English to Arabic Statistical Machine Translation System
Improvements using Preprocessing and Arabic Morphology Analysis

Shady Abdel Ghaffar¹, Mohamed Waleed Fakhir²
¹Faculty of computing and Information Technology
Arab Academy for Science and Technology
Sheraton, Cairo
Egypt
shady_fcis@yahoo.com, waleedf@aast.edu

Abstract: - In this paper we show how to achieve a significant increase in Bleu score in case of English to Arabic Statistical Machine Translation (SMT) by making some preprocessing for both English and Arabic and also using Morphological splitting of Arabic. The preprocessing involves numbers, dates and person names clustering. The morphological splitting uses Columbia University Arabic language analysis tool (MADA) and the SMT is using MOSES and GIZA++ tools.

Key-Words: - SMT, Bleu score. English to Arabic, Morphological analysis, proper nouns clustering.

1 Introduction

Machine Translation (MT) is the use of computers to automate some or all of the process of translating from one language to another. Many useful applications for MT including Cross-Language Information Retrieval (CLIR) which is a type of information retrieval where the language of the query and the language of the searched text are different; for example, searching Arabic text using English query. The World Wide Web nowadays contains tons of useful information presented in many languages. A typical internet user needs a machine translation system that is capable of delivering ideas and concepts presented in other languages to the user's language. Translating weather forecasting, News and computer manuals are very popular applications for MT. One-to-Many MT is applicable in translating manuals, books, and news. Many-to-one translation is required in translating the web content. An example for Many-to-many translation is the European Union where 23 official languages need to be intertranslated. Machine translation is a hard problem for several reasons; first languages are different at several levels; we have typological differences. At word level, words in different languages have different number of morphemes varying from one morpheme per word like Vietnamese (isolating languages), to many morphemes per word (polysynthetic languages). At syntactic level we have SVO languages (Subject Verb Object languages) like French, English and German, SOV languages (Subject Object Verb languages) like Hindi and Japanese, and VSO languages (Verb Subject Object languages) like Arabic and Hebrew. In addition we have lexical divergence; a word may have multiple senses, but only one in the context so, we need to have word sense disambiguation. On the other hand a word might be translated using one or more words in the target language [1].

Arabic is a highly inflected language where each word is inflected for gender and number. In addition a word may construct a meaningful sentence in its own. This makes word level alignment algorithms give bad alignment results [2]. For this reason we need to think of a way to improve the alignment quality to achieve good translation results. We can make use of morphological analysis as a preprocessing to resolve word level ambiguity and generate good alignment.

In this paper we discuss various preprocessing tasks that affect the Bleu score for English to Arabic Statistical Machine Translation in addition we show that using morphology analysis affects the Bleu score. Section II describes various Machine Translation techniques. Section III describes related works for both English to Arabic and Arabic to English SMT. In section IV we discuss several preprocessing tasks that affect the Bleu score when translating from English to Arabic. Then section V describes applying Morphology analysis. Section VI describes post processing. Then section VII describes the baseline experiment and how the preprocessing affects the Bleu score. Finally MADA splitting experiments and how we make use of Morphology analysis. Section VIII is the discussions and conclusions and finally section IX is the future work.
2 MT Approaches

The different MT approaches can be grouped into two main camps, the rule based (RBMT) and the statistical based (SMT) approaches [1, 3].

RBMT approaches based on explicit rules those are put by expert linguists. In its pure form RBMT can be applied at different levels including Syntactic Transfer which uses hard coded rules to figure out the syntactic mapping between the source and the target language, other technique is the Interlingua MT, which attempts to model semantics. In general RBMT requires rules and dictionaries which models the mapping between the source and the target language at the lexical and the syntactic levels those rules are developed manually or semi-automatically by language experts and software developers.

SMT is corpus based. SMT make use of translation samples called parallel/bilingual corpus. SMT in its basic form do the following. Given a sufficient sample of parallel text that is human translated the words are automatically aligned for each sentence pair. Then a translation model is learnt from the word alignment. The translation model basically models the words sequence mapping between the source and the target language. Then a decoder combines the translation model together with a language model for the target language to generate a ranked list of optimal translations.

RBMT was dominating the field of MT for many years; however over the last two decades researches for SMT have become very successful. The main motivation for this is the explicit linguistic rules can be probabilistic and can be learnt from parallel corpora. The last few years have witnessed an increasing interest in hybrid approaches between SMT and RBMT these approaches make use of both linguistic rules and statistical techniques. The most successful of such attempts so far are solutions that build on statistical corpus-based approaches by strategically using linguistics constraints or features [3].

