Bilingual Language Model for English Arabic Technical Translation

Marwa N. Refaie[*1], Ibrahim F. Imam[**2,] Ibrahim F. Eissa[**3]
[*]*Faculty of Computer Science, Modern University for Technology & Information*
Cairo, Egypt
[1]basmallah@hotmail.com
[**]*Department of Computer Science, Arab Academy for Science & Technology*
Cairo, Egypt
[2]ifi05@yahoo.com
****Department of Computer Science, Cairo University*
Cairo, Egypt
[3]i.farag@fci-cu.edu.eg

Abstract: The massive fast of new scientific publications increase the need to a reliable effective automatic machine translation (AMT) system, which translates from English, as the common language of publications, to other different languages. Statistical machine translation (SMT) model crafted to deal with certain domain of text often fails when subjected to another domain. The paper addresses the characterization of language domains and their behavior in SMT, experiments the management of SMT model to translate scientific text collected from artificial intelligence publications. The effectiveness of Bilingual language model is tested against the typical N-gram language model, in addition to utilizing the fill-up and back-off techniques to handle different phrase tables from different domains. As not every human capable to translate artificial intelligence book, should have strong knowledge in the field, We suggest that in order AMT can handle different domains it must be trained by in-domain parallel data, adjusting weights for the words on different domains to learn the model how to differentiate between different meaning of same word in different domains.

Keywords: Statistical machine translation, Domain adaptation, Bilingual Model, Fill-up phrase table.

1 INTRODUCTION

Our knowledge of science was built up over thousands of years. People of many cultures and civilizations have contributed to what we know today. Translation of scientific text is very important in transferring knowledge from one nation to another. Translating Arabic and Greek heritage to Europe on the 11thand 12th century led to the renaissance and scientific revolution on Western Europe. Now, it turns to Arabs to pass away others contributions to learn from and add to it. With the high speed of published papers and books, full depending on translators will be impractical; we need efficient AMT system for faster translation. Machine translation still not commonly used in technical or scientific documents translation. As there are specific words and terminologies may have totally different translation in different domains and function words surrounded, even some words have no correspondence in target language.

Statistical machine translation (SMT) is the state-of–the-art approach to machine translation of large vocabulary tasks. Since the approach was first presented in [1] and has been used in many translation systems since then. One drawback of this approach is sensitivity to the domain of the trained data, and that large amounts of training data are needed. SMT consists of two steps, first calculating probabilistic translation models, which are generally trained using sentence aligned parallel corpora for certain language pair. The second SMT step, calculating language models, that is responsible of reordering, which refers to the order of the translated word to generate good translation.

As the performance of the SMT system improves if this data is alike in topic and field, some researches proposed to use only in-domain data to build training model. METEO system [2] restricted the text type to weather data for French-English bilingual pair. As it was expected, the system is able to fully automatically produce high-quality translated output. Reference [3] claimed achieving BLEU score up to 56 translating movies subtitles as Swedish-Danish pair, using big corpus made up of only aligned movie subtitles for training. The challenge for SMT as studied in [4] is to create a system that can handle different domains, proposed the usage a log-linear combination of the in-domain and out-domain phrase table. Other used approach to handle domain adaptation is to use all available data to train a general system and to adapt the system to be trained on in-domain data for building language models [5].

Language models are used to automatically learn target words order patterns from data. Word orders can differ significantly across languages. Most of the used language model was built depending on the target language only. Given n-gram frequencies, then building a classical model that predicts the likelihood of a sequence of words given their preceding word [6]. Recently Bilingual model was introduced to build more efficient language models ([7], [8]). Claiming that most usually used n-gram model to extricate reordering knowledge built utilizing a target language only, so this does not take in consideration the translation correspondence but just models target-language fluency. Many researches attested that language models utilizing only monolingual corpus are not enough to have good translation [9].

This paper analyzing different approaches to overcome domain adaptation challenge, translating scientific text from English to Arabic, based on SMT steps translation and language models. We utilize a general English-Arabic corpus besides out-domain scientific corpus. The scientific data is collected from scientific publications and translated using Google, then Google translations were edited by a domain expert. First we examined the influence of different language models techniques, N-gram and Bilingual approaches, Secondly; we analyzed two different techniques to adapt the phrase pair scoring of different domains phrase tables. By studying the steps distinctly, we are able to combine the techniques from the experimented approaches in a new way and improve the translation quality.

2 RELATED WORK

Linguistics has mostly used the term sublanguages for specific types of texts within one language, while in the MT area, the term language domain is more common. Since the 90’s data driven approaches headed the research on AMT. One of the weaknesses is handling different text sublanguages to be translated; words in source language could have completely different meaning in different domains. For example the word “*deduction*” in general domain translated to “*خصم*”, while in scientific domain should translated to “أشتقاق”, “*Induction*” in general is “*تقليد*”, in science is “*استقراء*”, in medical “*تحريض*”. Human translators must have strong background knowledge about the domain field as much as knowledge in linguistics and vocabulary of the pair of languages, and as well machines.

