Research Article

Translating Idioms using Paraphrasing, Machine Translation and Rescoring

Tien-Ping, Tan[1], Jia Jun, Dong[2]

[1,2]School of Computer Sciences, Universiti Sains Malaysia, Penang, Malaysia
[1]tienping@usm.my, [2]dongjiajun26@gmail.com

Article History: Received: 10 November 2020; Revised: 12 January 2021; Accepted: 27 January 2021; Published online: 05 April 2021

Abstract: Idioms are rich multi-word expressions that can be found in many works of literature. The meaning of most idioms cannot be deduced literally. This makes translating idioms challenging. Moreover, the parallel text that contains idioms is limited. As a result, machine translation has difficulty in translating idioms correctly. Paraphrasing is a process to restate the meaning of a text or a passage using different words in the same language. Often, paraphrasing is used to give readers a clearer understanding of the original text. Paraphrasing can be used to assist machine translation in translating idioms. In this article, we attempted to improve the translation of idioms using paraphrasing. An approach that combine paraphrasing and rescoring with machine translation is proposed. The paraphrasing and rescoring improve the translation produced by neural machine translation from 12.03% to 12.92%.

Keywords: idiom, machine translation, paraphrasing, rescoring

1. Introduction

Idiom is a multi-word expression that is interesting. The meaning for an idiom cannot be deduced literally and very often they tell a story. Idiom is one of the rhetorical devices besides figure of speech, metaphor, and cliché. A sentence containing idioms often end up with a bad translation even with the state-of-the-art of machine translation (MT) system. This is due to scarce amount of parallel text containing idioms is available for building the translation model of an MT. As a result, the MT may not recognize the idiom in a text, and unable to translate it, or the MT may translate the idiom word by word, and the meaning of the idiom may end up being incorrect. Consider the Chinese idiom 打铁趁热 and the equivalent English idiom "struck while the iron was hot". Obviously, this idiom should not be translated literally because the meaning does not reflect from the words directly. Most of the idioms in Chinese consists of four characters, and they are known as chengyu. The total number of idioms in Chinese is quite big. The chengyu dictionary 汉语成语 (2018) for instance contains 8 thousand idioms, while the dictionary 新华成语大词典 (2017) and 20000 条成语大词典(全新版) (2016) have documented 20 thousands idioms in the dictionaries.

Paraphrasing is a process to restate the meaning of a text in a different form, often, the purpose is to give readers a clearer understanding. Most state-of-the-art paraphrasing approach applies neural networks architecture (Mallinson et al., 2017, Gupta et al., 2018, Hussain et al., 2019, Li et al. 2018). Paraphrasing has been receiving wide attention from researchers recently due to its importance especially in the areas of plagiarism detection (Barrón-Cedeño et al., 2013), information retrieval (Wallis, 1993), machine translation, computer-aided writing and others. In the domain of plagiarism and information retrieval, the main study of paraphrasing involves detecting texts that convey similar ideas. On the other hand, in machine translation, it is the generation of paraphrases that are the focus. For instance, paraphrasing can be used to rephrase a sentence that contains unknown word (UNK) words or multi-words expressions to a different text that the MT recognizes. The sentence (i) below is a Chinese sentence containing the idiom '人山人海'. If the idiom is not in the list of vocabulary of the MT system, an MT will not able to translate the idiom because the idiom is a UNK.With a paraphrasing system, the sentence however can be rephrased to sentence (ii), and subsequently the MT will be able to translate the paraphrased sentence (ii) to sentence (iii).

i. 百货公司里**人山人海**。
ii. 百货公司里**有很多人**
iii. The department store **has a lot of people**.

Since the resources (such as parallel text) containing idioms are limited, we are interested to investigate paraphrasing to rephrase the sentence before translating it.

