An Arabic to English Example-Based Translation System

K. Bar, Y. Choueka, and N. Dershowitz

Abstract—**An implementation of a non-structural Example-Based Machine Translation system that translates short sentences from Arabic to English, using a large parallel corpus aligned at the paragraph level, is described. Each new input sentence is matched to example patterns by using various levels of morphological data. We encountered several problems in the matching and the transfer steps, some of which were solved, partially or totally, sometimes by using linguistic tools for both languages. We discuss those problems and our proposed solutions.**

The system has been implemented and automatically evaluated. Results are encouraging.

Index Terms—**Example-Based Machine Translation, Arabic**

I. Introduction

THE example-based (or "memory-based") paradigm has become a fairly common technique for natural language processing (NLP) and especially for machine-translation applications, ever since it was first proposed by Nagao in [1]. That paper expressed the main idea behind an example-based machine translation (EBMT) paradigm, namely to emulate the way a human translator operates in some cases. Such a system exploits a large bilingual corpus to find similar examples for fragments of the input source-language (Arabic, in our case) text, and imitate its translations [2]. Searching for similar fragments is called *matching*. Given a group of matched fragments, the next step is to extract possible translations from the target-language (English, in our case) version of the corpus. This is the *transfer* step. The last step is *recombination*, which is the generation of a complete target-language text, pasting together translated fragments. Fig. 1 outlines an example-based system for Arabic to English. The reader may refer to the comprehensive survey of example-based machine-translation systems by Somers [3].

We describe an implementation of the major components of an EBMT system that translates short Modern Standard Arabic (MSA) sentences into English. It is a non-structural system, so it stores the translation examples as textual strings, with some additional morphology and part-of-speech information. Our work is still in progress. Currently, the system fragments any new introduced input sentence and translates each fragment separately. Recombining those translations into a final coherent form is left for future work.

Our final goal is to develop an automated assistant for Arabic-to-English machine translation systems that work within a rule-based or statistical paradigm, so as to better handle complicated cases and especially to improve the fluency of the generated translations.

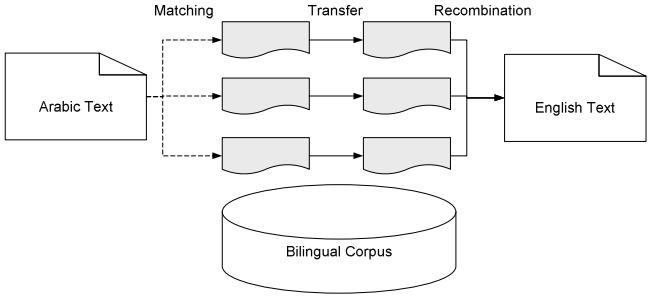

Fig. 1. Main steps of example-based translation system.

The following section is a general description of our system. In Section III, we give some experimental results using common automatic metrics. Conclusions are presented in Section IV.

II. System Description

A. Preliminaries: Storing Translation Examples

The translation examples we need were extracted from a collection of parallel unvocalized Arabic-English documents taken from the United-Nations document inventory available under the Official-Document-System (ODS) [4]. We automatically aligned each parallel document on the paragraph level and each parallel paragraph was taken as a translation example. These examples were morphologically analyzed using the well-known Buckwalter morphological analyzer (version 1.0) [5], and part-of-speech tagged using SVM-POS [6], in such a way that, for each word, we considered only the relevant Buckwalter analyses with the corresponding SVM-POS's part-of-speech tag. A special look-up table that maps Arabic words to their corresponding English words in each parallel paragraph was also created. Actually, for each Arabic word in the translation example, we look up its English equivalents in the lexicon and expand that with synonyms from WordNet. Then we search the English version of the

Manuscript received January 7, 2007.

K. Bar, Dept. of Computer Science, Tel Aviv University, Ramat Aviv, Israel (e-mail: kfirbar@post.tau.ac.il).

Y. Choueka, Dept. of Computer Science, Bar-Ilan University, Ramat Gan, Israel (e-mail: ycsarah@netvision.net.il).

N. Dershowitz, Dept. of Computer Science, Tel Aviv University, Ramat Aviv, Israel (e-mail: nachumd@post.tau.ac.il).

translation example for all instances on the lemma level and insert them in the table.