A. Language Models

Language model is one of the most important modules in statistical machine translation. LM normally includes a target-language model, which directs a translation decoder about the perfection of a given sentence and the fluency of the translation hypothesis.

1) Probabilistic Language Model:

It is usually formulated as a probability distribution p(s) over strings s that attempts to reflect how frequently a string S occurs as a sentence. Statistical language modeling is concerned by estimating the joint probability of a word sequence. P (W1, W2, W3, ….Wn). This is always converted into conditional probability: P (Next Word | History), as in (1):

$$\boldsymbol{\mathcal{P}(s) = \prod_{i=1}^{l} \mathcal{P}(w_i|w_1 \dots w_{i-1})} \tag{1}$$

For example, trigram model, probabilities looks like: P (الذكاء ,علم|الأصطناعي). One of the main problems in n-gram LM, even with large training corpora, there is many valid word sequence can have small or zero probabilities. Therefore, smoothing Kneser-Ney discounting technique [6] applied to n-grams to reassign probability value producing better prediction for unseen words sequence. Reference [12] shown that pre-ordering source language to follow to target language word order significantly improves translation quality. They applied the experiment translating from German to English, sentences are reordered in the train and test data. Results showed improvement from 25.2% BLEU score to 26.8% BLEU score.

2) Neural Probabilistic Language Model:

Research on language modeling for machine translation has increasingly focused on the application of neural network in recent years; the use of Neural Network LM has shown significant improvements over the traditional n gram models. Even with smoothing in n-gram based LM, the discrete n-gram language models still can’t handle words similarity, as words are treated as discrete variable. In contrast, the neural network language model set words in a continuous space in which probability estimation is performed using hidden layer neural network. The motivation based on that with accurate training of the word embedding, similar semantically or grammatically words will be mapped to similar points in the continuous space. And so, the NNLM can achieve better generalization for unseen words sequences. A main disadvantage of the NNLM is its very high computational cost during training. While traditional n-gram LMs can be trained in a few minutes using the SRILM or kenLM toolkit, it can take some hours to estimate a continuous space LM for a large vocabulary task.

The NNLM architecture proposed in [15], a feed-forward neural network with a single hidden layer was used to calculate the language model probabilities. The experiments done on two corpora, one with more than a million examples, and a larger one with up to 15 million words, proved that using NN get better results by 10% to 20% than using N-gram language model. Reference [18] compared results using different techniques translating English to Hindi. First they pre-order English sentence to follow Hindi sentence structure, second the paper compared neural network based language model with different morphological features and finally, they explored the use of lexical WordNet to overcome the effect of Out-of-Vocabulary (OOV) words on MT quality. Results showed improvement of BLEU score by 6 points using three types of features for building NNLM for Hindi (lemma POS, and NC), over Typical N-gram LM. Recurrent neural network architecture proposed in [14], which allow the model to use arbitrarily long histories. They claimed significant gains after scaling up RNNLM in term of data and model sizes, speech recognition experiments show 18% reduction of word error rate.

3) Bilingual Language Model:

Recently, the use of Bilingual LM has shown significant improvements over the monolingual target language based LM. In the bilingual LM, both the aligned target and source words build the tokens of the n-gram model. For example, when calculating the bilingual language model probability for the word *الذكاء الاصطناعي*, BiLM will take inconsideration the P(الأصطناعي_ Artificial | الذكاء _ Intelligence), as it is spectacle that through bilingual model tokens not only consider the previous word but also the previous source word. Bilingual n-gram LM approach presented in [7] using units of source and target words called tuples. Experiments reported improvement in the translation quality of both Spanish-English and English-Spanish tasks.

Different approach than SMT proposed in [16], using stochastic finite state transducer based on bilingual n-gram model. The experiment showed translation improvement on French-English translation task. The translation model is implemented as n-gram model over the tuples, state a probabilistic relationship between sequences of source and target words, defining the similarity to phrase pairs. Even though this model based on phrase-model approach, it differs from the well-known phrase-based approach in two basic points. First training data is distinctively segmented into bilingual units (tuples). And second, the model considers n-gram probabilities instead of relative frequencies. Based on the good results of recurrent neural network in machine translation, the authors in [17] investigate the utilization of the RNN training on bilingual word aligned sentences. They used English-Iraqi (Arabic) corpus, first aligned by GIZA, then the aligned sentence pairs sorted with the sequential order were fed into bilingual RNN training, while the Arabic text was fed into RNNLM training. Results of translation on different test sets proved that bRNN gain better BLEU score over the base line model and the RNNLM by0.9. They detected better results to rare unigrams words while decoding, they explained it as better learning of the context of English-Arabic pairs gain over the bilingual model training.