2. Machine Translation

The state-of-the-art machine translation systems are statistical machine translation (SMT) and neural machine

translation (NMT) (Luong, 2016). Both are data-driven approach, where the translation model is trained using a parallel text corpus. A SMT is given a sentence S = [s_1 s_2 ... s_n] in a particular source language, the SMT will produce a translated sentence T = [t_1 t_2 ... t_n] in the target language where si is a word/phrase in the source language, and t_j is a word/phrase in the target language. There are many combinations of T sentences that may be produced. During decoding, SMT selects the sentence T* that has the highest probability, given the source language sentence S as follows:

$$T^* = argmax(P(S|T) \times P(T))[1]$$

P(T) is the target language model. A text corpus is required to build a language model, where it stores the statistics of word n-gram. On the other hand, P(S|T) is the probability of the source language sentence given a target language sentence that is modeled by a translation model. A translation model consists of a phrase translation table and a reordering table trained using a parallel text corpus. Thus, to use an SMT for paraphrasing, the parallel text corpus must be a paraphrase corpus.

The state-of-the-art NMT uses a recurrent neural networks (RNN) architecture. A recurrent neuron looks like a typical neuron, but it has an additional feedback loops to allow present information to be used for subsequent neuron to make decisions. The encoder-decoder framework is a widely used in machine translation. To translate a source sentence, the words in the sentence is first converted to vectors through a process known as word embedding. The word vectors are input to the encoder RNN. The encoder RNN will accept the input word vector xt and the previous state, h_{t-1} and produce a new state, h_t. W_{xh} and W_{hh} are the weight matrixes learned during training.

$$h_t = \sigma(W_{xh} + W_{hh}h_{t-1}) \quad [2]$$

At the decoder RNN, during training, the translation with added special start of the sentence tag and end of the sentence tag (e.g. start tag is <GO>, end tag is <EOS>) is converted to word vectors before input to the decoder RNN. During decoding, the vector of the start of the sentence tag is input to the RNN, and the RNN will start producing translation in the target language until the end of the sentence tag is encountered [15].

$$h_t = \sigma(W_{xh} + W_{hh}h_{t-1}) \quad [3]$$

$$s_t = W_{hy}h_t \quad [4]$$

$$p_t = softmax(s_t) \quad [5]$$

where softmax() is the softmax function. W_{xh} and W_{hh} and W_{hy} are the weight matrixes learned during training.

3. Materials and Methods

In this paper, we propose an approach that combines paraphrasing and MT to translate sentences that contain idioms. The approach translates Chinese sentences to English. It takes advantage of the strength of both approaches to produce a better translation of idiom. The idea behind the approach is to first paraphrase the idioms before translating the Chinese sentence to English. At the same time, the original sentence will be translated to English directly. The two hypotheses will then be scored. The reason to score the hypotheses is to select one of them as the best answer because the translation quality produced by both approaches depends on the availability of the training data. Figure 1 shows the steps of our approach.

