Improving Example-based Machine Translation by Analogy Using Parse Trees

Yifei Zhou Yves Lepage
早稲田大学大学院情報生産システム研究科
yifei.zhou@ruri.waseda.jp, yves.lepage@waseda.jp

Abstract

We investigate an approach to example-based machine translation (EBMT) implemented by analogy in a low-resource scenario. This analogical approach requires analogies between sentences in the source language contained in the knowledge database. We use sentence analogies extracted by parse trees to improve the overall quality of translation. We demonstrate that our method is more effective by comparing the translation quality to other baseline systems. Our method surpasses the results of a recurrent neural network (RNN) model and a phrase-based statistical machine translation (PB-SMT) system and even achieves comparable results to those of a Transformer model, without the need for a large-scale training model.

1 Introduction

Distinguished from other techniques, EBMT is a method of machine translation in that translations take place by example. It employs a case-based reasoning strategy, where translations are generated by giving a set of sentences in the source language and other example sentences contained in the knowledge database.

Analogy is a way to implement reasoning. An analogical equation $A : B :: C : D$ expresses a relationship between four objects that pronounces as the following: A is to B as C is to D. Analogy has the ability to reason so that we can interpret words or sentences that we are not familiar with.

An EBMT system by analogy has been proposed in [1, 2], which is called `Beth`. This system requires analogies between sentences to perform case-based reasoning. The existing technique (`nlg`[1)] [3]) extracts sentence analogies only on the formal level. In this work, we propose to extract analogies at the syntactic level by using parse trees and use them in an EBMT system by analogy. Such analogies are shown to bring improvement.

1) http://lepage-lab.ips.waseda.ac.jp/media/filer_public/64/52/64528717-c3ce-4617-8208-c1fb70cf1442/nlg-v321.zip

2 Related Work

2.1 Types of Analogy

Analogy is usually categorized as formal analogy and semantic analogy. For formal analogy, we do not consider the meaning of the terms or the syntax of the sentences. We only care about the surface form of the strings, such as characters or words. Take (1) as an example.

$$\textit{I talk to him.} : \begin{matrix}\textit{I talked}\\ \textit{to him.}\end{matrix} :: \textit{I go to school.} : x \qquad (1)$$

On the level of form, the solution of (1) is: x = *I goed to school.* On that level, the only changes allowed are between characters. Whether the sentence is grammatically correct or makes sense is not taken into account.

In contrast to formal analogies, semantic analogies take into account the meaning encapsulated in the words or sentences. Thus, on the semantic level, the solution to (1) is x = *I went to school.* Here, the meaning attached to the strings is taken into account.

2.2 EBMT by Analogy

Formula (2) defines bilingual analogies between sentences in two languages, which are used by the EBMT system by analogy in [1]. $A : B :: C : D$ denotes a monolingual analogy in the source language and $A' : B' :: C' : D'$ is its corresponding translation in the target language.

$$\begin{array}{ccccccc} A & : & B & :: & C & : & D \\ \updownarrow & & \updownarrow & & \updownarrow & & \updownarrow \\ A' & : & B' & :: & C' & : & D' \end{array} \qquad (2)$$

Suppose that we want to get the translation D' of D. During the reasoning process, the Beth system retrieves three (problem, solution) cases (A, A'), (B, B'), (C, C') in which "A is to B as C is to problem D". The solution of the analogical equation "A' is to B' as C' is to x", $x = D'$, is the translation of D. It is clear from the above that the quality of analogies between sentences extracted from the corpus is critical for EBMT by analogy.

3 Analogy on the Level of Syntax

In this work, we propose to extract analogy at the syntactic level. Figure 1 shows an example of an analogy between sentences that happens on the level of syntax. There is no obvious interpretation to enforce analogy on both the level of form and meaning attached to these sentences. However, we can observe that there is an analogy on the level of syntax by considering their syntactic representations. These sentences form an analogy at the syntactic level when looking at their parse trees[2]: from personal pronoun (PRP) to proper noun (NNP).

