Hybrid Arabic-English Machine Translation to Solve Reordering and Ambiguity Problems

Khalid Shaker Alubaidi
Department of Computer Science, Computer College, University of Anbar, Iraq e-mail: khalidalhity@gmail.com

***Abstract*— The problem in Arabic to English rule-based machine translation is that the rule-based lexical analyzer leaves some amount of ambiguity; therefore a statistical approach is used to resolve the ambiguity problem. Rule Based Machine Translation (RBMT) uses linguistic rule between two languages which is built manually by human in general, whereas SMT uses appearance statistic of word in parallel corpora. In this paper, those different approaches are combined into Arabic-English Hybrid Machine Translation (HMT) system to get the advantage from both kind of information. In the beginning, Arabic text will be inputted into RBMT to solve reordering problem. Then, the output will be edited by SMT to solve the ambiguity problem and generate the final translation of English text. SMT is capable to do this because on the training process, it uses RBMT's output (English) as source material and real translation (English) as target material. The results showed that the quality of translation in HMT system is better than SMT system.**

Keywords*— *Machine translation, Arabic-English machine translation, Hybrid Machine Translation

I. INTRODUCTION

There are many languages in the world. As a consequence, the document can be written in various languages. In order to get a better understanding, people will translate the document written in foreign language into their native language. Manual translation by looking at the dictionary will need a big effort. So, using machine translation (MT) is a recommended option to perform the automatic translation. This situation then raises a need to improve the performance of MT in order to get better translation result. Two common approaches in MT to perform translation task are rule based and statistic based. Rule based is the earliest approach in the MT subject [1].

RBMT (Rule Based Machine Translation) system is constructed based on linguistic rule between two languages. It concerns about the morphological process (analysis and generation) and transformation process (structural and lexical). It has power on explicit linguistic knowledge that it can deeply analyze in both syntax and semantic levels [1]. This approach has some weaknesses, such as: it requires much linguistic knowledge to create the linguistic rule so that it has high development cost; the accuracy of result is depend on the accuracy of each sub stage [2]; and the output is less fluency than SMT [3]. The latter case is makes sense because the translation is generated strictly based on rule and the translation word is fixed for all input cases. There are many MT systems use this approach as their translation method, such as: MT for Romance Languages to Spain [4], Bulgarian-Macedonian [5], Indonesian-Malaysian [6], English to Sankrit [2], etc. Interested readers are referred to [13] for An Introduction to Machine Translation and [14] for a comprehensive survey of the Arabic to English machine translation in recent years.

On the other hand, SMT (Statistic Machine Translation) system is constructed based on parallel corpora. It performs training process on them to learn implicit knowledge that is contained in co-occurrence statistic. System will find translation for certain word in source language by looking the word in target language that is often occurs together with them in parallel corpora. The advantages of this approach are: it may be able to produce suitable translations in case input sentence is not similar to any sentences in a training corpus [1]; the output is more natural and fluency [3] and it is much easier to be built than RBMT system. However, because the system relies on information that is learned in training process, so the output faces a problem on unstructured syntax and grammatical mistake [3] and it is less literal. Some researches try to improve this approach, such as by using word sense disambiguation [7], using grammatical categories and word categorization to handle the error [8], etc. Based on the strengths and weaknesses of each approach, many researchers try to combine them by making the concept of hybrid machine translation (HMT) system in order to improve the performance of MT.