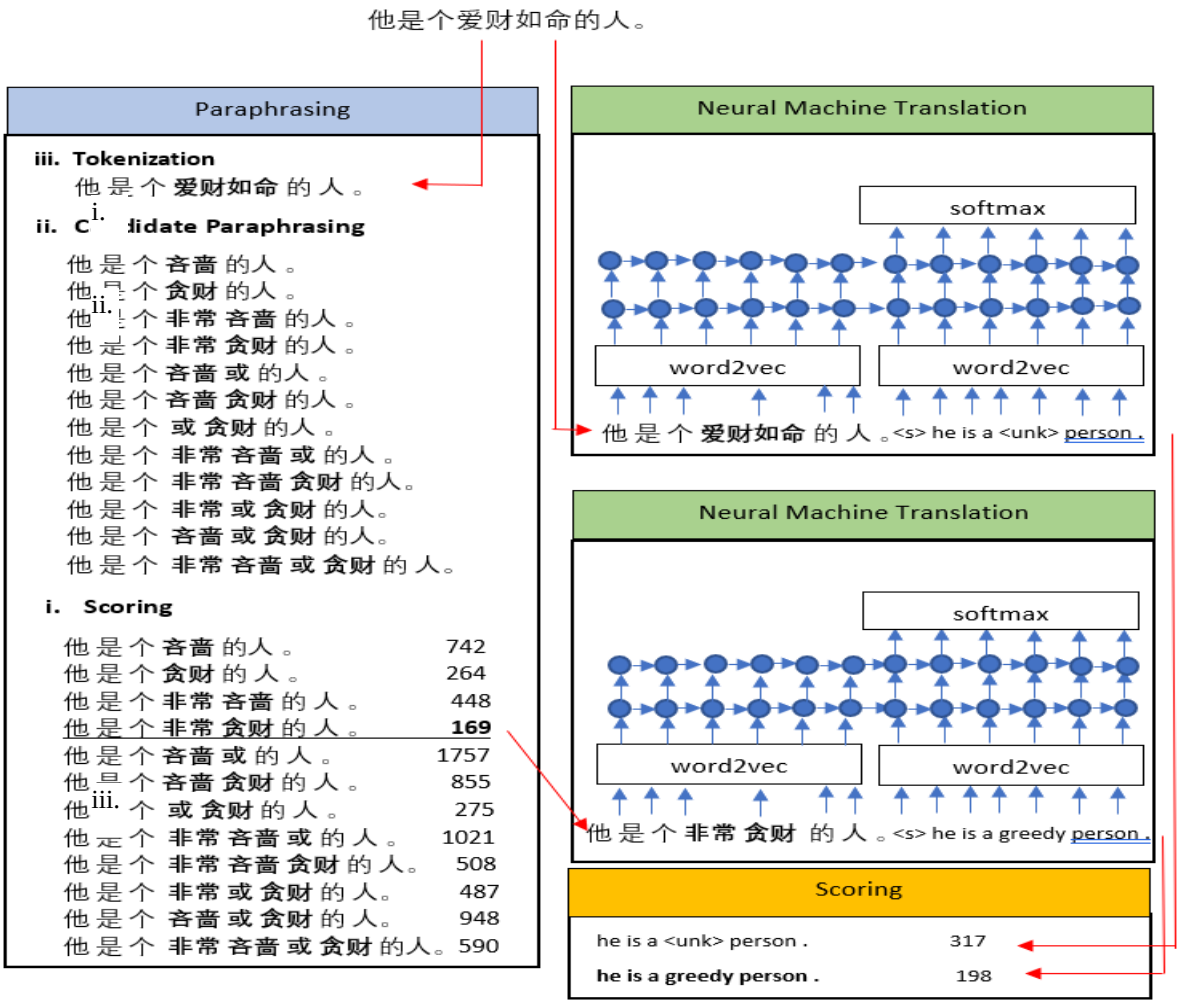

Figure 1. Combining Paraphrasing+MT and MT

In Figure 1, the example sentence '他是个爱财如命的人' (English: he is a person who loves money) is input to a a neural machine translation (NMT) and a paraphrasing system at the same time. The NMT is an attention based bidirectional LSTM encoder-decoder network. The sentence is tokenized with a Chinese tokenizer, and the tokens will be converted to vectors through word2vec and decoded by the NMT to English. In the example given in Figure 1, step 1b, the NMT decodes the sentence to "he is a love money person". The same sentence is processed by the paraphrasing system at step 1ai. At the paraphrasing system (Dong and Tan, 2020), the sentence is tokenized, and the idiom in the sentence is identified. Paraphrase candidates for the idiom will be generated, and the best matching paraphrase candidate for the context is the paraphrase candidate with the lowestscore. The output '他是个非常贪财的人' is produced. The sentence is then subsequently input to an NMT that will translate it at step 1aii. The sentence "he is a greedy person" is produced. There are two hypotheses produced at step 2. One of the outputs is produced directly by the MT, and another output is produced going through paraphrasing and MT. The final step is to score the two hypotheses to select the best hypothesis.