3.1 Tree Representation

To make the syntax of a sentence distinct, we first use dependency representation to discern essential information. All of the sentences contained in the corpus are converted into their dependency parse trees using the Universal Dependency parsers provided by spaCy[3] library. A sentence S is then represented by a feature vector $\overrightarrow{T_S}$ by counting the number of occurrences for all the branches found in its parse tree T_S. See Formula (3).

$$\overrightarrow{T_A} = \begin{pmatrix} |T_A|_{VBD \to PRP} \\ |T_A|_{VBD \to NNP} \\ \vdots \\ |T_A|_{VBD \to .} \end{pmatrix} \tag{3}$$

3.2 Ratios between Trees

In Formula (4), the ratio between two sentences A and B is defined as the difference between their vectors of syntactic features derived from their parse trees T_A and T_B. To extract analogies at the syntactic level, we check the conformity of an analogy $A : B :: C : D$. Formula (5) defines it. By doing this, we are able to extract sentence analogies from the corpus.

$$A : B \triangleq \overrightarrow{T_A} - \overrightarrow{T_B} = \begin{pmatrix} |T_A|_{VBD \to PRP} - |T_B|_{VBD \to PRP} \\ |T_A|_{VBD \to NNP} - |T_B|_{VBD \to NNP} \\ \vdots \\ |T_A|_{VBD \to .} - |T_B|_{VBD \to .} \end{pmatrix} \tag{4}$$

$$A : B :: C : D \quad \overset{\Delta}{\Longleftrightarrow} \quad \overrightarrow{T_A} - \overrightarrow{T_B} = \overrightarrow{T_C} - \overrightarrow{T_D} \tag{5}$$

3.3 Analogical Cluster

In [4], an analogical cluster is defined as a set of pairs of sentences with exactly the same ratio, as shown in Formula (6). With analogical cluster, we can measure how regular the transformations between sentences are. In particular, we use the nlg package to extract analogical clusters between sentences, at the level of syntax.

$$\begin{matrix} A_1 : B_1 \\ A_2 : B_2 \\ \vdots \\ A_n : B_n \end{matrix} \quad \overset{\Delta}{\Longleftrightarrow} \quad \forall (i, j) \in \{1, \ldots, n\}^2, \;\; A_i : B_i :: A_j : B_j \tag{6}$$

4 Experiments

4.1 Datasets

We use the English-French language pair from the Tatoeba[4] corpus and we perform translations from French to English. To simulate the low-resource setting, we use only 114,151 sentence pairs and randomly divide the entire data into three sets: training set (90%), validation set (9%), and test set (1%). Some statistics are shown in Table 1. The sentences contained in the corpus are very short, with an average of 7 words per sentence.

4.2 Metrics

We evaluate the translation quality automatically by comparing the output of the translation with the reference sentence in the test set. We apply four different metrics as follows.

2) In the parse trees, the terminals, which, by definition, appear on the leaves, are not considered.

3) https://spacy.io/

4) https://tatoeba.org/

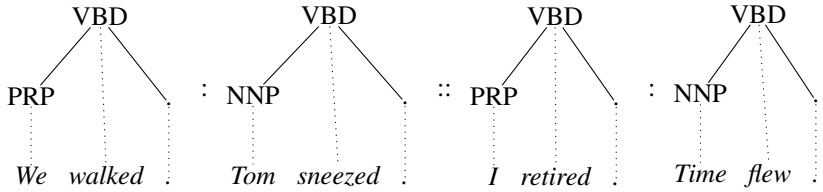

Figure 1 Analogy between sentences on the level of syntax using dependency representation

Table 1 Statistics on the data set used

Language	Number of			Avg. sentence length	
	sentences	tokens	types	in char	in word
en	114,151	760,274	11,372	27.87 ± 7.88	6.97 ± 1.77
fr	114,151	792,621	21,532	32.39 ± 9.31	7.76 ± 2.11

BLEU (BiLingual Evaluation Understudy) [5] evaluates the similarity between the translated sentence and the reference sentence. It has a scale of 0 to 100. The similarity between the two sentences increases with increasing BLEU scores. We use the implementation of `SacreBLEU`[5] in [6].

CHRF (Character n-gram F-score) [7] uses character n-gram F-score to automatically assess the result of machine translation. CHRF score is between 0 and 100. The quality of the translation will be better if it is higher.

TER (Translation Error Rate) counts the number of edit operations required to convert the translated sentence into the reference one. We report TER scores between 0 and 100. The lower the number, the better.

The Levenshtein Edit Distance[6] [8] counts the minimal number of three different edit operations (insertions, deletions and substitutions) between a translation result and a reference. Similarly, the lower, the better.

4.3 Baselines

We assess the performance of the EBMT system by analogy by comparing the translation results to those of other systems. Here, we give a brief overview of the NMT and PB-SMT systems used for experimentation.