Simard et al. [9] used SMT system as a layer to perform post-editing toward the output of RBMT system. SMT system will correct and adjust the translation output of RBMT system based on the most common translation that occurs in the parallel corpora. Dugas et al. [10] do the same method in [9] but they perform additional experiment by using SYSTRAN+ Moses[1]. Another approach to create HMT system is by incorporating the phrase table of Moses [15] with phrase table that is generated from alignment of source text and its translation output from RBMT [11]. Moses decoder then will choose the best combination of phrases. The result shows that hybrid system has better performance than baseline SMT system. Eiseleet al. [12] using same mechanism with [11] but they make some additional language pair on their experiment. This paper will describe the process of developing Arabic-English HMT system as a way to improve the performance of MT. Yulianti et al. [18] developed a Hybrid Machine Translation System for Indonesian-English language pair

[1] http://www.statmt.org/moses/

by utilizing SMT system as editing component of RBMT system's output. We use the hybrid approach by combine RBMT with SMT as editing component toward the output of RBMT. The remaining of this paper is organized as follows: Section 2 will describe the architecture of Arabic-English hybrid machine translation system; Section 3 will describe the implementation of HMT; Section 4 will present the experiment result together with its analysis; and Section 5 will give a conclusion about this research..

II. ARABIC-ENGLISH HYBRID MACHINE TRANSLATION SYSTEM

Arabic-English HMT system presented in this paper consists of RBMT system and SMT system that works sequentially. SMT system is utilized as editing component of RBMT system's output. Initially, Arabic text is inputted into RBMT system. Then the output will be edited by SMT system to generate the final translation of English text. SMT is capable to perform editing process because on the training process [21], it uses RBMT's output (English language) as source material and the real translation (English language) as target material. Because the source material and target material is actually in the target language, this process can be seen as target-to-target training. It is rather different with common training process of baseline SMT system that uses concept of source-to-target training.

A. RBMT System

Arabic-to-English RBMT system is ready to use. This research uses our RBMT (AE-TBMT) which developed in previous work [17]. Basically, the translation process of AE-TBMT consists of six main phases: 1). Text in the source language is transferred to tokenizer is to divide the text into tokens. 2). Then Start morphological analysis to provide morpho-syntactic information. 3). The syntactic parser builds a syntactic relevant tree, which represents relationships between the words of the phrase. 4).Lexical transfer will map Arabic lexical elements to their English equivalent. It will also map Arabic morphological features to the corresponding set of English features. 5).Structure transfer will map the Arabic dependency tree to the equivalent English syntactic structure, and 6). Finally Arabic synthesiser will synthesis the inflected English word-form based on the morphological features and traverses the syntactic tree to produce the surface English phrase.

1. Tokenization:

This an important step for a syntactic parser to construct a phrase structure tree from syntactic units. After inserting the source sentence in the system the tokenizer divides the text into tokens. The token can be a word, a part of a word, or a punctuation mark. A tokenizer requests to know the white spaces and punctuation marks.

2. Morphological analysis

After the tokenization process the morphological analyser will provide the morphological information about words. It provides the grammatical class of the words (parts of speech) and create the Arabic word in its right form depending on the morphological features.

3. Lexicon:

In this system the lexicon is accountable for inferring morphological and classifying verbs, nouns, adverb and adjectives when needed. It is the main lexicon translation; the source language searches in a dictionary and then chooses the translation. A lexicon provides the specific details about every individual lexical entry (i.e. word or phrase) in the vocabulary of the language concerned. Lexicon contains grammatical information which is usually have abbreviated form: 'n' for noun, 'v' for verb, 'pron' for pronoun, 'det' for determiner, 'prep' for preposition,'adj' for adjective, 'adv' for adverb, and 'conj' for conjunction. The lexicon must contain information about all the different words that can be used. If the word is ambiguous, it will be described by multiple entries in the lexicon, one for each different use.

4. Parsing:

The parser divides the sentence into smaller sets depending on their syntactic functions in the sentence. There are four types of phrases i.e. Verb Phrase (VP), Noun Phrase (NP), Adjective/Adverbial Phrase (AP), and Prepositional Phrase (PP). After the parsing process the sentence is represented in a phrase structure tree. Fig. 1 show the phrase structure tree for the sentence الطالب الذكي قرا الكتاب (the cleaver student reads the book).