For this purpose, we use perplexity as the metric to select the best English hypothesis. The best hypothesis is the one with the lower perplexity score. Perplexity is the metric normally used in language modeling to select the best language model. Given some test sentences, the lower the perplexity produced, the better the language model. The perplexity for a trigram language model can be calculated using equation [6].

$$PP_{trigram}(W) = \sqrt[N]{\prod_{i=1}^{N} \frac{1}{P(w_i|w_{i-1},w_{i-2})}} \quad [6]$$

$$W' = argmin(PP_{trigram}(W)) \quad [7]$$

In our case, the language model used to evaluate the sentences is fixed. We calculate perplexity for different sentence given the same language model, and select the sentence with the lowest perplexity as the best answer. See equation [7].

4. Results and Discussions

For testing, we collected Chinese sentences containing idioms from the Chinese idiom dictionary by Jiao et al. (2011), 500 Common Chinese Idioms: An Annotated Frequency Dictionary. 454 of the sentences were evaluated. The THULAC Chinese tokenizer was used to tokenize the sentences. The paraphrasing system was from Dong and Tan (2020). We trained a Chinese-English neural machine translation (NMT) that uses attention-based bidirectional LSTM encoder-decoder architecture (Bérard et al., 2016) using about 9 million parallel sentences from WMT17 (http://www.statmt.org/wmt17/). The number of layers was set at 2 and the size of the cell is set at 512. The optimization algorithm used in the training was Adam. The development data consists of 2000 parallel sentences also from WMT17. Three baseline experiments were conducted: 1) idioms in sentences were substituted with the <UNK> tag (, an OOV tag) 2) sentences that contain idioms were translated using NMT directly. 3) sentences were first converted to paraphrases and the paraphrases were translated using the NMT (Paraphrasing +NMT). We compared the translation quality using BLEU (Papineni et al., 2002).

The first experiment was conducted to know the lower bound BLEU result of the NMT, where the MT does not recognize any idiom. To achieve that, all idioms in the sentences were substituted with the <UNK> tag before translating them. We conducted the experiment, and the BLEU obtained was 12.03%. In the second experiment, the test sentences containing idioms were translated by NMT directly. The BLEU for the second experiment was 12.76%. From this result, we can deduce that the parallel sentences used in the training contains some examples of idioms, because the BLEU improve about 0.7% compared to the first baseline. In the third experiment, the idioms in the sentences were first paraphrase before feeding the input to an NMT. The BLEU obtained for Paraphrasing +NMT was 12.61%. The result was better than the baseline system in the first experiment, where the MT did not recognize any idiom, but slightly bad compared to using only NMT. Nevertheless, either one of the systems can perform better in reality depending on the amount of training data available.

In the final experiment, we rescored the translations in English produced by both the systems in experiment 2 and the system in experiment 3 to select the translation with a lower perplexity. For calculating the perplexity will require a language model. We build an English language model using about 2 GB of text corpus. The text was first tokenized using NLTK word tokenizer (Bird et al., 2009), before input to the SRILM toolkit (Stolcke et al., 2011). The modified Kneser-Ney discounting and interpolation was applied. The translation with the lowest perplexity was selected.The BLEU score after rescoring improved to 12.92%. Table 1 below summarize the results of the experiments. We select one idiom, “逍遥法外” and show the analysis of the idiom below.

- Original Sentence: 这起案件已经发生一年多了，可是凶手依然逍遥法外。
- Reference: this case occurred over a year ago , yet the murderer is still at large .
- Paraphrase Sentence: 这起案件已经发生一年多了，可是凶手依然犯了法的人没有受到制裁。
- Paraphrase & Translated Sentence: the case has been in place for more than a year , but the murderer did not have the sanction of the law .
- Translated Sentence (direct): the case has been in place for more than a year , but the murderer is still at large .
- Selected sentence by the system: the case has been in place for more than a year , but the murderer is still at large

Table 1. Comparing the results of translating idioms using different approaches

Approaches	BLEU (%)
1. Idioms in sentences were converted <UNK> before the sentences were decoded by NMT	12.03
2. Sentences containing idioms were decoded by NMT	12.76
3. Sentences containing idioms were paraphrased before decoded by NMT	12.61
4. Results from (2) and (3) were combined and scored using perplexity	12.92