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

B. 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. A matched fragment must contain at least two adjacent words in the same input sentence. The same fragment can be found in more than one translation example. Therefore, a special *match-score* is assigned to each fragment-translation pair, representing 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.

Words are matched on *text*, *stem*, *lemma*, and *part-of-speech* levels, with each level assigned a different score. Text (exact string) and stem matches credit the words with the maximum possible; a lemma match credits them with less and part-of-speech credits the fragment match-score with a minimal amount. Table I summarizes the several match levels we used in our experiments.

TABLE I
WORD MATCHING LEVELS

Match Level	Description	Match Score
Text	Exact match of the words.	1
Stem	Words match in their stems but not in their surface form. For instance, the words الدستورية (*Aldstwryp*, "the constitutionality") دستوريتي (*dstwryty*, "my constitutional") share the stem دستوري (*dusotuwriy*)	0.9
Lemma	Words share a lemma. For instance, the following words match in their lemmas: مارق (*mAriq*, "apostate") مراق (*mur~Aq*, "apostates") Note that the stems of these words are not the same.	Dynamic score
Content	This level is planned but not yet implemented. The idea is that, for example, two location names would get a higher score than two dissimilar proper nouns.	0.8
Part-of-Speech	Words match only in their part-of-speech. For instance, both are nouns. Actually, we require that both also have the same tags for their affixes. For example, if a word is tagged as a noun and has the definite article prefix ال (*Al*, "the"), the matched word must agree on both features – it must be a noun and also have the definite article prefix.	0.3
Common Word Match	This level is relevant only for common words and affixes, taken from a predefined list. These words/affixes are organized in groups that represent the same meaning. Clearly, a word/affix may be a member of more than one group. Words/affixes that are members of the same group are also matched on this level. For example the prefix ب (*b*, "with", "by", "in") is in the same group of the preposition في (*fy*, "in").	1

Text and stem match receive almost the same score since, currently, we do not yet handle the translation modification needed. When dealing with unvocalized text, there are, of course, complicated situations when both words have the same stem but different lemmas, for example, the words كتب (*katab*, "wrote") and كتب (*kutub*, "books"). Such cases are not yet handled, since we have not worked with a context sensitive Arabic lemmatizer and so cannot derive the correct lemma of an Arabic word. Still, the combination of the Buckwalter morphological analyzer and the SVM-POS tagger allows us to reduce the number of possible lemmas for every Arabic word so as to reduce the amount of ambiguity. Actually, by lemma match, we mean that words match on any one of their possible lemmas. The match-score in such a case is the ratio between the number of equal lemmas and the total number of lemma pairs (one per word). Further investigation, as well as developing and working with a context sensitive Arabic lemmatizer, is needed to better handle all such situations.

Fragments with a score below some predefined threshold are discarded, since passing low-score fragments to the next step dramatically increases total running time. Note that a larger corpus, with the concomitant increase in the number of potential fragments, would require raising the threshold.

Fragments are stored in a structure comprising the following: (1) *source pattern* – fragment's Arabic text, taken from the input sentence; (2) *example pattern* – fragment's Arabic text, taken from the matched translation example; (3) *example* – the English translation of the example pattern; (4) *match score* – of the fragment and its example translation.

For efficiency, fragments sharing the same example pattern are collected and stored in a higher-level, *general-fragment* structure. (Note that a general-fragment consisting of only one fragment is also possible.)

C. Transfer

The input to the transfer step consists of all the collected general-fragments that were found in the matching step, and its output is the translations of those general-fragments. The translation of a general-fragment is taken to be the best generated translation among the comprised 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 so that it will be the translation of the fragment's source pattern.

1) First Step – Translation Extraction

The first step is to extract the translation of the fragment's example pattern from the English version of the translation example. Here we use the prepared look-up 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. Then, we extract the *shortest* English segment that contains the *maximum* number of equivalence words. Usually a word in some 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, bound 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 II is the corresponding look-up table. Now, suppose the example pattern is ميدان حقوق الإنسان (*mydAn Hqwq Al<nsAn*, "the field of human rights"), so we want to extract its translation from the English version of the translation example. Using the extracted look-up, we mark the English equivalences 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 equivalent words, viz. "field of human rights".