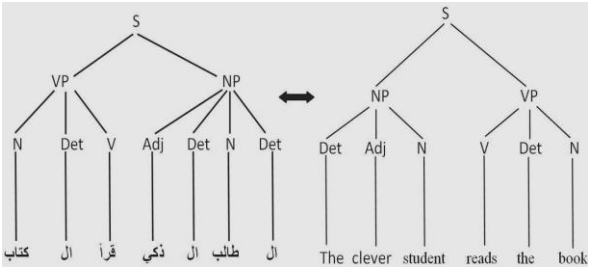


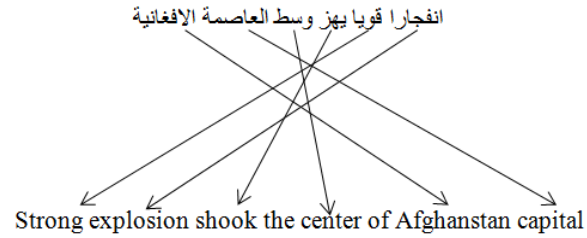
Fig. 1. Phrase structure tree

5. Syntactic rules:

A set of Arabic and English rules are fed into the system. In this step the reordering process will be found which will be based on the order of words in a sentence, and how the words are grouped.

6. Agreement rules:

After syntactic rules the agreement rules applied which are responsible about the additions of prefix and suffix in the sentences.



S: [NP VP] == S: [NP:[N Adj] VP:[V N1 Det N2 Adj]] Arabic rule

S: [NP VP] == S: [NP: [Adj N] VP:[V Det N1 Adj N2]] English rule

The architecture of this HMT system is illustrated on Fig.2.

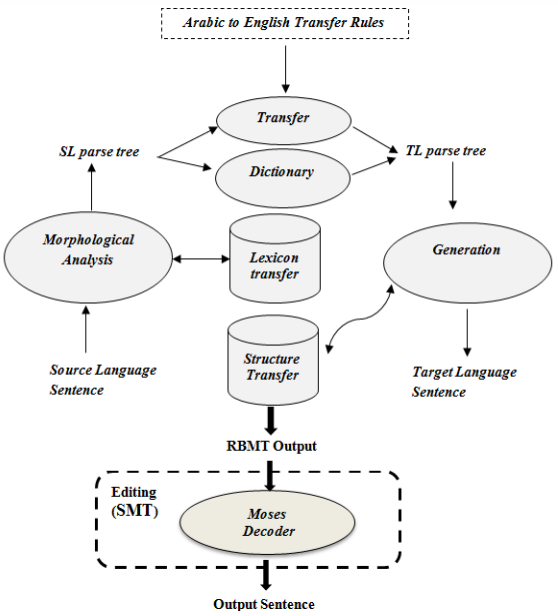


Fig. 2. Architecture of Arabic-English HMT System

B. *SMT System*

In this hybrid mechanism, we used Moses as our SMT system. Moses is an open source toolkit for machine translation that provides tools for training, tuning, and applying translation [15]. It is widely used in the research on machine translation area, such as in [10][11][12]. Having phrase-based translation paradigm, Moses decoder uses phrase table as main sources to find the phrase translation. It will choose the sentence with the highest score as a translated sentence [16]. Score is achieved from weighted-log probability that is product of 2 components: phrase translation model and language model.

III. IMPLEMENTATION

After the RBMT system for Arabic-English language pair is built, the next step we need to do is training process. We perform two kinds of experiment: SMT Experiment and HMT Experiment. We will compare the performance of HMT toward the baseline SMT. We use Moses toolkit with 3 gram language model in the experiments. The statistical machine translation system is trained using word alignments of parallel corpora of Arabic- English that obtained from Computational Linguistics Laboratory[2]. Totally, there are 1181 parallel sentences that we used in both of experiments. We use 10% of total data for testing corpus (120 sentences) and we use the rest for training corpus.

