

Data Augmentation for Machine Translation via Dependency Subtree Swapping

Attila Nagy¹, Dorina Petra Lakatos^{1,2}, Botond Barta^{1,2}, Patrick Nanys¹, Judit Ács²

attila.nagy234@gmail.com, patrick.nanys2000@gmail.com,
{dorinalakatos, botondbarta, acsjudit}@sztaki.hu

¹Department of Automation and Applied Informatics
Budapest University of Technology and Economics

²Institute for Computer Science and Control
Eötvös Loránd Research Network

Abstract. We present a generic framework for data augmentation via dependency subtree swapping that is applicable to machine translation. We extract corresponding subtrees from the dependency parse trees of the source and target sentences and swap these across bisentences to create augmented samples. We perform thorough filtering based on graph-based similarities of the dependency trees and additional heuristics to ensure that extracted subtrees correspond to the same meaning. We conduct resource-constrained experiments on 4 language pairs in both directions using the IWSLT text translation datasets and the Hunglish2 corpus. The results demonstrate consistent improvements in BLEU score over our baseline models in 3 out of 4 language pairs. Our code is available on GitHub¹.

1 Introduction

Parallel data is a necessity for building performant neural machine translation (NMT) systems. For high- and medium-resource languages, millions of parallel data points enable researchers to build translation models of high quality. However, in a resource-constrained setting, such as low-resource or domain-specific machine translation, the lack of data must be compensated with a variety of techniques to improve performance. These include training multilingual models for zero- and few-shot learning (Johnson et al., 2017; Sharaf et al., 2020), transfer learning (Kocmi and Bojar, 2018) and a wide variety of data augmentation techniques that alter existing parallel sentences to create more training data (Sennrich et al., 2016; Zhang and Zong, 2016; Fadaee et al., 2017).

In this paper, we present a framework for generating augmented samples with the help of syntactic information for machine translation. We benchmark our augmentation method in a low-resource setting on 4 language pairs, using low-resource datasets for English-German, English-Hebrew and English-Vietnamese and a subsample of a high-resource dataset for English-Hungarian.

¹ <https://github.com/attilanagy234/syntax-augmentation-nmt>

2 Related work

Data augmentation (DA) involves a set of techniques that enhance the data used to train a machine learning model both in size and variety. It has been widely applied in NLP for fixing class imbalance, mitigating bias, making the model more robust through adversarial examples or increasing model accuracy (Feng et al., 2021). DA methods that incorporate syntactic knowledge have been applied to a number of NLP tasks, but not particularly in the domain of machine translation. For part-of-speech tagging, Şahin and Steedman (2018) augment dependency trees by removing dependency links and moving tree fragments around the root. This dependency tree morphing method was also shown effective for dependency parsing by Vania et al. (2019). Xu et al. (2016) improve relation classification by making use of the directionality of relationships in a dependency tree. The idea of swapping compatible subparts of datapoints to generate augmented samples has shown performance improvements in dependency parsing (Dehouck and Gómez-Rodríguez, 2020), named entity recognition (Dai and Adel, 2020) and constituency parsing (Shi et al., 2020). Shi et al. (2021) introduce a generalized framework for substructure substitution, which produces new samples by swapping same-label substructures.