TABLE II
ALIGNMENT LOOK-UP TABLE

English	Arabic
Services	الخدمات
Advisory	الاستشارية
Cooperation	والتعاون
Technical	التقني
In	في
Field	ميدان
Rights	حقوق
Human	الإنسان

This is of course a simple example. More complicated ones would have more than one equivalent for each Arabic word.

Sometimes it is hard to find the corresponding English equivalents for a specific Arabic word. Usually this happens when the Arabic word is part of some phrase, whereas its translation does not follow word for word, as in, for example, the Arabic example pattern غير رسمي (*gyr rsm)*, meaning "not formal". In many cases, we might find "informal" in the English version instead. The problem is that neither the synonym list of the word رسمي (*rsmy*, "formal"), nor the list of the word غير (*gyr*, "not"), contains the word "informal". Such a situation is handled by a manually defined rule that is triggered whenever the word غير (*gyr*, "not") appears. The system checks the following word, and -- instead of building a synonym list -- builds an antonym list, using WordNet. In this example, the word "informal" appear as an antonym of the word "formal" in WordNet.

There are more complicated structures that are not handled yet, but capturing and writing rules for such cases seems quite feasible.

2) Second Step – Fixing the Translation

Recall that the match of a corpus fragment to the input fragment can be inexact: words may be matched on several levels. Exactly matched words are assumed to have the same translation, but stem or lemma matched words may require modifications (mostly inflection and prepositions issues) to the extracted translation. These issues were left for future work. Words matched on the part-of-speech level require 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") are matched on the part-of-speech 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*, "the security responsibility") and is 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: "the security council". The lexicon is implemented using the glossaries extracted from the Buckwalter morphological analyzer and expanded with WordNet synonyms as was explained above.

Sometimes the extracted translation contains some extra unnecessary words in the middle. Those words appear mostly because of the different structure of a noun-phrase in both languages. For example, consider the example, موضوع الامن الاقليمي (*mwDwE AlAmn AlAqlymy*), and its translation: "the subject of regional security". By extracting the translation of the pattern موضوع الامن (*mwDwE AlAmn*), we obtain: "the subject of regional security" (since it is the shortest segment that contains maximum word alignments). Clearly, the word "regional" is unnecessary in the translation because it is the translation of the word الاقليمي (*AlAqlymy*, "the regional") that does not appear in the pattern. So by removing that word from the translation we obtain the correct translation of the pattern. The word "regional" appears in the extracted translation due to the fact that Arabic adjectives come after the nouns they qualify, which is the opposite of English syntax. Here, the noun-phrase الامن الاقليمي (*AlAmn AlAqlymy*, "the regional security") is translated so that the translation of الاقليمي (*AlAqlymy*, "the regional") appears before the translation of الامن (AlAmn, security). Currently, identifying such situations is done by searching for the translation of the word "regional" in a fixed number of Arabic words that come immediately after the pattern in the translation example. However, this method is insufficient for more complex situations and is also very time consuming. Our plan is to apply an Arabic chunker to extract the boundaries of the noun-phrase and in that way delimiting the search area.

Removing unnecessary words from the extracted translation must preserve the correct English syntax of the remaining translation, which in some cases seems to be a difficult task.

For that purpose, we have compiled several rules to deal with different situations. These rules are based on the syntax of the English extracted translation and identify cases that need special care. First, we chunk the translation to discover its basic noun-phrases, using the BaseNP [7] chunker. To do that, we first apply Brill's part-of-speech tagger [8] to the translation. Then, by looking at the chunked English text, we can ascertain the effect of removing the unnecessary word. In the previous example, removing the word "regional" from the text, "the subject of regional security", may be done without any further modification, since by tagging and chunking the segment we get

[the/DT subject/NN] of/IN[regional/JJ security/NN]

(the phrases in brackets are noun-phrases) and "regional" is simply an adjective within a noun-phrase, which still has the same head. Prepositions and other function-words that relate to the phrase are still necessary, so we keep them.