In the training process of HMT system, Moses decoder applies the concept of target-to-target training in order to be able to perform editing process toward RBMT output. It uses the output of RBMT as source material and the real translation as target material. So, it will learn the mapping of RBMT translation into real translation. In case that RBMT translation is not common or does not exist, SMT will learn them on training process (as discussed in [15]) so that it will correct them on HMT system.

IV. EXPERIMENTAL RESULT

We evaluated our system using BLEU. Experiment toward baseline SMT system is also performed in order to evaluate the performance of HMT system. Bleu score for both of the system is calculated and it is described on Table 1. The table shows the values of BLEU obtained for phrase length: 1-gram, 2-gram, 3-gram, and 4-gram, respectively. Note that BLEU is in between 0 and 1 ($0 \leq$ Bleu<1). When BLEU value is close to 1, that mean the quality of translation is better and close to the manual translation. In this evaluation 1 candidate file (represent our system translation) and 2 references files (represent 2 different manual translation) have been used. It can be clearly seen that score of HMT system is higher than SMT system in all cases. When combining two approaches, HMT outperformed SMT in Bleu score by 1.19% with 1-gram, 2.01 with 2-gram, 4.27% with 3-gram.

We believe that a good translation could be achieved when combine RBMT with SMT as RBMT solves word ordering problem when translate from Arabic to English, however SMT solves the ambiguity problem.

After doing analysis toward the output of RBMT, we found that comprehensive reordering rules play an important role in the quality of translation. In addition, more data training makes the output of SMT more accurate.

TABLE I BLUE EVALUATION RESULTS OF HMT

Phrase length *n-gram*	HMT	SMT
1-gram	0.910932	0.89897
2-gram	0.802346	0.79113
3-gram	0.669451	0.64675
4-gram	0.57543	0.54642

[2] http://www.lllf.uam.es/ESP/

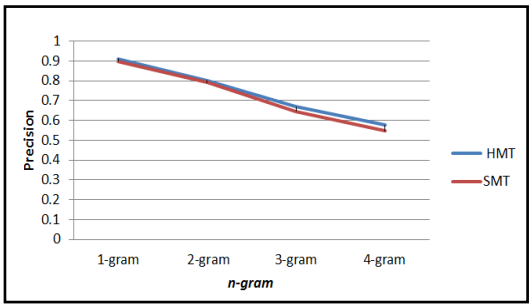


Fig. 3 illustrates how HMT system translation is closer than SMT system translation to manual translation with phrase length: 1-gram, 2-gram, 3-gram, and 4-gramr, respectively.

Fig. 3. Score of HMT with 1-gram, 2-gram, 3-gram, and 4-gram

V. CONCLUSION

In this paper, we have described an approach to develop HMT System for Arabic- English language pair by utilizing SMT system as editing component of RBMT system's output. The motivation behind this research is combining the advantage of information that is contained in each of the MT system to get better translation result. Evaluation by using Bleu score indicator shows that: 1). The size of the training data effects the statistic model on SMT and HMT system, so adding more training corpus can improve the performance HMT system. 2). HMT system outperforms SMT system in all cases. We analyze that hybrid solutions combine the advantages of the individual approaches to achieve an overall better translation. The approach is most useful to address one of Rule-Based MT greatest challenges – translation ambiguity. When a word/phrase can have more than one meaning, statistics can help identify the most suitable option.

References

[1] Charoenpornsawat, P., Sornlertlamvanich, V., Charoenporn, T.: Improving Translation Quality of Rule-based Machine Translation. In: Proceedings of COLING Workshop on Machine Translation in Asia, pp. 351-356, Taiwan (2002).

[2] Barkade, V. M., Devale, P. R.: English to Sankrit Machine Translation Semantic Mapper. In: International Journal of Engineering Science and Technology Vol. 2 Issue 10, pp. 5313-5318 (2010).

[3] Carrera, J., Beregovaya, O., Yanishevsky, A.: Machine Translation for Cross-Language Social Media. http://www.promt.com/company/technology/pdf/machine_translation_for_cross_language_social_media.pdf (2009).