Utilizing BiLM and part-of-speech (POS) examined on[7], extending the Bilingual n-gram approach to handle word factors. They designated other variant of the original BiLM, using POS tags instead of the words itself. For Arabic-English translation task, an extra bilingual LM on the POS tags instead of the surface word forms was formed led to translation improvements. In this work different language pairs analyzed, with different n-gram and bilingual LM, in addition to exploring results adding POS word factor. Most of the recorded result proves translation improvement when combining POS and BiLM. Based on [8] experiments using lexicalized BiLM, in [19]they proposed adding syntactic information to the BiLM tokens. They claimed that the existing bilingual n-gram models are not enough to differentiate between alternative reordering's. The idea is to build syntactic representation of a translation during decoding by adding fragments from the source parse tree. They proposed to characterize contextual syntactic roles of a word in terms of POS tags of the words themselves and their relatives in a dependency tree. Creating BiLM tokens sequence t1 …tn as (2):

$$\boldsymbol{t_i} = \langle \boldsymbol{ContE}(\boldsymbol{e_i}),\{\boldsymbol{ContF}(\boldsymbol{f})|\boldsymbol{f\epsilon\, A}(\boldsymbol{e_i})\}\rangle \qquad \textbf{(2)}$$

Where ei is the i-th target word, A: E → P(F) is an alignment function, F and E are source and target sentences, and ContE and ContF are target and source contextual functions. Different experiments sets were applied to Arabic-English and Chinese-English translation task. Results stated that utilizing source information yields to translation quality improvements.

B. *Machine Adaptation*

Machine translation systems are often built for very specific domains, such as movie and television subtitles [3], or Weather new [2]. A translation model may be trained only from sentences in a parallel corpus that are similar to the sentences to translated, obtaining language model by including only sentences that are similar to the ones in the target domain [10]. A lot of researches proposed methods to combine different domains, training by set of domains corpus, utilizing mixture models approaches to score weights among number of phrase tables. Text classification methods employed to identify different text domains, such as discriminant analysis [11], by discovering certain features to be learned in each text domains they can decide a domain for new text. The selected features based on certain word counts, and the frequency of certain words, results show that the text field identification task gets more difficult as the number of categories increase.

In rule-based system, there is set of rules especially crafted to handle language domains, whereas the statistical MT approaches using parallel training data dedicated for such a language domain. Domain adaptation techniques in SMT try to fully utilize the given set of data from source and target language pairs in different domains, combining knowledge from all available data to create a machine translation application capable of handling different source language text in different fields. Some proposed adopting only the language model, inspired by approaches in speech recognition [20], the main advantage is that only monolingual, target language, in-domain needed. Another approach proposed calculation vector similarity and adds it into the phrase table and use it as a tuning and decoding time features [21].The similarity is computed by comparing the vectorized representation of phrase pairs extracted from the development set and the training set.

By the claim that language model can’t arrest connection between source and target languages , researches proved the importance and big sensitivity of SMT to the availability of parallel data in different sublanguages, as the translation model can be adopted as well ([4], [5]). Used methods for training translation model conjoining knowledge from different domains may be applied at corpus level or phrase table level. At corpus level generative models and maximum likelihood were used, where adaptation at phrase table level is either off-line, by a linear mixture of weights, or at decoding step through a log linear combination.One of the techniques works on the corpus level, filtering down data to the parts that are more similar to the in-domain data [22].Claiming that bigger data not always offer better performance, and that the selected 1% of the corpus achieves better results. Other method depends on weights mixture model on text distances between in-domain data and mixture data set proposed in [23]. Different techniques explored in this paper, log-linear mixtures, dynamic adaptation, different text metrics to map weights, estimating set of features, two probabilities for the phrase table and one probability for each language model. Stated results show enhancement by the linear and log-linear mixtures over a baseline trained on the union of all training data, besides a set of held experiments using bilingual models results assuring the importance of considering both source and target languages.

Phrase Sense Disambiguation (PSD) is one of the used approaches to handle domain adaptation [24], PSD is a discriminative translation model, which scores translation candidates for a source sentence using source context, dissimilar to phrase table translation probabilities which are independent of the context.PSD concerns translation as a classification task, in decoding time, the PSD classifier uses the context to predict the correct translation of a source sentence in the target language. At training time, PSD uses word alignment to extract training features, same as in a standard phrase-based SMT system. However, the extracted training features are not just phrase pairs, but with an adding feature representing the source phrases context. In [25] they framed sense induction and disambiguation based on topic models, as learning topic distributions for a word type, while disambiguation consists of assigning topics to word tokens. This model can be used to detect newly gained senses for a word over an old domain. In a recent study they implement a system for sense spotting approach for SMT to spot tokens that have new senses in new domain [26].They used both general-domain sense dictionary, (French-English) and new-domain monolingual French text (medical, scientific & movies subtitles) to calculate two features: word-type and word-token. The model can determine which words demand a new translation in which domain, besides signifying which words need a new translation probability distribution when ensued to a new domain.

C. *Translation Model Combination*

In this technique it is required to adapt the translation model to capture new domain knowledge, without losing learned information from other domains or sublanguages. Domain adaptation for SMT can be implemented through translation model if parallel in-domain and out-domain data are available, phrase table level, many approaches used to combine several phrase tables by linear interpolation or instance weighting using fill-up.