As already mentioned, a general-fragment may contain several fragments sharing the same Arabic example pattern. Among the extracted translations of the comprised fragments, which are all translations of the same Arabic pattern, we choose the translation that covers the maximum number of Arabic words to represent the general-fragment. The *translation-score* calculated for the chosen translation is the ratio between the number of covered words and the total number of words in the Arabic pattern. The *total-score* of a general-fragment is the multiplication of its match-score and its translation-score.

D. 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 *N* best recombinations of the extracted translations that cover 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; the second is left for future work. By multiplying the total-scores of the comprised general-fragments, we calculate a *final-translation-score* for each generated recombination. The *N* best (where *N* is configurable) recombinations are reported.

III. Experimental Results

Experiments were conducted on a corpus containing 13,500 translation examples. The following results are based on 400 Arabic short sentences (5.5 words per sentence on average) that were taken from unseen documents of the United-Nations inventory. The ten best results were evaluated by some of the common automatic criteria for machine translation evaluation (BLEU [9], NIST, and METEOR [10]), although our system is still under construction. Also, we used only two different translation references for the evaluation. Table III shows some preliminary experimental results. The first row contains the results of evaluating the system's highest ranked translation for each input sentence. The second is the same, but on the best translation from the viewpoint of a human referee. In most cases, the best translation chosen by the referee had a close (or even the same) final-translation-score as the system's best translation.

TABLE III
Evaluation Results

	BLEU (4-gram)	NIST	METEOR
Best translation chosen by the system	0.1849	4.1792	0.4851
Best translation chosen by a human referee	0.2488	5.1281	0.5363

IV. Conclusion

We believe we have demonstrated the potential of the example-based approach for Arabic, with only minimum investment in Arabic syntactical and linguistic issues. We found that matching fragments on the level of lemma and stem, as well as part-of-speech, enabled the system to better exploit the small number of examples in the corpus we used. More work is needed to enlarge and enrich the corpus, as well as to formulate rules to deal with various problematic situations that are not yet handled. This all appears quite feasible. Finally, we do not claim that the example-based method is sufficient to handle the complete translation process. It seems that, for Arabic, it should work together with some kind of rule-based engine, as part of a multi-engine system, so as to better handle more complicated situations.

References

[1] M. Nagao, "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, 1984.

[2] S. Sato, and M. Nagao, "Toward memory-based translation," *COLING 13*, vol. 3, pp. 247-252, 1990.

[3] H. L. Somers, "Review article: Example-based machine translation", *Machine Translation* **14**, pp. 113-157, 1999.

[4] United Nations Official Document System (ODS), URL - http://www.ods.un.org (viewed on 29/11/06).

[5] T. Buckwalter, "Buckwalter Arabic Morphological Analyzer Version 1.0". Linguistic Data Consortium, Philadelphia, 2002. URL http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2002L49 (viewed on 21/11/2006)

[6] M. Diab, K. Hacioglu and D. Jurafsky, "Automatic tagging of Arabic text: from raw text to base phrase chunks", The National Science Foundation, USA, 2004.

[7] L. Ramshaw and M. Marcus, "Text chunking using transformation-based learning", In *Proceedings of the 3rd ACL Workshop on Very Large Corpora*, MIT, 1995.

[8] E. Brill, "A simple rule-based part of speech tagger", In *Proceedings of the DARPA, Speech and Natural Language Workshop*. pp. 112-116. Morgan Kauffman. San Mateo, California, 1992.

[9] K. Papineni, S. Roukos, T. Ward and W. J. Zhu, "Bleu: a method for automatic evaluation of machine translation", *In proceedings of the ACL 40th annual meeting*, pp. 311-318, Philadelphia, PA, July, 2002.

[10] S. Banerjee and A. Lavie, "Meteor: an automatic metric for MT evaluation with improved correlation with human judgments", In *Proceedings of the ACL 43th Annual Meeting.Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization*, pp. 65-72, Ann Arbor, MI, June, 2005.