that experiments in which synonyms did not help improve translations significantly show a reduction in final scores when using classification. 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 3 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.

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

Table 3. Experimental results – Uncovered N-grams in the small corpus

5. Verb Experiment

For the second experiment we extract Arabic verb synonyms using corpus of comparable documents. The evaluation of this process was performed both manually and automatically. As in the noun experiment, we measure the translation quality of a system that uses the extracted synonyms in the matching step. In the next phase we will extend our work for the purpose of finding longer equivalents, also known as *paraphrases*, extracted from the same comparable corpora.

We chose to use comparable documents rather than monolingual or parallel corpus for finding Arabic verb synonyms. Existing parallel corpora for Arabic are usually translated to English and since we currently focus on

the Arabic-to-English direction, any additional parallel corpus could have been preprocessed by a translation system in the usual way. The contribution of additional equivalents that were extracted from a parallel corpus, which is loaded into the system in the traditional way, is expected to be very limited. However, we could use other parallel corpora, which pair Arabic with other languages, allowing us to use the other language as a pivot and find paraphrases in cases where two Arabic phrases share the same translation. Bannard and Callison-Burch (2005) implemented this idea using parallel corpora of French and Spanish paired with other languages. However, the parallel corpora we could find so far, that pair Arabic with other languages than English, contain a very limited quantity of sentences, which makes it irrelevant for the synonym extraction task. Since Arabic is one of the UN official languages, we could have built such corpora using the formal published documentation by the UN, provided in seven different languages. Using the same algorithm implemented by Bannard and Callison-Burch on automatically sentence-aligned Arabic-(the other 6 UN languages) corpora is something that we consider trying in the near future.

On the other hand, a large monolingual corpus is easy to obtain, but the context of candidate pairs is not defined. To overcome this problem, in most of the cases, a full syntax parser is used to find relevant syntactic features for describing the context of a possible synonym match. This fact

limits the language set of such an algorithm. Moreover, the lack of an indication of context forces the algorithm to be more restrictive when trying to match a potential candidate, usually resulting with fewer extracted synonyms. However, using syntax dependency information helps in the identification of discontinuous synonyms, which are very common in Arabic.

Our corpus was derived from the Arabic Gigaword corpus. Among all articles published by two news agencies (al-Nahar and al-Hayat) on the same dates, we took only those whose titles matched lexically. The matching criterion was simple: for every candidate pair of articles, we count the number of matched stems (using a morphological analyzer for Arabic) appearing in their titles and for each single article we choose another article having the larger number of matched stems to be its match. For the time being, we eliminated cases in which one article matched more than one document. Currently, we have only created small corpus, which is being used in the development steps.

We based our ideas on the work of Barzilay and McKeon (2001) on paraphrasing in English. The first task is to obtain a (partial) word alignment of every document pair. Since they used various English translations of the same source, the alignment could be obtained with less effort. This is not the case when using a corpus of comparable documents as a source for synonym extraction. Documents that deal with the same

event are not necessarily a translation of the same source, thus finding sentence alignments is impossible. Therefore, we decided to align together every verb pair sharing the same stem. Obviously, many-to-many alignment is also possible. In addition to this initial alignment, we created a list of potential synonym relations for a large list of Arabic verbs. This list was extracted using the English glosses provided with the Arabic stem list of the Buckwalter morphological analyzer, exactly as was done in the noun experiment using the English WordNet. In this case, stems that share at least one translation in common or whose translations are English synonyms were deemed to be synonyms. This is equivalent to taking all the synonyms of level 5 as defined in the noun experiment. Not like nouns, Arabic verbs tend to change their senses with different attached prepositions; therefore the extracted list of synonymous verbs is error-prone. Note that the Buckwalter's stem list does not contain prepositional information so this list cannot be treated as a thesaurus at all, but it will be used in a different way by our extraction algorithm.

Given the initial alignment for every document pair within the corpus, we start to look for those *contexts* in which verb synonyms exist (at this stage, only similar verbs). A context is defined as a list of features extracted from the *n* (*n*, a parameter, determines the context size) words on the left and right sides of the verb within the sentence and mainly contains morpho-syntactic information. The features that we intend to use are the

words' stem (since Arabic is a highly inflectional language) and part-of-speech tags, estimated automatically by AMIRA. Figure 4 shows an example for a verb context when *n*=2.