1) Phrase Table Fill-up:

The original method of fill-up was conceived in [27], to train speaker system and proved to outperform classical linear interpolation handling the problem of language model adaptation for speech recognition. And then was recently presented in SMT [28]. In this work, the phrase tables calculated from in-domain (news) and out-domain (Europarl) corpus are combined by keeping all the phrase pairs unchanged from the in-domain phrase table, and only adding in the phrase pairs from the out-domain phrase tables that are not occurred at the in-domain phrase table, as in (3):

$$\boldsymbol{F\,ill - up\{PT\} = \{PT_{\_in}\} \cup \{PT_{\_out} - PT_{\_in}\}} \tag{3}$$

Where PT$_{\_in}$ and PT$_{\_out}$ are the in-domain and out-domain phrase table. Results off all experiments were improved by combining in-domain and out-of-domain data phrase tables.

Another approach using the Fill-Up technique was described in [29]. They used the in-domain and out-domain scores and an indicator feature, the out-domain scores were only used if no in-domain probabilities were available for a sentence in the phrase table. They also extend the fill-up approach into the SMT reordering model, proposing as well study for pruning options. The experiments show that the fill-up approach is able to produce better translation and increases the efficiency of minimum error rate. Probabilities calculated such that T$_1$ and T$_2$ are the in-domain out-domain phrase tables, the translation model assigns a feature vector to each phrase pair φ ($\tilde{f}$, $\tilde{e}$), where $\tilde{f}$and$\tilde{e}$ are the source and target phrases.

$$\boldsymbol{\varphi(\tilde{f}\ ,\tilde{e}) = (P_{ph}(\tilde{e}|\tilde{f}), P_{ph}(\tilde{f}|\tilde{e}), P_{lex}(\tilde{e}|\tilde{f}), P_{lex}(\tilde{f}|\tilde{e}), pp(\tilde{f}|\tilde{e}))} \tag{4}$$

Where P$_{ph}$ refers to the phrase translation probability, P$_{lex}$ is the lexical weighting probability, and pp is a constant, so the fill-up model T$_F$ is defined as in (5):

$$\forall(\boldsymbol{\tilde{f}}|\boldsymbol{\tilde{e}}) \in \boldsymbol{T_1} \cup \boldsymbol{T_2}:$$

$$\emptyset_F(\boldsymbol{\tilde{f}}|\boldsymbol{\tilde{e}}) = \begin{cases} (\boldsymbol{\emptyset_1}\left(\boldsymbol{\tilde{f}}|\boldsymbol{\tilde{e}}), \boldsymbol{exp(0)}\right) & \boldsymbol{if}\ (\boldsymbol{\tilde{f}}|\boldsymbol{\tilde{e}}) \in \boldsymbol{T_1} \\ (\boldsymbol{\emptyset_2}\left(\boldsymbol{\tilde{f}}|\boldsymbol{\tilde{e}}), \boldsymbol{exp(1)}\right) & \boldsymbol{Otherwise} \end{cases} \tag{5}$$

Reference [30] utilized support vector machine for estimating probabilistic feature, combining phrase tables from in-domain and general corpus for English-French language pair. They used the fill-up approach considering their calculated probabilistic feature instead of the original binary feature. They claim improvements of BLEU score up to 0.8 point using their proposed probabilistic feature fill-up approach

2) Linear Interpolation:

Linear interpolation is another approach for phrase table combination, based on computing the weighted average of multiple phrase tables and combining it on one probability model [5]. The new calculated probability model calculated as in 6:

$$\boldsymbol{P(x|y;\lambda) = \sum_{i=1}^{n} \lambda_i p_i(x|y)} \tag{6}$$

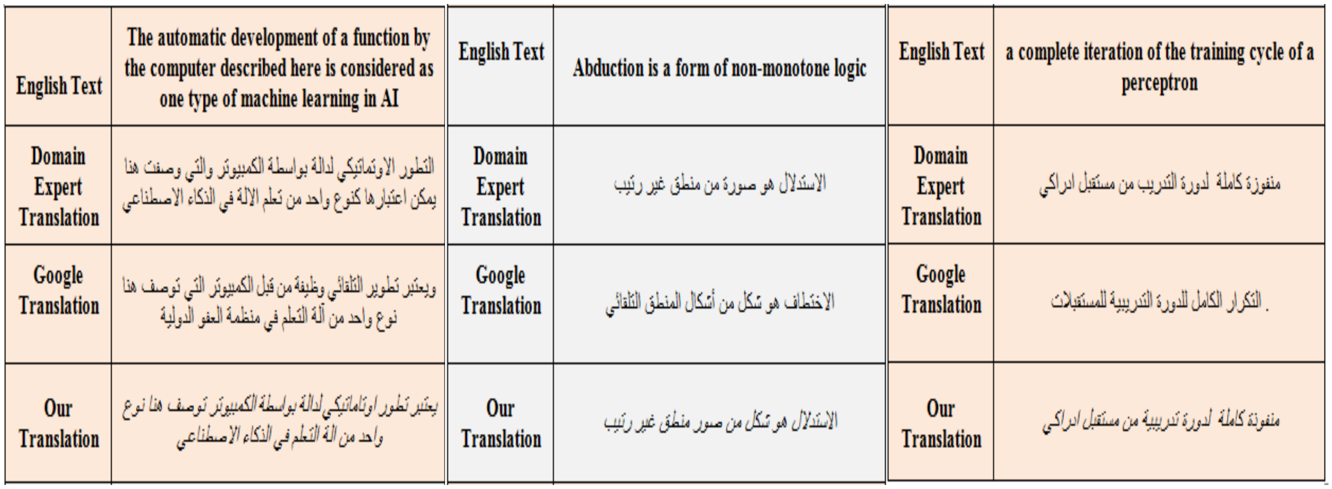
Where λ_i is the interpolation weight of each model I, and ($\sum_i \lambda_i$) =1. The interpolation technique has been implemented in many different systems according to how they set the interpolation weight, such as considering uniform weights or to set different coefficients. In [5] the author introduced set of comparative results using in-domain LM and interpolation combination technique with different pairs of languages French-German, Haiti Creole-English. Results shown that depending on pure in-domain LM, or even using small part of it, improve the BLEU score, as well as using more training data. In addition to another set of experiments, for domain adaptation proved that the modified calculated interpolation weights, depending on perplexity minimization scales, led to better translation.

Resent empirical study [31] targeting grammatical error correction over SMT, reported results show increasing of BLEU score by more than 32 points combining different approaches. They add different linguistic knowledge to the parallel corpus, such as lemma, part of speech (pos) suffix, and prefix, in addition to combination of TM of phrase-based and factor-based using linear interpolation. Results verify the efficiency of combining different factored-based and phrase-based TM, besides proving the efficiency of adding the pos as a factor, that the model trained with this factor outperforms the others.

3 Experiment

There are several English corpora, in different sublanguages which have been created for the purpose of research of English linguistics since the 1960s. Certainly, these efforts led to the advance of different fields of English linguistics and especially machine translation. In opposite to Arabic language, we suffer from lake of corpora on different domains and enough size for empirical study.

Our experiments applied to scientific text from English into Arabic. The data is collected from scientific publications and books, especially in Artificial Intelligence domain. English text translated using Google, then Google translations were edited by a domain expert. Google translated text proved that SMT perform poorly when applied to a new domain. Many words and scientific terminologies were missedtranslated.Figure1 shows an English sentences and its Arabic translation by Google, domain expert translations and our model translation output. These examples show that there is a problem with the training data, this problem is either due to missing data or inappropriate statistical distribution of the training data. It is clear that there is a significant difference between translations.



English Text	The automatic development of a function by the computer described here is considered as one type of machine learning in AI	English Text	Abduction is a form of non-monotone logic	English Text	a complete iteration of the training cycle of a perceptron
Domain Expert Translation	التطور الاوتماتيكي لدالة بواسطة الكمبيوتر والتي وصفت هنا يمكن اعتبارها كنوع واحد من تعلم الالة في الذكاء الاصطناعي	Domain Expert Translation	الاستدلال هو صورة من منطق غير رتيب	Domain Expert Translation	منفوزة كاملة لدورة التدريب من مستقبل ادراكي
Google Translation	ويعتبر تطوير التلقائي وظيفة من قبل الكمبيوتر التي توصف هنا نوع واحد من آلة التعلم في منظمة العفو الدولية	Google Translation	الاختطاف هو شكل من أشكال المنطق التلقائي	Google Translation	التكرار الكامل للدورة التدريبية للمستقبلات .
Our Translation	يعتبر تطور اوتاماتيكي لدالة بواسطة الكمبيوتر توصف هنا نوع واحد من الة التعلم في الذكاء الاصطناعي	Our Translation	الاستدلال هو شكل من صور منطق غير رتيب	Our Translation	منفوزة كاملة لدورة تدريبية من مستقبل ادراكي

Figure 1: Scientific text Human, Google, & our model translation examples

Our corpus consists of in-domain and general-domain parallel corpus. The in-domain scientific text was prpared by us, around 1700 K parallel sentence pairs as training data set and about 5000 target language sentences for n-gram LM , in addition to 5000 sentence pairs used for building bilingual language models. The tuning set are around 200 K sentence pairs, this set was filtered by skipping very short or very long sentences. We apply decoder on two different test files. For the general corpus we used nearly same size of the scientific parallel text, utilizing English-Arabic text from the WMT 2013 news-commentary corpus.

The whole set of data were tokenized, lowered case and cleaned using Moses preprocessing scripts, all experiments applied through Moses decoder[32]. The training data were aligned in both directions using GIZA++ [33]. We trained individual LMs for each experiment, using in-domain monolingual only or out-domain monolingual only. KenLM [34] was employed to 5-gram language models. The second set of trained LM based on BiLM,based on both source and target in-domain aligned text using Bilingual Neural LM. As the decoder constructed of different models, such as language model, translation model and reordering model a mean of weight adjusting tools were needed. Minimum error rate, MERT [35] used to find the optimal features weights, it adjusts and finds the set of linear models weights to maximize translation performance on a small set of parallel sentences.