For machine translation, the most common augmentation method is back-translation (Sennrich et al., 2016), where additional training data is obtained by translating a target-language monolingual corpus into the source language, using a baseline model trained with the originally available data. Fadaee et al. (2017) targets low-frequency words by generating new sentence pairs with rare words in a new context. Wang et al. (2018) propose SwitchOut, a technique, which randomly replaces words in both the source and target sentences with other random words from their corresponding vocabularies. Gao et al. (2019) introduce Soft Contextual DA, where they augment a randomly chosen word in a sentence by its contextual mixture of multiple related words. Nguyen et al. (2020) diversify the training data by using the predictions of multiple forward and backward models and then merging them with the original dataset on which the final NMT model is trained. Moussallem et al. (2019) use knowledge graphs to enhance semantic feature extraction and hence the translation of entities and terminological expressions. Chen et al. (2020) create constraint-aware training data by first randomly sampling the phrases of the reference as constraints, and then packing them together into the source sentence with a separation symbol. Wei et al. (2020) propose an uncertainty-aware semantic DA method, which explicitly captures the universal semantic information among multiple semantically-equivalent source sentences and enhances the hidden representations with this information for better translations. Sánchez-Cartagena et al. (2021) present an approach, where augmented sentence pairs are generated with simple transformations and used as auxiliary tasks in a multi-task learning framework with the aim of providing new contexts where the target prefix is not informative enough to predict the next word. Wei et al. (2022) suggest Continuous Semantic Augmentation (CSANMT), which augments each training instance with an adjacency semantic region that could cover adequate variants of literal expression under the same

meaning. A method, which uses dependency parsing for data augmentation was introduced by Duan et al. (2020). They perform simple token-level manipulations, such as blanking, dropout and replacement and use the depth of tokens in the dependency tree as guidance for computing word-selection probabilities.

3 Methodology

In this section, we introduce the data augmentation and the graph-based filtering methods used in our experiments in detail. Figure 1 gives an overview of the augmentation algorithm starting from pairs of parallel sentences or bisentences.

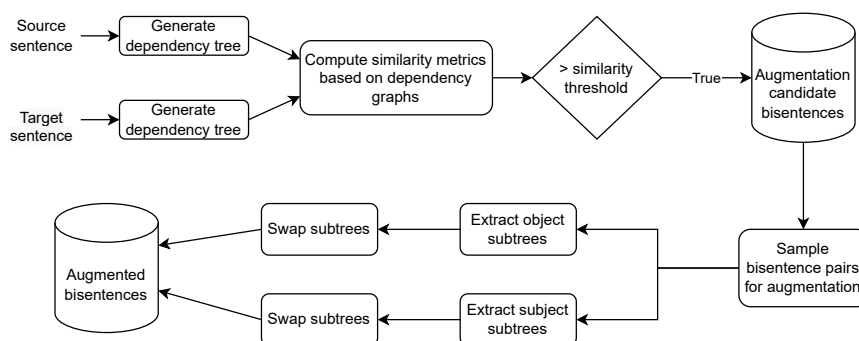


Fig. 1: Overview of our data augmentation method

3.1 Dependency subtree swapping

We propose a simple, yet effective data augmentation method for machine translation based on dependency trees. The intuition behind our approach is that elements of a translation pair - although they are of a different language - show syntactic similarities. The central idea is to extract corresponding syntactic substructures in the source and target sentences and swap these across bisentences to produce augmented datapoints. One clear advantage of this method compared to many other non-model based data augmentation methods for NMT, is that this approach alters the source and target sentences simultaneously, so it has a better chance of maintaining parallelism in the augmented samples. Depending on the language and the phrasing, associated substructures could have different composition, although we hypothesize that subjects and objects have a similar enough representation across many language-pairs to be used for this kind of augmentation. We consider object and subject subtrees in the dependency trees that correspond to OBJ and NSUBJ edges defined in the Universal Dependencies (Nivre et al., 2020). Figure 2 shows an example of the object subtrees in a

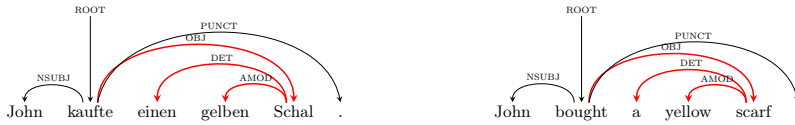


Fig. 2: Example of a object substructures in the dependency trees of a German-English translation pair. The corresponding subtrees are highlighted in red.

German-English translation pair. Figure 3 explains the technique of object and subject subtree swapping through an example bisentence pair.