2.1 Statistical Machine Translation

SMT make use of the Bayesian Noisy Channel model. For example in case we are translating from English to Arabic the model assumes that the Arabic sentence has been distorted by the noisy channel as a result we have the English sentence [1, 3]. Our task is to recover the original Arabic sentence. In other words we need to find the proper Arabic sentence that is the most probable translation for a given English sentence as shown below using the Bayes probability rules:

\[ A^* = \arg \max_A P(A | E) = \arg \max_A P(E|A) \times P(A) \]  

P(A|E) represents the faithfulness of mapping between the source and target languages, while the P(A) represents the fluency of the translated target language sentence.

The noisy channel model requires three components. Translation model, language model and a decoding algorithm to find the sentence that maximizes the above equation.

P(E|A) is the translation probability (the probability that the given English sentence is mapped to the generated Arabic sentence). We can estimate it by multiplying phrase translation probability and the distortion probability (reordering probability). We can think of any other models that maximizes the translation probability.

We call phrase translation probabilities Phrase Table is a bilingual mapping between source and target phrases and the mapping probability. Phrase table is extracted from the word level alignment. A phrase is a group of contiguous words.

Many models have been developed to generate word alignment given large parallel copra including EM algorithm, IBM model 1, 2, 3 and HMM based word alignment [1, 3].

Decoding algorithm searches the phrase table for the set of phrases that translates a given sentence and maximizes equation (1). Best first search algorithms are used like A* and beam-search algorithm.

3 Related Work

Arabic is a highly inflected language. Words are inflected for gender, number and some grammatical cases, but English is not. This mismatch between English and Arabic makes automatic word alignment between sentences pairs is a non-trivial problem. Therefore, efforts have been made to make English phrases match Arabic phrases to improve automatic alignment quality. In prior work [2] it has been shown that morphological segmentation of Arabic source makes a significant increase in Bleu score in Arabic to English SMT. However, English to Arabic SMT requires recombination. The better the recombination is the higher Bleu score is achieved. English to Arabic SMT is more difficult than Arabic to English SMT since the output in this case is segmented Arabic which requires recombination to construct Arabic words. The Recombination problem is non-trivial problem.
because Arabic is highly inflected language. In prior work [4] several recombination techniques were introduced. Those techniques are recombination table and a set of hard coded morphological rules that are obtained from the training set. In this paper, we compare the word-based system with and without preprocessing with the splitting-based system with and without preprocessing.

4 Preprocessing

Before we do training for our machine translation system we have done some preprocessing to the parallel corpus. We do simple tokenization, removing punctuations, normalizing all forms of Alef Hamza to bare Alif and final ‘Y’ Alif Maksora to Yaa. Numbers, numeric dates, times and percentages are not translated. In addition there is a very large number of values for these categories and only few of them may appear in the training and tuning data. This decreases the quality of language model and alignment. As a preprocessing we replaced all numbers, numeric dates, times and percentages by special tags (B) for numbers, (C) for percentages and (Q) for dates. To improve alignment quality we choose the maximum sentence length to be 40 words.

In another experiment we replaced all person names in both Arabic and English by the tag (PRN). We will show that this preprocessing affects the Alignment quality and the Bleu score in a positive way.

4 Morphology Analysis

Each Arabic word has multiple possible analyses. When a word appears in a sentence it has only one analysis. We used MADA (an SVM based morphological analyzer by Nizar Habash [5]) to select the correct sequence of analysis for each word. This step is important because choosing the wrong analysis results in wrong prefix, suffix segmentation. In MADA experiments we used the following splitting scheme.

S1: Decliticization by splitting off each conjunction clitic (w+, f+, b+, k+, l+), definite article (Al+) , pronominal clitics including possession pronoun (+p) and object pronoun (+O: ). Note that plural and subject pronouns are not splitted. S1 is summarized by (w+ f+ b+ k+ l+ Al+ REST +P: +O :). For example wlAwlAdh (‘and for his kids’) it would be (w+ l+ Awlad +h) according to S1.

5 Postprocessing

The generated translations need to be recomposed to match the non-splitted Arabic. This is done by the recombination model. To build a recombination model rules are extracted from both Training and Tuning sets. We observed the most frequent recombination patterns. In addition a recombination table is extracted from both Training and tuning sets.

The reason why recombination is not a simple process is that some letters are eliminated when we do splitting. For example lnkny is splitted to ln +y in this case we have two possible recombination lnk and lnkny. The advantage of splitting is sparseness reduction on the other hand the recombination is difficult because more than one possible word can be generated from a given stem affixes depending on the case ending. We can rely on a word based language model to choose the best recombined word, however this technique require a very strong language model that is built from a huge Arabic text to cover all case endings.

The recombination techniques have been addressed in a prior work [4]. We used the same recombination techniques which are recombination table extracted from training and tuning data and recombination rules. An example to recombination rules is when a suffix is attached to the end of a word which ends with Taa Marboota (P) the Taa Maroota is replaced by Taa Maftoooha (t).