For MT adaptation we experiments the results of two approaches for combining phrase tables of the two corpuses. The Fill-up combination and the Back-off approach are tested [29].Fill-up method keeps all the weights and scores coming from the first model, and adds entries from the other models only if new, and then add a binary feature symbolize to the attribution of an entry. Forming new phrase table and then applying MERT to adjust features weights. The Back-off is a simplification of the fill-up method, didn’t add a binary feature rather it keep same number of scores. This step is held after building two separated translation models, one based on the scientific text and the other one based on the news-commentary corpus.

We implemented a BiLM as a feature function inside Moses, following closely the implementation delineated in [36]. For these experiments we used a target text of four words, and an aligned source window of nine words. As NPLM does not support separate source and target tuples, a parallel corpora used to extract 14-grams which consist of 9 source and 5 target words. Once the 14-grams are extracted we train NPLM on them as if it were a monolingual dataset. That’s clue to a decoder which is about twice as slow as the phrase-based decoder without BiLM.

Table I summarizes results of the first model, using only our scientific corpus for train, language model and tuning. BILM improves the BLEU score, while human revising show some loose of the correct sentence components compared to KENLM. In Table II we list the score of translating general text and scientific test, while only general text from the News-commentary corpus used for training. BLUE score decreased than the first experiment, while using general-domain in both train and test can be due to using scientific-based text for building both KENLM and BiLM language models. Finally results in table III represents the BLUE score translation scientific and general text, while using Fill-up combination method, again it seems higher score for scientific data set is due to the usage of our science data for building LM and tuning step. We compared our results by Google translation, calculating BLEU score for same test text, it's noticeable how it give far than accurate translation for scientific domain. It is clear the misunderstanding for many of the sense and new meaning gained to English words and terminologies when occurred in different domains.

TABLE I TRAIN ON SCIENTIFIC CORPUS

LM	Test Set	
	General-Text	**Scientific-Text**
KENLM	24.89	75.47
BiLM	26.02	78.52

TABLE II TRAIN ON GENERAL CORPUS

LM	Test Set	
	General-Text	**Scientific-Text**
KENLM	69.60	6.49
BiLM	71.63	8.47

TABLE III PHRASE TABLE COMBINATION & GOOGLE BLEU RESULTS

LM	Test Set	
	General-Text	**Scientific-Text**
KENLM	64.47	75.47
BiLM	63.62	78.52
Google	68.54	30.91

4 CONCLUSION

There are very small of Arabic linguistic research based on corpora compared to English and Europe languages researches, as we don't have enough Arabic corpus compared to other languages. As well as the need to a faster transfer of knowledge from English to our language, in order to start our scientific journey adding our contribution. From this concern we plane for more effort building our scientific English-Arabic corpus, to cover different topics in computer science field.

In this paper we addressed the issue of MT domain adaptation in SMT, proposing using different approaches together for as much as possible better translation for scientific text. We examine the efficiency of using the fill-up phrase table combination method besides the Bilingual language models. Both approaches prove efficiency when used together; fulfill better BLEU score than Google for scientific text translation. Combination method show robustness of collecting new domain knowledge, as the new sense a ward gained in the new domain. No significant difference shown when using Back-off instead the Fill-up approach. BiLM proves better BLEU score as was expected, compared to the n-gram language models. It can due to BiLM built upon both source and target languages, as translation is more complicated than using monolingual LM.

Overall, the experiments results prove how the state-of-the-art SMT translation is sensitive to the sublanguage or the text domain used in training the decoder. And that the automatic machine translation performs best depends on how well the test and in-domain training data matches.

REFERENCES

[1] Brown P.F., Stephen A. Della Pietra, Vincent J. Della Pietra, & Robert L. M., *The Mathematics of Statistical Machine Translation*: *Parameter Estimation*, Computational Linguistics, 19(2), pp. 263-312, 1993.

[2] Thouin B., *The METEO system*, In V. Lawson, Practical Experience of Machine Translation, Amsterdam Holland, pp. 39–44, 1982.

[3] Volk M., & Harder S., *Evaluating MT with Translations or Translators. What is the Difference?,* In Proceedings of the MT-Summit, Copenhagen, 2007.

[4] Koehn P. and Schroeder J., “Experiments in domain adaptation for statistical machine translation,” In Proceeding of the Second Workshop on Statistical Machine Translation, StatMT ’07, , Association for Computational Linguistics, Stroudsburg, USA, pp. 224–227, 2007.

[5] Sennrich, Rico, *Perplexity minimization for translation model domain adaptation in statistical machine translation*, In Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, Avignon, France, 2012, 539-549, 2012.

[6] Chen S.F. & Goodman J.,*An empirical study of smoothing techniques for language modeling*, Computer Speech and Language, 4 (13), pp. 359–394, 1999.