We utilize the `OpenNMT`[7] toolkit [9] to build our NMT systems. We experiment with RNN [10] models and Transformer [11] models. Bidirectional RNN is used as an encoder to design the RNN system and both the encoder and the decoder are 6 layers in size. Also, we use a 4-layer Transformer and follow the set-up recommendations in [11] to construct the Transformer system. The early stopping criteria are used for both RNN and Transformer models. These above-mentioned NMT systems are trained from scratch on the same training set that simulates a low-resource scenario.

Our PB-SMT system is constructed using the `Moses`[8] toolkit [12]. We train a 3-gram language model with `KenLM`[9] [13] and smooth it with modified Kneser-Ney smoothing [14]. `GIZA++`[10] included in `Moses` is applied as an alignment tool to generate bilingual phrase tables. Similar to NMT systems, the PB-SMT system is built using sentences from our training set and does not rely on any external data.

4.4 Using Analogies at the Level of Syntax in Example-Based Machine Translation by Analogy

We examine whether analogical clusters extracted by syntactic information indeed increase the quality of machine translation by conducting experiments with the EBMT system by analogy proposed in [1]. The core idea of EBMT is that translation is carried out by comparing a given target sentence with the existing cases in the knowledge database. Analogies are used to perform the reasoning process in `Beth` (See Section 2.2). It is also possible to compile analogical clusters into the `Beth` system in advance. With the help of this retrieval knowledge, we can accelerate the selection of the most similar cases.

Particularly, we first conduct experiments with the original `Beth` without using any analogical clusters. Then, we extract analogical clusters on two levels: (1) considering characters (`char`) only, (2) combining characters and parse trees (`char` ∩ `tree`) together. We can eliminate some ana-

5) https://github.com/mjpost/sacrebleu
6) https://github.com/roy-ht/editdistance
7) https://opennmt.net/
8) http://www2.statmt.org/moses/
9) https://github.com/kpu/kenlm
10) https://github.com/moses-smt/giza-pp

Table 2 Translation results of different systems on the test set (fr → en)

System	Analogical clusters	Size (Mb)	BLEU	CHRF	TER	Edit distance	
						in char	in word
RNN	-	458	33.96	46.09	42.87	11.83	3.02
Transformer	-	511	**53.42**	**62.57**	**29.61**	**8.42**	**2.13**
PB-SMT	-	28	39.96	64.53	35.24	9.92	3.34
Beth	without clusters	0	44.61	54.09	39.13	10.31	2.81
	char	4	52.48	61.52	33.36	8.76	2.42
	char ∩ tree	36	**53.55**	**61.92**	**32.65**	**8.54**	**2.35**

logical clusters that are only character transformations by extracting them from the intersection of on the levels of characters and parse trees. We expect that involving linguistic information, such as parse trees, will improve case reliability. Notice that, we only keep the first 3,000 analogical clusters with the largest number of ratios due to the distribution of the number of analogical clusters with the same size.

4.5 Translation Quality

Table 2 displays the results that we obtained with all the different systems mentioned above. The translation quality of the original Beth is already reasonable, with a BLEU score of 44.61. This BLEU score far outperforms RNN's 33.96 and PB-SMT's 39.96. It is worth mentioning that because this EBMT system requires no linguistic knowledge, it can be applied to any language pair. Furthermore, by adding analogies both on the level of syntax and form, we achieve a BLEU score of 53.55 in the EBMT system by analogy. The Transformer model has a BLEU score similar to our proposed method, but it has a lower TER and edit distance, implying a smaller gap between the translated sentence and the reference sentence.

4.6 Model Size vs. Translation Quality

We also discuss the trade-off between model size and translation quality. We compute the number of parameters that are trainable in RNN and Transformer models. We calculate the size of the KenLM language model and generated phrase tables for the PB-SMT system. For the original Beth system, there is no need for extra data. In the experiments involving analogical clusters in advance as retrieval knowledge, we count the size of analogy clusters extracted from different levels.

As shown in Figure 2, although the performance of a Transformer model slightly exceeds our proposal, the model size of a base Transformer is already 511 Mb. From this perspective, we can say that our EBMT by analogy is a lightweight approach that is more effective for low-resource machine translation.