Applying the method in a more generic setting to more complex syntactic substructures is possible. However, as the performance of the algorithm is dependent on both the translation quality and the quality of the dependency parser, finding correspondance between more complex subtrees could be hard and that may result in injecting too much noise into the training data. Nevertheless, this is an interesting future direction of research, which should be explored.

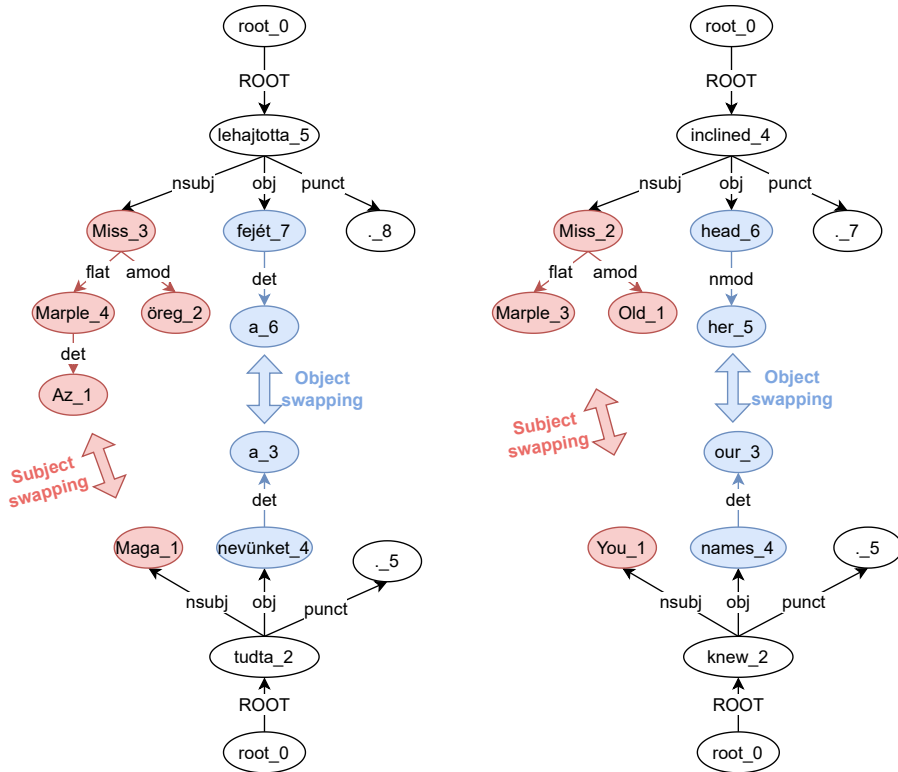


Fig. 3: Two kinds of augmentation techniques: object and subject subtree swapping.

3.2 Generating augmentation candidates

Our prior experiments have shown that although the number of samples that we can generate grows exponentially with respect to the size of the training data, the generated bisentences can be very noisy (Nagy et al., 2022). In this section, we discuss the rules and heuristics that we implemented to ensure that the generated bisentences are high quality.

Filtering heuristics As the augmentation method is based on swapping certain subtrees, it is only possible to sample from a subset of the training data, where the dependency trees contain the required edges. We consider a bisentence eligible for augmentation, if the dependency trees of both the source and target sentences have at least one NSUBJ and one OBJ edges. We also constrain that they cannot contain more than one of each dependency edge type. In order to further reduce noise, we exclude bisentences, where the root of the selected subtrees of the source and target sentences do not belong to the same part-of-speech tag. We also constrain, that every subtree eligible for swapping must contain either a noun or a proper noun.

Graph-based sampling As the method is sensitive to both the translation quality and the performance of the dependency parser, we need to ensure that the selected subtrees capture the same meaning in both the source and target sentences. We hypothesize that if the syntactic composition of the subtrees of the source and target sentences are significantly different, the subtrees likely have a different meaning. To study the syntactic similarities of the dependency trees, we explore two graph-based methods: Graph Edit Distance (GED) (Sanfeliu and Fu, 1983) and Edge Mapping (EM).