[7] Marino J.B., Banchs R.E., Crego J.M., De Gispert A., Lambert P., Fonollosa R., & Costa-jussa M.R., *N-gram based machine translation*, Computational Linguistics, 32 (4), pp. 527–549, ACL, 2006.

[8] Niehues J., Herrmann T., Vogel S., &Waibel A., *Wider context by using bilingual language models in machine translation*. In Proceedings of the Sixth Workshop on Statistical Machine Translation, pp. 198–206, Association for Computational Linguistics, 2011.

[9] Al-Onaizan Y. & Papineni K., *Distortion models for statistical machine translation*, In Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, pp. 529–536, Sydney, Australia, July, 2006.

[10] Sethy A., Georgiou P. & Narayanan S, *Selecting relevant text subsets from web-data for building topic specific language models*, Proceedings of the Human Language Technology Conference, 2006.

[11] Karlgren J. & Cutting D., *Recognizing text genres with simple metrics using discriminant analysis*, In Proceedings of the 15th International Conference on Computational Linguistics, volume 2, pages 1071–1075, Kyoto, Japan, Aug. 1994.

[12] Collins M., Koehn P., & Kucerov I., Clause restructuring for statistical machine translation, Proceedings of the 43rd annual meeting on association for computational linguistics, pp. 531–540, Association for Computational Linguistics, 2005.

[13] Singla K., Sachdeva K., Yadav D., Bangalore S., Misra D.S., *Reducing the Impact of Data Sparsity in Statistical Machine Translation*, Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation, Association for Computational Linguistics pp. 51–56, 2014.

[14] Mikolov T., Karafi’at M., Burget L. Cernock y. J. & Khudanpur S., *Recurrent neural network based language model*,11th Annual Conference of the International Speech Communication Association, 2010.

[15] Bengio Y., Ducharme R.E., & Vincent P., *A neural probabilistic language model*, Journal of Machine Learning Research, 2003, pp. 1137–1155, 2003.

[16] Allauzen A., Crego J., El-kahlout I.D., & Yvon F., *LIMSI’s statistical translation systems*, In fifth workshop on statistical Machine Translation(WMT’10), Uppsala, Sweden, 2010.

[17] Zhao B. & Tam Y-C., *Bilingual Recurrent Neural Networks for Improved Statistical Machine Translation*, In Proceedings of the IEEE Spoken Language Technology Workshop, South Lake Tahoe, pp. 68-70, 2014.

[18] Singla K., Sachdeva K., Yadav D., Bangalore S., Misra D.S., Reducing the Impact of Data Sparsity in Statistical Machine Translation, Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation, Association for Computational Linguistics pp. 51–56, 2014.

[19] Garmash E. and Monz C., Dependency-Based Bilingual Language Models for Reordering in Statistical Machine Translation, Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1689–1700, 2014 Association for Computational Linguistics, 2014.

[20] Bulyko I., MatsoukasS., Schwartz R., Nguyen L., & Makhoul J., *Language Model Adaptation in Machine Translation from Speech*, In ICASSP, Honolulu, USA, 2007.

[21] Chen, B., Kuhn, R., & Foster, G., *Vector Space Model for Adaptation in Statistical Machine Translation*, In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pages 1285–1293, Sofia, Bulgaria, 2013.

[22] Amittai A., Xiaodong H., & Jianfeng G, *Domain Adaptation via Pseudo In-Domain Data Selection*, Proceedings of the Conference on Empirical Methods in Natural Language Processing mentioned in Domain Adaptation, 2011.

[23] Foster G., & Kuhn R., *Mixture-model adaptation for SMT*, in Proceedings of the Second Workshop on Statistical Machine Translation. Prague, Czech Republic: Association for Computational Linguistics, pp. 128–135, 2007.

[24] Carpuat M.& Wu D., *Improving Statistical Machine Translation using Word Sense Disambiguation*, In Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), pp. 61–72, Prague, June 2007.

[25] Lau J., Cook P., McCarthy D., Newman D. & Baldwin T., *Word sense induction for novel sense detection*, In Proceedings of the 13th Conference of the European Chapter of the Association for computational Linguistics, pages 591–601, 2012.

[26] Carpuat M., Daume H., Henry K., Irvine A., Jagarlamudi J., Rudinger R., *Sense Spotting: Never let your parallel data tie you to an old domain*, In proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pages 1435–1445, Sofia, Bulgaria 2013.

[27] Besling S & Meier H., *Language model speaker adaptation*, in Proceedings of the 4th European Conference on Speech Communication and Technology, vol. 3, pp. 1755–1758, Madrid, Spain, 1995.

[28] Nakov P., *Improving English-Spanish Statistical Machine Translation: Experiments in Domain Adaptation, Sentence Paraphrasing, Tokenization, and Recasing*, in Workshop on Statistical Machine Translation, Association for Computational Linguistics, 2008.