[4] Corbi-Bellot, A. M., Forcada, M. L., Ortiz-Rojas, S., Perez-Ortiz, J. A., Ramirez-Sanchez, G., Sanchez- Martinez, F., Alegria, I., Mayor, A., Sarasola, K.: An Open-Source Shallow-Transfer Machine Translation Engine for the Romance Languages of Spain. In: Proceedings of the Tenth Conference of the European Association for Machine Translation, pp. 79-86, (2005).

[5] Rangelov, T.: Rule-based Machine Translation between Bulgarian and Macedonian. In: Proceedings of the Second International Workshop on Free/Open-Source Rule-Based Machine Translation, pp. 53-59, Barcelona (2011).

[6] Larasati, S. D., Kuboň, V.: A Study of Indonesian-to- Malaysian MT System. In: Proceedings of the 4th International MALINDO Workshop, Jakarta (2010).

[7] Carpuat, M., Wu, D.: Improving Statistical Machine Translation using Word Sense Disambiguation. In: Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural

Language Learning (EMNLP-CoNLL), pp. 61-72 (2005).

[8] Farrús, M., Mariño, J. B., Poch, M., Hernández, A., Henríquez, C., Fonollosa, J. A. R., Costa-Jussà, M. R.: Overcoming Statistical Machine Translation Limitations: Error Analysis and Proposed Solutions for the Catalan---Spanish Language Pair. In: Journal Language Resources and Evaluation, Vol. 45 Issue 2, pp. 181-208 (2011).

[9] Simard, M., Ueffing, N., Isabelle, P., Kuhn, R.: Rulebased Translation with Statistical Phrase-based Postediting. In: Proceedings of the Second Workshop on Statistical Machine Translation, pp. 203-206, Prague (2007).

[10] Dugast, L., Snellart, J., Koehn, P.: Statistical Post- Editing on SYSTRAN's Rule-Based Translation System. In: Proceedings of the Second Workshop on Statistical Machine Translation, pp. 220-223, Prague (2007.)

[11] Chen, Y., Eisele, A., Federman, C., Hasler, E., Jellinghaus M., Theison, S.: Multi-Engine Machine Translation with an Open-Source Decoder for Statistical Machine Translation. In: Proceedings of the Second Workshop on Statistical Machine Translation, *pp. 193-196, Prague (2007).*

[12] Eisele, A., Federman, C., Saint-Amand, H., Jellinghaus, M., Hermann, T., Chen, Y.: Using Moses to Integrate Multiple Rule-Based Machine Translation Engines into a Hybrid System. In: Proceedings of the Third Workshop on Statistical Machine Translation, pp. 179- 182, Colombus (2008).

[13] Hutchins, W. J., Somers, H. L.: An Introduction to Machine Translation Vol. 362. Academic Press. New York (1992).

[14] Alqudsi A, Omar N., and Shaker K. 2012, "Arabic Machine Translation: a Survey", Artificial Intelligence Review (July 2012), pp.1-24.

[15] Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., Herbst, E.: Moses: Open Source Toolkit for Statistical Machine Translation. In: Proceedings of the ACL 2007 Demo and Poster Sessions, pp. 177-180, Prague (2007).

[16] Koehn, P.: Moses Statistical Machine Translation System: User Manual and Code Guide. http://www.statmt.org (2010).

[17] Hatem A, Omar N (2010) Syntactic reordering for Arabic-English phrase-based machine translation. In: Database theory and application, bio-science and bio-technology. Springer Lecture Notes in Computer Science, vol 118. Verlag, Berlin, pp 198–206.

[18] Yulianti, M. Adriani, H. M. Manurung, I. Budi, and A. N. Hidayanto, "Developing Indonesian-English Hybrid Machine Translation System," in Proc. International Conference on Advanced Computer Science and Information System, 2011, pp. 265-270.