Sentence 1:
mktb Alsnywrp wdywAn Âwlmrt ***ynfyAn*** *xbrA ϛn lqA' fy šrm Alšyx.*

Translation:
Seniora's office and Olmert's administration deny a story about a meeting in Sharm al-Sheikh.

Sentence 2:
mktb Alsnywrħ ***ynfy*** *xbrA ϛn lqAŷh msŵwlyn ĂsrAŷylyyn.*

Translation:
Seniora's office denies a story about a meeting with Israeli officials.

Verb: *nfy* (*ynfyAn*, *ynfy*)

Context:
Left-1: (NN, NNP) Right-1: (NN_1, IN_2)
Left-2: (NN, NNP) Right-2: (NN_1, IN_2)

Figure 4 – An example of a context

In this example, the verb *ŋfy* ("deny") appears in the first sentence in its imperfect dual form, while in the second sentence it is in its imperfect singular form. The stems are not used as direct features, but only to indicate equality of words located in both context parts. In such a case, the part-of-speech tag will contain the index of the matched word in the other part of the context, as can be seen in this example: on the hand right sides, the two following words (*xbrA*, *ϛn*) are exactly the same, however on the left hand side there are no stem-based similar words.

We will consider adding other features. Matching context content words (not functional) based on a direct WordNet hypernym relation is possible. In such cases we will match the Arabic equivalents to “blue” and “green” as both words being part of synsets with direct hyponym relation to the same synset, namely, color. The main challenge here is the lack of a robust and extensive WordNet for Arabic. Instead, we will use the English WordNet for the gloss entries of the words’ stem.

Arabic verbs often use additional particles to highlight its object. Different particles can completely change the meaning of the verb. For instance, the meaning of the direct-object version of the transitive verb *qDy* is “to judge” and the meaning of the same verb using the preposition *ςly* to mark the object is “to kill”. Therefore, we should also use the word that appears right after the first preposition as part of the context.

Although parsing Arabic text is a difficult task, there have been recent works on dependency parsing in Arabic that we will consider using to locate the subject and object of each verb and then consider them as the context, instead of choosing the immediate words surrounding the verb. Arabic sentences are often written with many noun and verb modifiers and descriptors, so we think that using such parser will help produce more accurate synonyms.

Based on the ideas of Barzilay and McKeon (2001), we can 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 (similar verbs), n is the number of times it appears in a negative example (non-similar verbs), and N is simply the frequency of the context in the entire corpus. We then select the most frequent k positive and negative contexts (k, a parameter) that their strength is higher than a predefined threshold and use them for extracting synonymous verbs. We do this by finding all instances of each selected positive context that are not covered by a negative context in every document pair. The verbs that are surrounded by those contexts are deemed synonymous. Since we do not use word alignment of any kind, finding negative examples seems to be non-trivial. For this reason, we previously created the potential synonym verbs list. In every document pair, we look for verb-pair candidates, which are not even synonyms based on the potential list. Such verb pairs are marked as negative examples.

To evaluate this process, we will examine a random number of the resultant synonyms with their extracted contexts. Callison-Burch et al. (2008) proposed ParaMetric, an automatic metric for measuring paraphrasing techniques. The evaluation in this setting is done using manually word-aligned groups of sentences in which sentences are translations of the same source. This gold-standard set was used to test different paraphrasing techniques, calculating the relative recall and the

lower bound precision of each one of them. However, this was done based on a corpus of multiple English translations of the same Chinese texts, which was made to serve machine translation evaluation campaigns. Unfortunately, there is no similar corpus available for Arabic, so for the time being at least we have had to abandon this direction.

An expert will evaluate our results by examining some number of pairs, given along with the contexts in which they were found, and decide whether the verbs are synonyms (at least in one context) or non-synonyms (wrongly identified as synonyms in all the given contexts). We want to measure the precision and relative recall, based on manually tagging paraphrase relations between all candidate pairs within a limited set of compared documents.

Using Synonyms in Translation

As in the previous experiment, we allow the system to find matches on a synonym level. As before, the evaluation of this process was done automatically using the common metrics (BLEU, METEOR). Some features were tested for their effect on the final results separately. (In this experiment, we did not check the effect of using a topic classifier as a tool for deciding when synonym should be considered and when they should not.)

Since the automatic metrics may hide the improvement of using paraphrases in the internal matching step, we also performed a manual evaluation. An expert evaluated a predefined small test-set. The expert scored the synonym-based extracted patterns, given with their contexts. First, the source-language part corresponding to the input sentence fragment was examined, and then the quality of the translations was.

