

A Framework for Neural Machine Translation by Fuzzy Analogies

Liyan Wang^{1,*}, Bartholomäus Wloka² and Yves Lepage¹

¹Waseda University, Kitakyushu, 808-0135, Japan

²University of Vienna, Vienna, 1190, Austria

Abstract

This paper introduces a novel translation technique, driven by modeling fuzzy analogies that capture approximate conformity to parallel transformations between fragments in sentences. We conduct preliminary experiments on English-Japanese translations with a data set of limited size. The results show the potential of using fuzzy analogies for translation, achieving an increase of about 6 BLEU points compared to NMT.

Keywords

Machine translation, Fuzzy analogy, Limited data

1. Introduction

Low resource settings pose significant challenges to modern Machine Translation (MT) systems [1, 2]. Neural MT (NMT) with large-scale models require large amounts of parallel data to fine-tune learnt weights of two language spaces [3]. MT by analogy (i.e. example-based MT) [4, 5], enables tracing translations by structuring knowledge from examples. It relies on strict analogies that involve ratios with the exact same transformation rule [6]. However, finding sentence analogies with strictness on form can be difficult, particularly in cases where there are less correlated sentences in relatively small sized corpora. In this paper, we propose to explore partial analogies between sentences, which capture approximate conformity between ratios relying on fuzzy matches, i.e., ratios which are partial transformations are matched. For example, *I feel ridiculous.* : *That is untrue.* :: *I feel funny.* : *That is funny.* is a quadruple that captures parallel transformation on sentence fragments. We call this **fuzzy analogy**.

2. Methodology

The proposed method is built on the indirect paradigm of example-based MT in [5]. Similar to this, given translation queries D , we first construct sentence analogies as $A : B :: C : D$,

IARML@IJCAI'2023: Workshop on the Interactions between Analogical Reasoning and Machine Learning, at IJCAI'2023, August, 2023, Macao, China


*Corresponding author.

✉ wangliyan0905@toki.waseda.jp (L. Wang); bartholomaeus.wloka@univie.ac.at (B. Wloka);

yves.lepage@waseda.jp (Y. Lepage)

🆔 0000-0002-9561-5037 (L. Wang); 0000-0002-7484-878X (B. Wloka); 0000-0002-3059-4271 (Y. Lepage)

© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

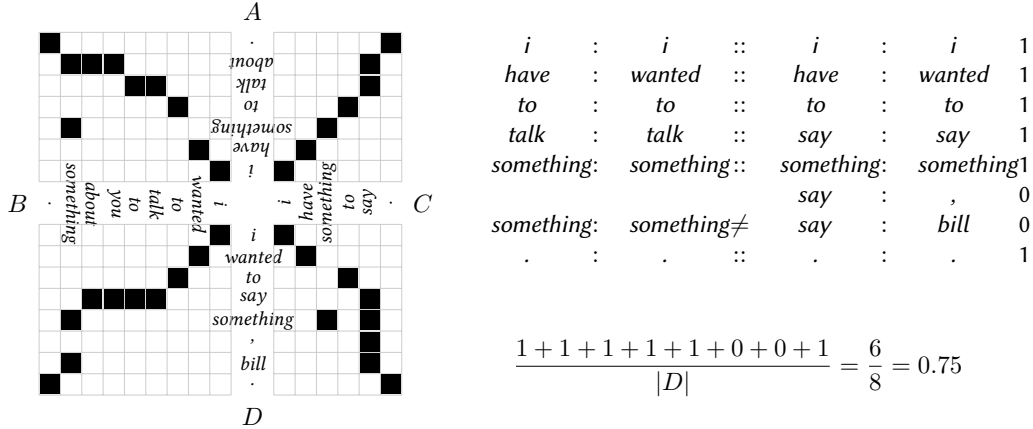


Figure 1: Computation of analogical score. Black cells in matrices indicate alignment points between tokens. Each token quadruple linked by alignments, one for each (sub-)word in D , is checked for trivial analogy. We divide by the length of D to get the analogical score. If $A : B :: C : D$ is a strict analogy, the score is 1.

where A , B , and C are source examples retrieved from translation memory, that will maximize analogical score with D . By looking up the annotated translations of (A, B, C) , we can obtain corresponding analogical equations in the target language. Following this, we exploit a previously learnt model to generate solutions of target analogies as translation results, i.e., $A' : B' :: C' : x \Rightarrow x = D'$.

To **retrieve** sentence analogies, we first pre-compute candidate pools for terms A , B , and C by collecting the k nearest neighbors of D using cosine similarity between sentence embeddings. Theoretically, there will be a cubic number of possible combinations of sentence quadruples (A, B, C, D) . To reduce the computational cost, we prune candidate quadruples. We leave out the quadruples with no lexical overlap between A and C , and between B and D . Finally, for each D , we rank the quadruples by analogical score, and select the first n ones. As in [7], we use alignments between (A, B, C, D) considered as sequences of (sub-)words. We count the number of trivial analogies of the form $a : a :: b : b$ or $a : b :: a : b$ for every aligned (sub-)word quadruple. Figure 1 illustrates the computation of analogical scores.

Next, we **train** a sequence-to-sequence model to solve analogies, so as to derive translation answers. Suppose $A : B :: C : D$ and $A' : B' :: C' : D'$ are a retrieved source analogy and its corresponding translation. We concatenate 7 sentences (excluding D') in two monolingual analogies as input X , to train the model to generate the solution D' by optimizing cross-entropy (CE) between probability distributions conditional on the context of input and preceding target tokens:

$$\mathcal{L}_{CE} = - \sum_{i=1}^{|D'|} \log P(D'_i | D'_{<i}, X) \tag{1}$$

To encourage the model to be more confident in reconstructing target fragments that are in analogical relationships, while being flexible to non-analogical relationships, we introduce a

weighting scalar in (1). Formally, the aim is to minimize weighted CE (WCE):

$$\mathcal{L}_{\text{WCE}} = - \sum_{i=1}^{|D'|} w_i \log P(D'_i | D'_{<i}, X) \quad (2)$$

where w_i takes the value of 1 for trivial analogies, and 0.5 else. For each target token, a weighted value is determined by its corresponding aligned token in D .

3. Preliminary Experiments

3.1. Datasets

We experiment with parallel sentences from the Japanese-English Subtitle Corpus¹, with 50,000 pairs for training, 2,000 for validation, and 2,000 for test. In this work, we primarily investigate the translation quality from English to Japanese. The source sentences contain approximately nine words on average. For each data set, we take source sentences as queries and look for fuzzy analogies from the source part of translation memory (i.e., the training set). The strictness in analogies depends on how closely the queries match the examples in memory. We assess the closeness between the data sets and the memory by computing the similarity using the length of longest common subsequence between sentences at the word level. Specifically, we compare the query sentence to the twenty most similar examples in the memory, excluding itself in the case of the training set. Table 1 shows the statistics of three data sets. On average, the three data sets exhibit similar characteristics, where source sentences are found to have an overlap of four words with their corresponding similar sentences in the memory.

Table 1

Data statistics for the English-to-Japanese translation task, specifically pertaining to the source side of the data sets. Closeness to memory indicates the average number of words that overlap with each of the twenty most similar examples retrieved from the memory.

	Training	Validation	Test
Number of parallel sentences	50,000	2,000	2,000
Sentence length	9 ± 3	9 ± 3	9 ± 3
Number of word types	24,689	3,425	3,348
Closeness to memory	3.89 ± 0.96	3.89 ± 0.95	3.91 ± 0.94

3.2. Implementation Details

In order to retrieve analogies from the corpus, we first use a Sentence-BERT [8] model² to represent sentences as vectors. Subsequently, for each query D , we collect twenty examples as the candidates of B and C , which are the nearest neighbors to D in the embedding space.

¹<https://nlp.stanford.edu/projects/jesc/>

²https://www.sbert.net/docs/pretrained_models.html

Sentences A are selected from the twenty closest neighbors to each candidate for B . We pre-tokenize sentences into sub-words using a SentencePiece [9] model with the vocabulary size of 250,000³. We then enumerate (A, B, C, D) from collected candidates and filter possible quadruples by the overlap constraint between A and C , and between B and D at the sub-word level. Next, We use mGIZA [10] and Moses⁴ to estimate sub-sentential alignments. Based on that, we compute analogical score for each possible quadruple. For each D , we select one fuzzy analogy for translation.

To learn from analogy, we fine-tune a pre-trained mBART [3] model⁵ on fuzzy analogies that are retrieved from the training set. We utilize the large-scale mBART model consisting of a 12-layer encoder and a 12-layer decoder. The target sentences are generated using a beam size of 5 during decoding. To fine-tune the model, we freeze the encoder part and update the parameters of the last 6 layers of the decoder. The frozen model is trained using a batch size of 8 for a maximum of 20 epochs. In the case there are no improvements for three consecutive epochs, we halt the training process before completing all the epochs (early stopping). Finally, we save the model that demonstrates the best performance on the validation set.

3.3. Results and Analysis

We compare to an NMT system by fine-tuning the same pre-trained mBART model on the data sets of parallel sentences. The baseline NMT model is trained using the consistent settings as described above. On 50,000 parallel sentences, NMT obtains a BLEU score of only 2.9. Our system using (1) achieved an improvement of 5.6 and the use of (2) leads to a further gain of about 0.4 BLEU points. Even though fuzzy analogies relax the strictness, the inclusion of partial evidence in parallel transformations still helps in deducing possible translation.

In the retrieved analogies, query sentences are covered by examples under the analogy constraint to different extents with analogical scores ranging from 0 to 1. Figure 2 shows the number of fuzzy analogies constructed for the sentences in the three data sets, categorized by their respective scores. In general, three sets of analogy data demonstrate a comparable distribution in the extent of fuzzy matches between sentence transformations. The majority of analogies fall within the score range of 0.3 to 0.7. This indicates that approximately 30%-70% of tokens in query sentences are associated with examples in the analogy relationship.

Next, we examine the model performance in inferring translation answers by solving fuzzy analogies with different scores. Figure 3 shows that our model is capable of reasoning analogies with lower scores, where less than half of a query sentence is linked to translation examples through analogical associations. This suggests that fuzzy analogies can capture relative knowledge of two languages, which can even assist in translating queries that are distant from memory. We also compare to an NMT baseline on translating each test sentence. In Figure 3, blue points (415 out of 2,000) indicate the cases where our model performs worse than NMT in BLEU. Relatively, there are fewer underperforming cases when analogies have higher scores (>0.7). In Table 2, we list examples of two methods in translating sentences that are either close

³To enable the learning model (e.g., mBART) to identify analogical transformations in quadruples, we use the SentencePiece model with the same tokenization as in mBART.

⁴<http://www2.statmt.org/moses/>

⁵<https://huggingface.co/facebook/mbart-large-50>

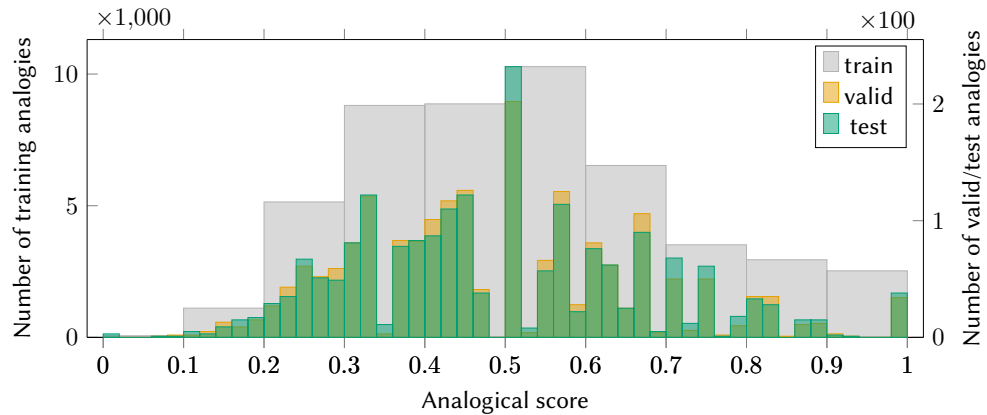


Figure 2: Distribution of analogical scores for fuzzy analogies retrieved from three data sets. Note that there are two different vertical scales: one for training, one for validation and test. The scales for training is ten times more than the second one.

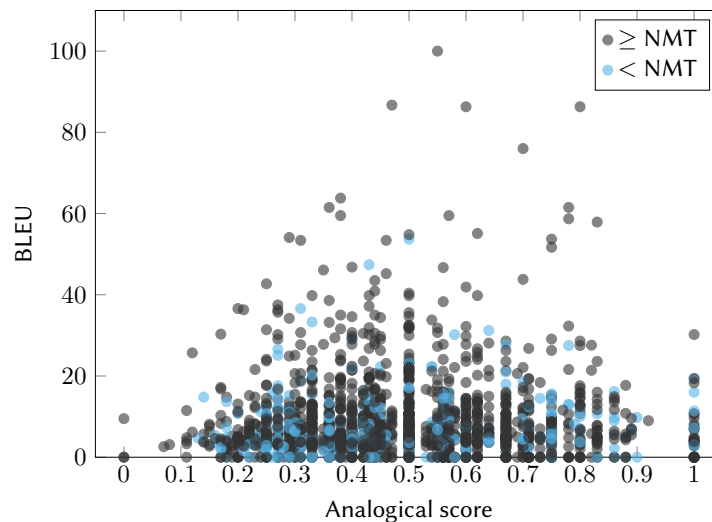


Figure 3: BLEU scores against analogical scores for test analogies. The test cases where our model outperforms or achieves equal performance to NMT are represented by gray points, while the remaining cases are denoted in blue.

to or distant from translation memory.

Do different source analogies constructed for the same query result in diverse translation outputs? We conduct additional experiments to address this question. For each test sentence, we retrieve five fuzzy analogies with the maximum scores and then employ the model trained specifically to handle one analogy per query to solve each of these analogies. Table 3 presents five distinct translations to a query sentence. As shown by the example, it is possible for the model to generate more idiomatic translations that closely convey the intended meaning, using

4. Conclusion and Future Work

In this paper, we introduced a novel translation approach based on the mechanism of using indirect analogies for translation. Unlike the work in [5], we proposed to handle partial analogies that capture approximate conformity between sentence transformations. We call that fuzzy analogies. To solve fuzzy analogies between sentences, we trained an mBART model to generate translations given source quadruples and three known translations in the target analogies. We conducted a comparison between our approach and an NMT baseline under low resource constraints. Additionally, we investigated the impact of analogical quality on translation.

In future work, we will conduct ablation studies to search for optimal configurations for modeling analogies. In addition, we will expand this work to different language pairs and directions, as well as investigate the influence of corpus size on performance.

Acknowledgments

The research reported in this paper was supported in part by a grant for Kakenhi (kiban C) from the Japanese Society for the Promotion of Science (JSPS), n° 21K12038 “Theoretically founded algorithms for the automatic production of analogy tests in NLP”.

References

- [1] R. Aharoni, M. Johnson, O. Firat, Massively multilingual neural machine translation, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 3874–3884. URL: <https://aclanthology.org/N19-1388>. doi:10.18653/v1/N19-1388.
- [2] J. Gu, Y. Wang, Y. Chen, V. O. K. Li, K. Cho, Meta-learning for low-resource neural machine translation, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 3622–3631. URL: <https://aclanthology.org/D18-1398>. doi:10.18653/v1/D18-1398.
- [3] Y. Liu, J. Gu, N. Goyal, X. Li, S. Edunov, M. Ghazvininejad, M. Lewis, L. Zettlemoyer, Multilingual denoising pre-training for neural machine translation, Transactions of the Association for Computational Linguistics 8 (2020) 726–742. URL: <https://aclanthology.org/2020.tacl-1.47>. doi:10.1162/tacl_a_00343.
- [4] M. Nagao, A framework of a mechanical translation between Japanese and English by analogy principle, Artificial and human intelligence (1984) 351–354.
- [5] Y. Lepage, E. Denoual, Purest ever example-based machine translation: Detailed presentation and assessment, Machine Translation 19 (2005) 251–282. doi:10.1007/s10590-006-9010-x.
- [6] P.-A. Murena, M. Al-Ghossein, J.-L. Dessalles, A. Cornuéjols, Solving analogies on words based on minimal complexity transformation, in: C. Bessiere (Ed.), Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20, Interna-

- tional Joint Conferences on Artificial Intelligence Organization, 2020, pp. 1848–1854. URL: <https://doi.org/10.24963/ijcai.2020/256>. doi:10.24963/ijcai.2020/256, main track.
- [7] L. Miclet, S. Bayoudh, A. Delhay, Analogical dissimilarity, *J. Artif. Int. Res.* 32 (2008) 793–824.
- [8] N. Reimers, I. Gurevych, Sentence-BERT: Sentence embeddings using Siamese BERT-networks, in: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Association for Computational Linguistics, Hong Kong, China, 2019, pp. 3982–3992. URL: <https://aclanthology.org/D19-1410>. doi:10.18653/v1/D19-1410.
- [9] T. Kudo, J. Richardson, SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing, in: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 66–71. URL: <https://aclanthology.org/D18-2012>. doi:10.18653/v1/D18-2012.
- [10] Q. Gao, S. Vogel, Parallel implementations of word alignment tool, in: *Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, Association for Computational Linguistics, Columbus, Ohio, 2008, pp. 49–57. URL: <https://aclanthology.org/W08-0509>.

Table 3

Discrepancy between translation answers for the query "*i think it's better if you frankly say your intentions.*" by solving the first five fuzzy analogies with different analogical scores. The reference translation is "あの...ズバツと訴えたい事を伝えたほうがいいと思うんです.". We highlight the answer that closely matches the reference according to the human evaluation.

score	Analogy and solution			BLEU
0.47	<i>i think that's a good idea.</i>	<i>i think that's best.</i>	<i>i think it's great to do what you like.</i>	<i>i think it's better if you frankly say your intentions.</i>
	はい。それがいいかと思ひます。	確かに、それが良いかもな	好きなことを頑張るっていいことだと思いますけど	x
	⇒ $x =$ じゃあはっきり言うなら...。 (well, to put it bluntly...)			1.3
0.40	<i>i guess it's better on my side.</i>	<i>yep, i guess it's probably for the best.</i>	<i>i think that's best.</i>	<i>i think it's better if you frankly say your intentions.</i>
	まだ俺の方がいいや。へへッ。何だよ?これ。痛っ。	ああ、最良だったと思うよ	確かに、それが良いかもな	x
	⇒ $x =$ はっきり言うなら... 意地悪いのは悪いけど。 (to put it bluntly... sorry for being mean.)			2.4
0.40	<i>i think that's best.</i>	<i>yep, i guess it's probably for the best.</i>	<i>i think it's great to do what you like.</i>	<i>i think it's better if you frankly say your intentions.</i>
	確かに、それが良いかもな	ああ、最良だったと思うよ	好きなことを頑張るっていいことだと思いますけど	x
	⇒ $x =$ はっきり言うなら... 意図は正直に言うべきよ (to put it bluntly... you should be honest about your intentions)			2.0
0.40	<i>i guess it's better on my side.</i>	<i>i think that's best.</i>	<i>yep, i guess it's probably for the best.</i>	<i>i think it's better if you frankly say your intentions.</i>
	まだ俺の方がいいや。へへッ。何だよ?これ。痛っ。	確かに、それが良いかもな	ああ、最良だったと思うよ	x
	⇒ $x =$ はっきり言うなら君の意地悪は悪いけどな。 (to put it bluntly, your meanness is bad.)			2.0
0.40	<i>yep, i guess it's probably for the best.</i>	<i>i think that's best.</i>	<i>i guess it's better on my side.</i>	<i>i think it's better if you frankly say your intentions.</i>
	ああ、最良だったと思うよ	確かに、それが良いかもな	まだ俺の方がいいや。へへッ。何だよ?これ。痛っ。	x
	⇒ $x =$ 正直なところ... 意地悪いって言ったら (honestly... if you say mean)			1.8