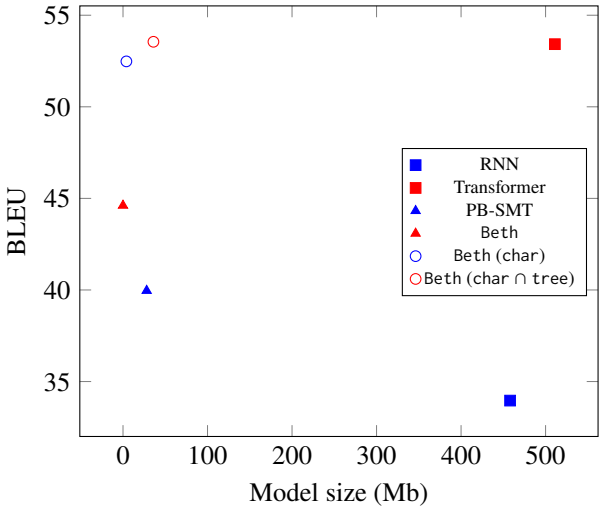

Figure 2 Model size and BLEU scores for different systems

5 Conclusion

In this work, we worked on an approach to extract analogies between sentences at the syntactic level by using parse trees. We were able to extract sentence analogies both on the level of form and the level of syntax. We showed that using analogies at the syntactic level has a positive impact on the translation quality of an EBMT system by analogy. By contrasting the translation results with existing baselines, our approach outperformed an RNN model and a PB-SMT system by showing higher BLEU scores in a low-resource scenario. Although we got similar evaluation results to those of a Transformer model, our approach still has advantages in model size since we do not need pre-training of a large model.

References

[1] Yves Lepage and Jean Lieber. Case-based translation: First steps from a knowledge-light approach based on analogy to a knowledge-intensive one. In **Case-Based Reasoning Research and Development - 26th International Conference, ICCBR 2018, Proceedings**, pp. 563–579. Springer Verlag, 2018.

[2] Yves Lepage and Jean Lieber. An approach to case-based reasoning based on local enrichment of the case base. In **Case-Based Reasoning Research and Development - 27th International Conference, ICCBR 2019, Proceedings**, pp. 235–250. Springer Verlag, 2019.

[3] Rashel Fam and Yves Lepage. Tools for the production of analogical grids and a resource of n-gram analogical grids in 11 languages. In **Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)**, Miyazaki, Japan, 2018. European Language Resources Association (ELRA).

[4] Yves Lepage and Chooi Ling Goh. Towards automatic acquisition of linguistic features. In **Proceedings of the 17th Nordic Conference of Computational Linguistics (NODALIDA 2009)**, pp. 118–125, Odense, Denmark, 2009. Northern European Association for Language Technology (NEALT).

[5] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In **Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics**, pp. 311–318, Philadelphia, Pennsylvania, USA, 2002. Association for Computational Linguistics.

[6] Matt Post. A call for clarity in reporting BLEU scores. In **Proceedings of the Third Conference on Machine Translation: Research Papers**, pp. 186–191, Brussels, Belgium, 2018. Association for Computational Linguistics.

[7] Maja Popović. chrF: character n-gram F-score for automatic MT evaluation. In **Proceedings of the Tenth Workshop on Statistical Machine Translation**, pp. 392–395, Lisbon, Portugal, 2015. Association for Computational Linguistics.

[8] V.I. Levenshtein. Binary codes capable of correcting deletions, insertions and reversals. **Soviet Physics-doklady**, Vol. 10, No. 8, pp. 707–710, February 1966.

[9] Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. OpenNMT: Open-source toolkit for neural machine translation. In **Proceedings of ACL 2017, System Demonstrations**, pp. 67–72, Vancouver, Canada, July 2017. Association for Computational Linguistics.

[10] Thang Luong, Hieu Pham, and Christopher D. Manning. Effective approaches to attention-based neural machine translation. In **Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing**, pp. 1412–1421, Lisbon, Portugal, September 2015. Association for Computational Linguistics.

[11] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, L ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In **Advances in Neural Information Processing Systems**, Vol. 30. Curran Associates, Inc., 2017.

[12] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. Moses: Open source toolkit for statistical machine translation. In **Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions**, pp. 177–180, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

[13] Kenneth Heafield, Ivan Pouzyrevsky, Jonathan H. Clark, and Philipp Koehn. Scalable modified Kneser-Ney language model estimation. In **Proceedings of ACL 2013**, pp. 690–696, Sofia, Bulgaria, August 2013. Association for Computational Linguistics.

[14] R. Kneser and H. Ney. Improved backing-off for m-gram language modeling. In **1995 International Conference on Acoustics, Speech, and Signal Processing**, Vol. 1, pp. 181–184 vol.1, 1995.