Experimental Results

We used 5500 document pairs, extracted from the Arabic Gigaword. The total number of words is about three million. The context window that we are using is of size 2, meaning that we consider all the possible contexts surrounding a verb with the limitation of one or two words before the candidate verbs and after. The strength threshold for selecting the best contexts of both categories is 0.95, as suggested by Barizlay and McKeon, and the number of best contexts (defined as k above) we use is 20. (We found that, k=10, as used by Barzilay and McKeon, is too restrictive in our settings, and that, therefore, the result set was relatively small).

Two experts evaluated the resultant verb pairs. The verb pairs were given along with the different contexts in which they were found. For each candidate pair, each expert was requested to make one of the following decisions: *correct*—verb instances are exchanged in some contexts; or

incorrect—verb instances are not exchangeable at all. There needed to be at least one context in which a verb pair is semantically replaceable in order for them to be marked as correct paraphrases by the experts. They were also allowed to say that verbs are correct paraphrases even if their meaning is modified by another word in the context. Recall that the system was instructed to identify whether two verbs have the same meaning in a given context. Therefore, it decides so even if the meaning is defined by an expression of more than one word, including the target verb. Table 4 shows the expert decisions.

Unique Candidates	**Unique Synonyms**	**Expert 1: Correct Synonyms**	**Expert 2: Correct Synonyms**
15,101	139	120 (**86%** precision)	103 (**74%** precision)

Table 4. Experts' evaluation

In all, we found about 15,000 unique (based on the verbs' lemmas) candidates. Of these, the classifier decided that only 139 are synonyms. While 120 were found to be correct by the first expert, the second expert found only 103 to be correct, yielding precision values of 86% and 74%, respectively. Since we do not know how many of the 15,000 candidates are actually synonyms, we have not calculated recall. Table 5 shows some synonym examples that were extracted by this technique and Table 6 shows two of the best contexts for the positive candidates.

Synonyms
Ăçtql / Âwqf (“arrest”)
bθ~ / nq~l (“broadcast”)
Ăstqbl / Ăltqy (“meet”)

Table 5. Examples of extracted synonyms

Best Contexts
Left-1: (NN0) Right-1: (IN, NN) Left-2: (NN0) Right-2: (IN)
Left-1: (NN, WP0) Right-1: (NN0) Left-2: (WP1) Right-2: (NN0)

Table 6. Some of the best contexts for the positive candidates

Once we found the abovementioned synonyms, we used them in translation, under settings similar to those described in the noun experiment. In this case, we tested the system only using the corpus containing about 1.2 million Arabic words and on the same test-set: 586 sentences from 68 documents (17370 words) taken from the 2009 NIST MT Evaluation data comparing to four reference translations. We automatically evaluated the results under BLEU and realized that there was only a slight insignificant improvement in the final results. We decided to examine the results manually, this time it was done only by the first author. Overall, we found 193 input sentence fragments for which at least one of the words matched on a synonym level. For simplicity, we will call these *syn-fragments*. For 162 syn-fragments, the synonym matching is correct

under the sentence context. Out of these syn-fragments, 57 are covering parts in the input sentence that are not covered by other fragments of at least the same size. That means they might help to better cover the input sentence in the matching step, however our current recombination algorithm was not able to capture that. We further looked at the extracted translation for the 193 syn-fragments and found that only 97 were actually translated correctly. All the other syn-fragments received wrong translations by the system. From a first look, in most cases, the synonyms were not the main reason for the wrong translation. It seems more like the traditional problem of error-prone word alignment, affecting the translation of the fragments. Only 63 syn-fragments participated in final translations; out of them, only 42 were translated correctly. Seeing these results, one can conclude that, unsurprisingly, the system is making bad choices when it tries to select the best fragments for incorporation in the final translations. Remember that our current example-based system is using a very simple recombination technique, and it still needs to be adapted to use a more standard model. That can explain why we could not see a real improvement in BLEU score.

6. Conclusions

The system we are working on has demonstrated a promising potential for using contextual synonyms in an example-based approach to machine translation, for Arabic, in particular. We found that noun synonyms benefit from being matched carefully by considering the topic of the sentence 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. Sometimes, 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.

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 standard 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 will continue working on different methods for automatic extraction of semantic equivalents for Arabic.