[29] Bisazza A., Ruiz N, & Federico M., *Fill-up versus Interpolation Methods for Phrase based SMT Adaptation*, In International Workshop on Spoken Language Translation (IWSLT), San Francisco, USA, 2011.

[30] Zhang J., Liangyou L., Way A., Liu q., *A Probabilistic Feature-Based Fill-up for SMT*, Proceedings of AMTA, vol. 1, Vancouver, Canada, 2014.

[31] Wang Y., Wang L., Wong F. D., Chao S. L., Zeng X., Lu Y., *Factored Statistical Machine Translation for Grammatical Error Correction*, In proceedings of the 18th Conference on Computational Natural Language Learning, pp. 83–90, Baltimore, Maryland, 2014.

[32] Koehn P., Hoang H., Birch A., Callison-Burch C., Federico M., Bertoldi., Cowan B., Shen W., Moran C., Zens R., Dyer C., Bojar O., Constantin A., Herbst E), *Moses: Open Source Toolkit for Statistical Machine Translation*, Annual Meeting of the Association for Computational Linguistics (ACL), demonstration session, Prague, Czech Republic, 2007.

[33] Och F. J., & Ney H, *A Systematic Comparison of Various Statistical Alignment Models*, Computational Linguistics, pp. 19-51, 2003.

[34] Heafield K., *KenLM: Faster and Smaller Language Model Queries*, In proceedings of the Sixth Workshop on Statistical Machine Translation, 2011.

[35] Bertoldi N., Haddow B. and Fouet J-B., *Improved minimum error rate training in Moses*, The Prague Bulletin of Mathematical Linguistics, 2009.

[36] Devlin J., Zbib R., Huang Z., Lamar T., Schwartz R., & Makhoul, J., *Fast and robust neural network joint models for statistical machine translation*, In proceedings of the 52nd annual meeting on association for computational linguistics, 2014.

BIOGRAPHY

Ibrahim F. Eissa

Professor & Former Dean of Faculty of Computers and Information – Cairo University.

Ibrahim F. Imam

Professor of Computer Science, Arab Academy for Science, Technology and Maritime Transport. Dr. Imam received his B.Sc. in Mathematics and Statistics in 1986 and a graduate diploma in Computer Science and Information in 1989 from Cairo University. He received his M.Sc. in Computer Science in 1992 and his Ph.D. in Information Technology and Engineering in 1995 from George Mason University. Dr. Imam edited three international books and several international journals. He chaired three international conferences and workshops. He authored co-authored over 60 papers in refereed journals, conference proceedings, and workshop proceedings. His research focused in the fields of artificial intelligence, data mining, text mining (Arabic & English), pattern recognition, and machine learning. Dr. Imam is a steering committee member of the international journal of artificial intelligence and machine learning. He served as program committee member for many international conferences.

Marwa N. Refaie

Lecturer Assistant in Modern University for Technology and Information, Faculty of Computer Science. She studied B.SC at Modern Academy, 2002. Received Master of Computer Science from Arab Academy for Science, Technology and Maritime, 2006. Researches concern on Machine learning, Computer Vision and Natural Language processing.

نموذج ثنائي اللغة لترجمة النصوص العلمية من الانجليزية الي العربية

مروة رفاعي[*1] , إبراهيم أمام[**2], إبراهيم عيسي[***3]

**الجامعة الحديثة للعلوم و التكنولوجيا, كلية علوم الكمبيوتر*
[1]basmallah@hotmail.com

***الاكاديمية العربية للعلوم و التكنولوجيا و النقل البحري , قسم علوم الكمبيوتر*
[2]ifi05@yahoo.com

****كلية الحاسبات و المعلومات, جامعة القاهرة*
[3]i.farag@fci-cu.edu.eg

الملخص:

ان سرعة اصدار الأبحاث العلمية تزيد من الحاجة الي نظام ترجمة آلي يمكن الاعتماد اليه في اصدار ترجمة صحيحة , خاصة من الانجليزية كاللغة الأكثر شيوعا للأبحاث العلمية الي اللغات الأخري. غالبا ما تفشل الترجمة الآلية الإحصائية التي صممت للتعامل مع نصوص من مجال معين عند ترجمة نصوص من مجالات مختلفة. ان هذا البحث يعرض الخصائص لمجالات النصوص و كيفية التعامل معها من خلال الترجمةالآلية الإحصائية, و يعرض البحث من خلال مجموعة من التجارب طريقة التناول لترجمة أبحاث علمية في مجال الذكاء الاصطناعي. كما تم تقييم و مقارنة مدي كفاءة استخدام نماذج اللغة المختلفة , بالاضافة الي الاستفادة من طرق دمج نماذج الترجمة في مجالات مختلفة. كما ان ليس كل انسان يستطيع ترجمة كتاب في الذكاء الاصطناعي , حيث يجب ان يستمتع بمعلومات كافية في نفس المجال , هكذا الترجمة الآلية. حيث يجب تدريب برنامج الترجمة الآلي علي مجالات مختلفة من اللغة, و تعلم الفروق بين ترجمة نفس الكلمة في عدة مجالات.