5. Conclusion

In this paper, we propose an approach to translate sentences containing idioms by first paraphrasing the sentences. The experiment result shows thatparaphrasing the idiom is useful when the MT does not recognize the idioms in the sentences. However, when the MT was trained with some translation examples containing the idiom, the paraphrasing step may not be necessary. One reason is because the paraphrasing process may

introduce some errors into the translation. However, if the MT is not trained with sufficient examples, the additional paraphrasing step may still be beneficial. Thus, we introduce a rescoring step to evaluate the translations produced by both approaches. The result show that the overall result can be improved by scoring the hypotheses using perplexity metric. Our future work is to evaluate the proposed approach on Malay idioms (*peribahasa*),

6. Acknowledgment

The authors would like to express appreciation for the support of the Universiti Sains Malaysia [Project Number = 304.PKOMP.6316283].

References

1. Barrón-Cedeño, A., M. Vila, M. A. Martí and P. Rosso. 2013. Plagiarism meets paraphrasing: Insights for the next generation in automatic plagiarism detection. Computational Linguistics, 29(4): 917-947.
2. Bérard, A., O. Pietquin, L. Besacier and C. Servan. 2016. Listen and translate: A proof of concept for end-to-end speech-to-text translation, Proceedings of NIPS: 1–5.
3. Bird, S., E. Klein and E. Loper. 2009. Natural language processing with Python. Sebastopol, O'Reilly Media.
4. Chinese Academy of Sciences, ed. 2002. 现代汉语词典 [Contemporary Chinese Dictionary], Beijing: Foreign Language Teaching and Research Press.
5. Dong, J. J. and T.-P. Tan. 2020. Paraphrasing chinese idioms: Paraphrase acquisition, rewording and scoring. (Submitted) Applied Informatic International Conference.
6. Duboue, P.A. and J. Chu-Carroll. 2006. Answering the question you wish they had asked: The impact of paraphrasing for question answering. Proceedings of HLT-NAACL, New York: 33-36.
7. Géron A., 2017. Hands-on machine learning with Scikit-Learn and TensorFlow: Concepts, tools, and techniques to build intelligent systems, O'Reilly Media.
8. Gupta, A., A. Agarwal, P. Singh, P. Rai. 2018. A Deep Generative Framework for Paraphrase Generation. Proceedings of AAAI-18: 5149-5156.
9. Hussain, A., Surendar, A., Clementking, A., Kanagarajan, S., Ilyashenko, L.K. (2019). Rock brittleness prediction through two optimization algorithms namely particle swarm optimization and imperialism competitive algorithm. Engineering with Computers, 35 (3), pp. 1027-1035.
10. Jiao, L., C.C. Kubler and W. Zhang. 2011. 500 Common Chinese Idioms: An annotated Frequency Dictionary, Routledge, Abingdon.
11. Koehn, P., H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin and E. Herbst. 2007. Moses: Open source toolkit for statistical machine translation. Proceedings of the ACL, Prague: 177–180. Luong, M.-T. 2016. Neural machine translation, Stanford University, Thesis.
12. Mallinson, J., R. Sennrich and M. Lapata. 2017. Paraphrasing revisited with neural machine translation. Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, Valencia: 881–893.
13. Mihalcea, R., C. Corley and C. Strapparava. 2006. Corpus-based and knowledge-based measures of text semantic similarity. Proceedings of the Artificial intelligence: 775–780.
14. Luong, M.-T. 2016. Neural Machine Translation, Stanford University, Thesis.
15. Papineni, K., S. Roukos, T. Ward and W.-J. Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. Proceedings of ACL, Philadelphia: 311-318.
16. Li, Z., X. Jiang, L. Shang and H. Li. 2018. Paraphrase Generation with Deep Reinforcement Learning. Proceedings of 8 Conference on Empirical Methods in Natural Language Processing, Brussels: 3865-3878.
17. Stolcke, A., J. Zheng, W. Wang and V. Abrash. 2011. SRILM at sixteen: Update and outlook. Proceedings of ASRU Workshop, Hawaii.
18. Wallis, P. 1993. Information retrieval based on paraphrase. Proceedings of PACLING, Vancouver: 118-126.