As demonstrated in the verb experiment, the classifier, which has been trained to find new verb synonyms in a corpus of comparable documents, performs pretty well in terms of precision. Even though we have not yet calculated the recall, by manually overlooking at the candidates, we could see some that there are true synonyms that were not yet discovered. We are now working on different techniques for building a new classifier for extracting semantic equivalents from a corpus of comparable documents. We even plan to consider more context features. For instance – marking words that are derived from the same WordNet type. A dependency parser seems to be a powerful tool for finding equivalents. Had we had one, we could have used the subject and the object phrases as the context of the verbs rather than using a fixed-size surrounding window. Since there are some recent works in this area, we will be considering incorporating such tools.

The two experiments described here are just first steps toward the larger goal of deriving longer paraphrases for Arabic and using them to improve machine translation.

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.

Bannard, Colin and Chris Callison-Burch. 2005. Paraphrasing with bilingual parallel corpora. In *Proceedings of the Annual Meeting of the Association of Computational Linguistics (ACL 2005)*. Ann Arbor, MI, 597-604.

Bar, Kfir and Nachum Dershowitz. 2010. Using synonyms for Arabic-to-English example-based translation. In *Proceedings of the Ninth Conference of the Association for Machine Translation in the Americas (AMTA 9)*. Denver, Colorado.

Barzilay, Regina and Kathleen McKeown. 2001. Extracting paraphrases from a parallel corpus. In *Proceedings of the Annual Meeting of the Association of Computational Linguistics (ACL 2001)*, Toulouse, France, 50-57.

Brill, Eric. 1992. A simple rule-based part-of-speech tagger. In *Proceedings of the DARPA Speech and Natural Language Workshop*. Morgan Kauffman. San Mateo, CA, 112-116.

Brown, Ralf D. 1996. Example-based machine translation in the Pangloss system. In *Proceedings of International Conference on Computational Linguistics (COLING-96)*, Copenhagen, Denmark, Vol. 1, 169-174.

Brown, Ralf D. 1999. Adding linguistic knowledge to a lexical example-based translation system. In *Proceedings of the International Conference on Theoretical and Methodological Issues in Machine Translation (TMI)*, Chester, UK, 22-32.

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

Callison-Burch, Chris. 2007. *Paraphrasing and Translation.* PhD dissertation, University of Edinburgh.

Callison-Burch, Chris, Trevor Cohn, and Mirella Lapata. ParaMetric: An automatic evaluation metric for paraphrasing. In *Proceedings of the 22nd International Conference on Computational Linguistics (COLING 2008),* 97-104.

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*, 49, 1, 311-326.

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, 81-88.

Glickman, Oren and Ido Dagan. 2003. Identifying lexical paraphrases from a single corpus: A case study for verbs. In *Proceedings of* the *Recent Advantages in Natural Language Processing (RANLP-03),* 81-90.

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.

Lance A. Ramshaw and Mitchell P. Marcus. 1995. Text chunking using transformation based learning. In *Proceedings of the Third ACL Workshop on Very Large Corpora.*

Lee, Young-Suk. 2004. Morphological analysis for statistical machine translation. In *Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics (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 Annual Meeting of the Association for Computational Linguistics and International Conference on Computational Linguistics (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 the 4th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI)*, 173-184.

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, 173-180.

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 *Proceedings of the International Conference on New Methods in Language Processing (NeMLaP)*, Manchester, UK, 78-87.

Lin, Dekang and Patrick Pantel. 2001. Discovery of inference rules for question answering. *Natural Language Engineering*, 7(4):343-36.

Papineni, Kishore, Salim Roukos, Todd Ward and Wei-Jing Zhu. 2002. **B**LEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the 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*, New York, 49-52.

Sato, Satoshi and Makoto Nagao. 1990. Toward memory-based translation. In *Proc. 13th International Conference on Computational Linguistics (COLING)*, Vol. 3, 247-252.

Stroppa, Nicolas, Declan Groves, Kepa Sarasola and Andy Way. 2006. Example-based machine translation of the Basque language. In *Proceedings 7th Conference of the Association for Machine Translation in the Americas*, 232-241.

Sumita, Eiichiro and Hitoshi Iida. 1995. Heterogeneous computing for example-based translation of spoken language. In *Proceedings of the International Conference on Theoretical and Methodological Issues in Machine Translation (TMI)*, 273-286.