Graph Edit Distance is a generalization of the Levenshtein distance (Levenshtein et al., 1966) for graphs. The GED value represents the minimal cost of graph edit operations, which is required to transform the first graph into the second one. The operations we used for nodes and edges are insertion, deletion and substitution. A deletion or an insertion costs 1, while the substitution costs 2. To compare these numbers concerning the size of the graphs, we normalized the edit distance with the following equations:

$$\begin{aligned} d_{\max} &= 2 * |V_1| - 1 + 2 * |V_2| - 1 \\ \text{sim}(G_1, G_2) &= \frac{d_{\max} - \text{GED}(G_1, G_2)}{d_{\max}} \end{aligned} \quad (1)$$

where d_{\max} corresponds to the maximum distance between 2 graphs, when all of the edges and nodes are deleted from G_1 and every node and edge in G_2 is inserted. V_1 and V_2 are the set of vertices in G_1 and G_2 respectively. The $\text{sim}(G_1, G_2)$ is a similarity measure between G_1 and G_2 , and its value is between 0 and 1, 1 if the two graphs are isomorphic.

The Edge Mapping (EM) algorithm is based on a general graph similarity algorithm (Champin and Solnon, 2003). A modified version of this is described in

Algorithm 1 Edge mapping

Require: $G_1(V_1, E_1), G_2(V_2, E_2)$
 $m \leftarrow \{\}$
for all $e_1 \in E_1$ **do**
 $\text{cands} \leftarrow \{e_2 \mid e_2 \in E_2, e_2 \notin m, e_1 = e_2\}$
 if cands is empty **then**
 continue
 end if
 $\text{cands} \leftarrow \arg \max_{c \in \text{cands}} \text{score}(e_1, c)$
 $\text{cands} \leftarrow \arg \max_{c \in \text{cands}} \text{route_sim}(e_1, c)$
 $m[e_1] \leftarrow \text{cands}[0]$
end for
return m

Algorithm 1. The $\text{score}(e_1, e_2)$ represents how many nodes are common between the two edges (2 at max). $\text{route_sim}(e_1, e_2)$ is the Levenshtein distance between the root – e_1 and root – e_2 routes, where every route is defined by the part-of-speech tag of the visited nodes. The algorithm gives a mapping between the edges of the graphs, which we consider as the intersection. This way we can calculate a Jaccard index between the edges, which gives us another similarity measure:

$$J(G_1, G_2) = \frac{|m|}{|E_1| + |E_2| - |m|} \quad (2)$$

where m is the mapping, E_1 and E_2 are the set of edges in G_1 and G_2 respectively. Figure 4 compares the two kinds of similarity measures through the object subtree of a sentence.

4 Experiments

We conduct experiments on 4 language pairs in both directions: English–German, English–Hebrew, English–Vietnamese and English–Hungarian. We trained baseline models for all language pairs, which we benchmark against models trained using additional augmented data.