6 Experiments

We carried out two main experiments. The first is the baseline experiment which does not involve morphology analysis. The second experiment involves using morphological analyzers MADA. We used the Arabic sentences in the training set to construct a 7-gram modified Kneser-Ney language model for the baseline and MADA experiments. We used SRI toolkit for language modeling [6]. Then GIZA++ [7] is used to obtain word alignment. MOSES scripts [8] are used then to extract the phrase table from the word aligned sentences we choose the maximum phrase length to be 8 words in case of the baseline experiment and 15 words in case of MADA experiment. MOSES scripts have been used to evaluate parameters together with the tuning set. Parameters are language model weight; phrase table weight and reordering table weight are tuned to achieve the highest bleu score over the tuning set. Bleu score is calculated after translating the test set using the tuned model.
We used an LDC parallel corpus catalog number LDC2004T18 and ISBN 1-58563-310-0. This corpus contains Arabic news stories and their English translations LDC collected via Ummah Press Service from January 2001 to September 2004. It totals 8,439 story pairs, 68,685 sentence pairs, around 2M Arabic words and 2.5M English words. The corpus is aligned at sentence level.

The Arabic sentences have been used to develop the language model. 2000 sentences pairs have been selected randomly for Tuning and another 2000 sentences pairs for Testing. The rest are left for training. Training data has been filtered to include sentences whose length is between 1 and 40 words for better alignment by GIZA++. 40,000 sentence pairs have been used for training.

6.1 Baseline Experiment
In this experiment we just used a simple tokenization for both Arabic and English. We applied the normalizations described in the preprocessing section. We repeated the experiment with and without using the numeric normalization. We repeated the experiment with and without person names clustering. We used Stanford English Named Entity Recognizer (NER) [9] to tag all person names in English text in the training set. Then we used Google translate to translate these names from English to Arabic. Finally all person names in both Arabic and English text are replaced by tag (PRN).

6.2 MADA Experiment
Training and tuning Arabic sentences are analyzed using MADA. Prefixes and suffixes are split. Prefixes are marked by a trailing plus sign and suffixes are marked by a beginning plus sign. So each word split into prefixes, stem and suffixes separated by spaces. After phrase table is constructed we removed all phrase table entries whose target phrase either starts with a suffix or ends with a prefix. We repeated this experiment with and without this post processing.
A set of recombination rules is extracted from the training data. A recombination table is extracted from the training data. Rules and the recombination table are tested on the testing set.

7 Discussion and Conclusion
A significant increase in Bleu score can be achieved by doing simple numeric and date normalization. This is because numbers increase sparseness and is considered as out of vocabulary. If we group all numbers in a single token (B) the language model quality increase as shown in table 1. In addition word alignment quality increases as a result a higher Bleu score is achieved.

In MADA experiment phrase table filtering increases the Bleu score because it forces the decoder to output compatible affixes/stems as a result a well formatted Arabic words are generated. Person names clustering in the baseline experiment decreases language model perplexity and improved the alignment quality. Person names are transliterated and they are infinite. They increase the number of vocabulary. Grouping these names in a single token (PRN) achieves 2 points in Bleu score.

Table 1 compares the Baseline experiments and the MADA-based experiments.

8 Conclusions and Future Work
It is obvious that clustering numbers and proper person names have significant effect on enhancing the SMT system. Splitting using MADA also introduces some enhancement through decreasing the perplexity of the Arabic language. However, MADA does not recognize person names as a result Named Entities like qrDAn get wrongly segmented to qrDa + y. This behavior introduces more ambiguity and negatively affects both alignment quality and language model.

We will repeat MADA experiment, but we should do person names clustering as a preprocessing step so that they would not be splitted by MADA. Also, it is our plan to cluster names that only occur less than a specific threshold and leave the more high frequency names intact.

Table (1) Bleu scores

<table>
<thead>
<tr>
<th></th>
<th>LM Perplexity</th>
<th>Bleu score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline with basic letters normalization and basic tokenization</td>
<td>303</td>
<td>19.1</td>
</tr>
<tr>
<td>Baseline (Numbers/Dates Normalization + basic letters normalization)</td>
<td>269</td>
<td>24.8</td>
</tr>
<tr>
<td>Baseline (Number/Dates Normalization + basic letters normalization) + person names clustering</td>
<td>136.2</td>
<td>26.5</td>
</tr>
<tr>
<td>MADA using S1 splitting scheme (Without phrase table filtering)</td>
<td>139.2</td>
<td>27.05</td>
</tr>
<tr>
<td>MADA using S1 splitting scheme (With phrase table filtering)</td>
<td>139.2</td>
<td>27.39</td>
</tr>
</tbody>
</table>
Acknowledgement
We would like to thank Dr. Hany Hassan, for providing the data and for his advices, Dr. Nizar Habash for providing MADA.

References: