A Proposed Model for Enhancing Lexical Statistical Machine Translation (ELSMT)

Prof. AlaaEl-Din M. El-Ghazli
Prof. of Computer and Information Systems - SAMS

Dr. Ahmed S. Salama
Teacher of Computer and Information Systems - SAMS

Ahmed G. Elsayed
MSc. Candidate - SAMS

ABSTRACT

Statistical Machine Translation (SMT) deals with automatically mapping sentences in one human language into another human language. This means that it translates from the source language to the target language, so the goal of SMT is automatically analyze existing human sentence translations, to build general translation model for translation.

A model presented for efficiently incorporate models, which used before in statistical machine translation such as language model, alignment model, phrase based model, reordering model, and translation model. These models combined to enhance the performance of statistical machine translation (SMT). One of the advantages of the statistical approach to machine translation is that it is largely language agnostic. Machine translation models used to learn automatically translation patterns from data. This research introduces a model, which might be used to translate from the source to the target sentence automatically.

There are many tools have been used in this work such as Gizaa++. All these tools used to take the advantage of the previous mentioned models combined together with each other.

Finally, based on the implementation of this model, it has proved that this model has improved the result of the statistical machine translation.

General Terms

Machine Learning, Machine Translation

Keywords

Machine Learning, Machine Translation, Linguistics

1. INTRODUCTION

Statistical Machine Translation (SMT) deals with automatically mapping sentences in one human language into another human language.[8] Therefore, it translates from the source language to the target language.

The goal of SMT is to analyze automatically existing human sentence translations, to build general translation rules.

The problem of machine translation has not solved yet. Much research and development still needed to earn the reliability of humans by preforming a fluency translation as humans do.

Therefore, a method represented to tune the translation parameters and improve the translation quality.

2. LITERATURE REVIEW

In general, many models have been used in machine translation for many years. SMT was the preferred approach for research and development of machine translation systems within this period.

In 2005, Joseph Olive started a program in US Defense Advanced Research called GALE project “Global Autonomous Language Exploitation”. [7] Rabeh Zbib conducted a research in 2010 for how to uselinguistic knowledge in statistical machine translation in MIT, USA.[14]Ahmed Ragab Nabhan and Ahmed Rafea in 2004 have done a research on tuning statistical machine translation parameters in central laboratory for agricultural expert system (CLAES) in Egypt.[5]Philipp Koehn in 2004 presented pharaoh: a beam search decoder for phrase-based statistical machine translation models, MIT, USA.[3]Rabih Zbib, Spyros Matsoukas, Richard Schwartz, and John Makhoul.in 2010 conducted a research for using decision trees for lexical smoothing in statistical machine translation in MIT, USA.[13]Papineni, K.; Roukos, S.; Ward, T.; Zhu, W. J. in 2002 introduced the "BLEU: which is a method for automatic evaluation of machine translation" in ACL40thannual meeting of the association for computational linguistics in USA.[9] Ibrahim Badr, Rabih Zbib, and James Glass in 2008 done a research on segmentation for English-to-Arabic statistical machine translation in MIT, USA.[1] Franz Och and Hermann Ney have done a research in 2000 about the improved statistical alignment models in ACL.[6]Kenneth Heafiel, Philipp Koehn,and Alon Lavie did a research in 2013 aboutgrouping language model boundary words to speed K-Best extraction from hypergraphs in proceedings of the conference of the north American chapter of the association for computational linguistics (NAACL).[2]In 2004 Noah A. Smith has done a research on Log-Linear Models in Johns Hopkins University in USA.[11]

3. THE MODEL ARCHITECTURE

Architecture of the model for enhancing lexical statistical machine translation shown inFig 1. Main model components are: data preparator, learner, translation model tuner, and translation decoder. These components are discussed briefly in the following subsection.

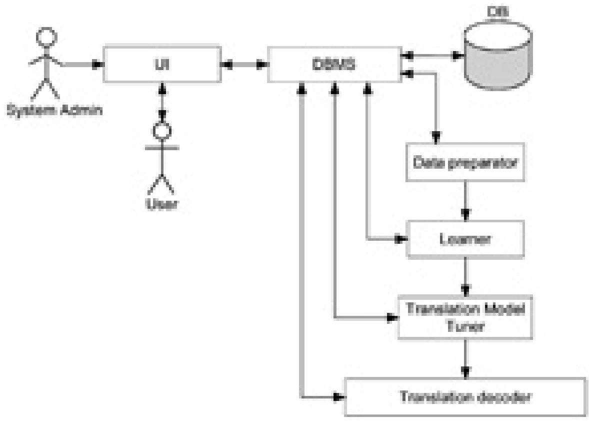

Fig 1 Architecture Diagram of the Proposed Model for enhancing SMT

1.1 Data Preparator

This component focuseson converting the input parallel corpus into format, which is suitable for the next step to use them in this model. Data preparator of this model consist of four components and they are as follows:

- Normalization by converting all words of source and target language to upper or lower cased in all sentences.

- Tokenization for all data in the both language corpus data by inserting putting spaces between words and punctuation.

- True-casing by generating probabilities for all words in the parallel corpus and building a model for this truecasing to be used for generating the file of the truecasing for each language.

- Cleaning data by delete long sentences, which are longer than specific number or remove empty sentence or misaligned sentences, which can affect the translation quality.

1.2 Learner

This is the core of this model. In this component, many models combined to perform a hybrid model for data training to get better translation result. These components are the following:

1.2.1 Language Model (LM)

Language model built to make sure that translation has been performed correctly with fluency, so it has been built in Arabic

language, which is the target language. Language model built based on a combination of ngram models.The language model is based on a combination of ngram models. It consists of the 5gram, quadgram, trigram, bigram, and unigram language models plus adding symbols to sentence boundaries.

The "language model" or "lm" is a statistical description of one language that includes the frequencies of token-based n-grams occurrences in a corpus. Language modelistrained from a monolingual corpus and saved as a file. The language model file is a required component of every translation model.[4]

1.2.2 Alignment Model

Alignment models used in statistical machine translation to determine relation between translation words in a sentence in one language compared to the words in a sentence with the same meaning in a different language.The alignment model form an important part of the translation process, as it's used to produce word-aligned text, which is used to create machine translation systems.

IBM five models are famous models for alignment. They were proposed about 25 years ago from now but they still form the state of the art models for alignment. They consist of five models for word-to-word alignment called "IBM 1-5 models".With alignment model, a generalization of the procedures of extracting the phrase of phrase-based systems with its corresponding sequence can be done. Alignment models will be used to align words to enhance translations.

Aligned data are elements of a parallel corpus in two languages. Each element in one language matches the corresponding element in the other language.

This model consists of the training model for inspecting the alignments for data and create alignment table which used later to align words and help in perform reordering model.

1.2.3 Translation and Phrases Extraction

This component used to get the direct translation for words then create table with maximum likelihood for words translations and extract phrases in one file with their scores.

It extract the phrase table, which is a statistical description of a parallel corpus of source-target language sentence pairs. It extract the phrase table, which is a statistical description of a parallel corpus of source-target language sentence pairs.

1.2.4 Reordering Model

After phrases extraction with their scores then it has used the alignment model to generate reordering table. This reordering table is the component of the reordering model component.Reordering table will be used to reorder phrases, and will be used in the translation model. Reordering table contains statistical frequencies that describe changes in word order between source and target language.

1.2.5 TranslationModel

This is the final components generated as output after training all the previous components, which will be tuned later or can be used to generate the translation directly. The translation model containsalignment model, phrase table, translation table, reordering table, and language model, which will be used for translation.

1.3 Translation Model Tuner

The main purpose of tuning the trained SMT model is to enhance statistical machine translation results and improve its quality. Since SMT, training uses a linear model, the tuning aims to find the optimal weights for this linear model and minimize error rating, where optimal weights are those, which maximize translation performance.

Therefore, that tuning is a process of finding the optimized settings for the translation model. The tuning process translates thousands of source language phrases in the tuning set with a translation model, compares the model's output to a set of reference human translations, then it adjusts the settings with the intention to improve the translation quality. This process continues through iterations. With each iteration, the tuning process repeats the steps until it reaches an optimized translation quality and minimization of error rate.

1.4 Translation Decoder

This is the final component of this model, which applies the translation model with its components on the entered English sentence by user to translate it to the appropriate translation then send the Arabic translation back to user interface in order to display it to the user.

2. THE PROPOSEDMODEL TRAINING AND LEARNING ALGORITHMS

1. System administrator will support the dictionary and feed it with vocabularies and translated sentences or parallel corpora, which used as a reference for the translation model. He will enter this data through the user interface (UI) and Database Management System (DBMS) tool.

2. Prepare the parallel corpora for the training process by normalizing words, tokenization, truecasing, and cleaning.

3. Create the language model for the source language to help in translation process.

4. Train the word alignment model with parallel corpus data.

5. Save the alignment model files for references and to help create the reordering model.

6. Generate translation table with maximum likelihood estimation.

7. Save the translation table.

8. Build phrases table for translated phrases with their scores.

9. Save the phrases table.

10. Build reordering model.

11. Save the reordering table.

12. Build the translation model.

13. Save the translation model.

14. Tune the translation model and minimize the error rate for translation to improve translation quality and add model upgrade.

15. Binarise the phrase table and reordering table.

By performing all the previous steps, the translation model would be trained for parallel corpus.

3. IMPLEMENTATION OF THE MODEL

This model presents an integration between models for enhancing lexical statistical machine translation including language modeling, alignment model, phrase based translation model, reordering model, translation modeling, and tuning the final translation model. All these models combined together in order to improve the quality of the output of statistical machine translation and build a more reliable model.

The Implementation of this model with an experiment presented to prove that the proposed model enhances lexical statistical machine translation result from English verb phrase to Arabic. In this experiment, a data have been used from collection of translated documents from the United Nations originally compiled into a translation memory by Alexandre Rafalovitch, and Robert Dale for the training process. A data also used from a collection of translated sentences from Tatoeba for the tuning process. In addition, a parallel corpus from Ubuntu localization files have been used for testing and getting the BLEU result.

3.1 The Proposed Model Implementation Tools

This model used a various set of tools to generate the final model. These tools are as follows:

- IRSTLM Toolkit for language modeling.

- KenLM also for language modeling.

In language modeling both IRSTLM Toolkit and KenLM tool were used together combined with the main tool for language modeling and training all of them have to be compiled with each other. Both of them do the same purpose but the only difference is IRSTLM is better in query and editing the model, but KenLM is faster in model creation because it is multithreaded.

- GIZA++ word aligning tool to align the parallel corpus and train the alignment model.

This toolkit is an implementation of the IBM models that started statistical machine translation research and the state of the art techniques.

- MOSES as a complete statistical machine translation system.

In the rest of this model for training and tuning, Moses have been used which allows training translation models for language pairs. Once creating the translation model, an efficient search algorithm finds quickly the highest probability translation among the exponential number of choices.

- Bi-Lingual Evaluation Understudy (BLEU) score tool, which is the most famous scoring, and testing tool for statistical machine translation quality was used for model evaluation. The BLEU score indicates how closely the token sequences in one set of data for example in machine translation output, correlate with or match the token sequences in another set of data, such as a reference human translation.

3.2 Model Testing

In this, section an illustration of how the model was implemented with the training steps.

3.2.1 DataUsed

As input for this experiment to test this model, a small amount of data used from a collection of translated documents from the United Nation. For model training a set of 74067 sentences were used before data cleaning which have been shortened to 67575 sentences after data cleaning.[10] For tuning a collection of translated sentences from Tatoeba were used.[12] Finally for model evaluation a parallel corpus of Ubuntu localization files were used to get the BLEU result and assess the model.[12]

After the collection of these data, it must be prepared in order to train the model and tune it then evaluate it.

3.2.2 Evaluation Results

To evaluate this model, the BLEU tool has been used to compare the human translation of parallel corpora from Ubuntu with this model translation and get the BLEU score. As known that more data used in training the model, the better result we get. In case of using data from the same domain for testing, the less errors would be achieved from the model and the better translation BLEU result accomplished. When using small number of corpus, then the expected BLEU score is very low.

For evaluating this model, a parallel corpus from Ubuntu localization files consists of 6000 sentences have been used. The score for this model before tuning was 8.99. This considered a very good result compared to the number of data used for training in this model. In addition, the data used for the model evaluation is not from the same domain for model training, which decrease the BLEU result.After tuning this model, the BLEU score shows a significant improvement as it achieved a result of 19.52.

As a conclusion for this model, it shows that after tuning the model, the better result achieved. The model can be trained many times and every time it will perform better and give better result. This proves that this model has enhanced the statistical machine translation quality.

3.2.3 Implementation Results Using Case Study

This model for enhancing lexical statistical machine translation will be under manual test and evaluation by humans using sentences from the same domain of the training domain to be able to make sense for human evaluation.

1.1.1.1 English Sentences used for evaluation

- There are crimes against women in Brazil.
- He did the same thing last summer.

1.1.1.2 Model Translation before Tuning

- The Arabic translation for the sentence “There are crimes against women in Brazil” before tuning as shown inFig 2 is

“هناك الجرائم ضد المرأة فى البرازيل”

Fig 2Arabic translation of “There are crimes against women in Brazil” before tuning

- The Arabic translation for the sentence “He did the same thing last summer” before tuning as shown in

Fig 3is

“التي لم نفس thing الصيف الماضى”

Fig 3Arabic translation of “He did the same thing last summer” before tuning

1.1.1.3 Model Translation after Tuning

- The Arabic translation for the sentence “There are crimes against women in Brazil” after tuning as shown inFig 4is

“هناك الجرائم المرتكبة ضد النساء فى البرازيل”

Fig 4 translation of “There are crimes against women in Brazil” after tuning

- The Arabic translation for the sentence “He did the same thing last summer” after tuning as shown in

Fig 5 is”ولقد فعل نفس الشىء في الصيف الماضى”

Fig 5Arabic translation of “He did the same thing last summer” after tuning

As shown from the previous translations, the translation before and after tuning are not the same. A huge improvement in translation noticed in translation after tuning than the translation before tuning. Tuning can be done for this model several times to improve the model quality, and many parallel corpus can be trained to give better translation results.

2. CONCLUSIONS AND RECOMMENDATIONS

The presented model incorporate efficiently some other models at different levels namely: the language model, the alignment model, phrase based model, reordering model, translation model, and finally the tuning model.

2.1 Conclusions

From the implementation of this proposed model for enhancing lexical statistical machine translation, we can conclude that when combining different models in building and training statistical machine translation model, ithelps in achieving good results. In addition, translation results have been improved after tuning and the error rate has been minimized.New words, vocabularies, phrases, and sentences can be added and new models can be trained for these entered data to learn from these human translation.Model can be trained and tuned several times with new data to get better and better results.Context dependent language model achieves better results and predict the translation more accurate, but also the more time cost. The more models combined to train data with a certain order can help in achieving better translation results.

The proposed model uses a combination of models to generate the translation model. These models are the alignment models used in statistical machine translation to determine translational correspondences between words in a sentence in one language with the words in a sentence with the same meaning in a different language.Language models increase the efficiency of the word alignment by using words depending on their context in the sentence.Phrase based model increases the capability of the proposed model by dealing with words and their correspondences or phrases in both languages at the same time. It adds value to the effectiveness of this model by translating a group of contiguous words in one language to a contiguous sequence of words in the other language.Reordering model uses all the mentioned previous models to generate reordering table by determining the orientation of two phrases based on word alignments at training time. This adds extra points to the reliability of this model and increases its dependability.The implementation of this model proves by an evaluation of an experiment done that it has improved the quality of statistical machine translation. It has proved that this model is a reliable and can be used for enhancing lexical SMT systems.

2.2 Future Work

The demand of faster and cheaper translation between languages will only increase with the need to share information between nations.

Future work and recommendations, building on the results of this model can be done through many ways such as using large amount of parallel corpus data to train the proposed model, to achieve better results. Use large amount of data for tuning from the same domain of the training data to get better tuning result. Every time this model is tuned, the better results it can achieve, so it would be better to tune it as much as possible.Use testing data from the same domain to get better actual results.

3. REFERENCES

[1] Badr, Ibrahim. Zbib, Rabih. Glass,James. (2008. Segmentation for English-to-Arabic Statistical Machine Translation. MIT, USA).

[2] Heafield,Kenneth. Koehn, Philipp. Lavie, Alon. (2013. Grouping Language Model Boundary Words to Speed K-Best Extraction from Hypergraphs. Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL). US).

[3] Koehn, Philipp. (2004. Pharaoh: A Beam Search Decoder for Phrase-Based Statistical Machine Translation Models. MIT, USA).

[4] Koehn, Philipp. (2010. Statistical Machine Translation. Cambridge University Press. Cambridge, United Kingdom).

[5] Nabhan, Ahmed. Rafea, Ahmed. (2004. Tuning Statistical Machine Translation Parameters, Central Laboratory for Agricultural Expert System (CLAES). Egypt).

[6] Och, Franz. Ney,Hermann. (2000. Improved Statistical Alignment Models. In Proc. of ACL, US).

[7] Olive, Joseph. (2005. GALE Program Manager, Defense Advanced Research).

[8] Osborne, Miles.(Statistical Machine Translation. University of Edinburgh, UK).

[9] Papineni, K.. Roukos, S.. Ward, T.. Zhu, W. J.. (2002. “BLEU: a method for automatic evaluation of machine translation”. ACL-2002: 40thAnnual meeting of the Association for Computational Linguistics. USA).

[10] Rafalovitch,Alexandre. Dale, Robert. (2009. United Nations General Assembly Resolutions: A Six-Language Parallel Corpus. In Proceedings of the MT Summit XII, pages 292-299, Ottawa, Canada, August).

[11] Smith, Noah. (2004. Log-Linear Models, Johns Hopkins University, USA).

[12] Tiedemann, Jörg.(2012, Parallel Data, Tools and Interfaces in OPUS. In Proceedings of the eighth International Conference on Language Resources and Evaluation (LREC 2012)).

[13] Zbib, Rabih. Matsoukas, Spyros. Schwartz, Richard. Makhoul, John. (2010. Decision Trees for Lexical Smoothing in Statistical Machine Translation. MIT, USA).

[14] Zbib, Rabih. September (2010. Using Linguistic Knowledge in Statistical Machine Translation. MIT, USA).