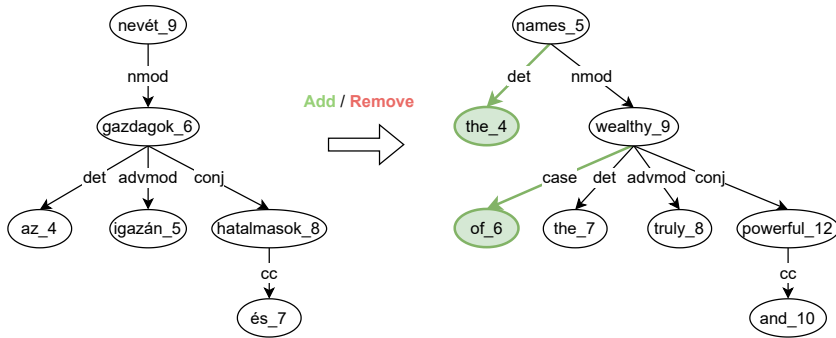
4.1 Datasets

Following other works in the field (Sánchez-Cartagena et al., 2021; Gao et al., 2019; Guo et al., 2020; Wang et al., 2018), we use datasets from the text translation track of IWSLT. The IWSLT datasets are transcripts of Ted and TedX talks and are regularly used for benchmarking data augmentation methods for NMT in a low-resource setting. For English–German and English–Hebrew, we use the IWSLT 2014 (Cettolo et al., 2014) text translation track for training and we use the *tst2013* and *tst2014* datasets for development and test sets respectively.

Ritkán tudod meg **az igazán gazdagok és hatalmasok nevét.**

You seldom learn **the names of the truly wealthy and powerful.**

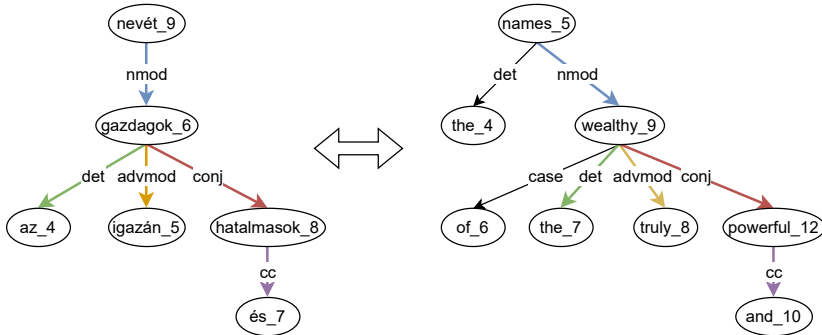
GED



$$\text{GED} = 4 \quad \longrightarrow \quad \text{sim} = (26 - 4) / 26 = 0.92$$

$$d_{\max} = 2 \cdot 6 - 1 - 2 \cdot 8 - 1 = 26$$

EM



$$|\text{mapping}| = 5 \quad \longrightarrow \quad \text{sim} = 5 / (5 + 7 - 5) = 0.71$$

Fig. 4: The two kinds of graph similarity measure of a bisentence’s object subtree. For the GED method the added and removed vertices/edges are colored green and red respectively. In the case of the EM measure, the colors of the edges represent the mapping.

For English-Vietnamese, we select the IWSLT 2015 (Cettolo et al., 2015) text translation track for training, whereas the *tst2012* and *tst2013* datasets are used for development and test sets respectively. For English-Hungarian, we perform experiments on the Hunglish2 corpus (Varga et al., 2007), which is a sentence-aligned dataset. It was constructed by scraping and aligning bilingual text in several domains, such as literature, movie subtitles, software documentation and legal text. To simulate a low-resource experiment, we created a stratified subsample of 250k sentences based on the domains and created a train-dev-test splits from this, also with stratified sampling.

We apply the same preprocessing steps for all language pairs. We remove sentences longer than 32 tokens. We also remove sentences where the difference of the source- and target-side token counts are more than 7 and their ratio is more than 1.2. We also remove leading and trailing quotation marks and dashes. We perform language detection on both the source and target sentences using *fastText* (Joulin et al., 2016) and remove datapoints whenever the languages mismatch with what is expected. We normalize punctuations with *sacremoses*². The statistics of the the preprocessed datasets can be seen in Table 1.

Language pair	train	dev	test
En-De	127,506	993	1,305
En-He	132,105	1,382	962
En-Vi	89,188	1,553	1,268
En-Hu	212,500	37,500	21,700

Table 1: Number of bisentences in the train/dev/test sets for each language pair.

4.2 Training details

We use the same Transformer-based encoder-decoder architecture (Vaswani et al., 2017) for every experiment. The hyperparameters used in each experiment can be seen in Table 2. All of the models were trained on a single A100 GPU for 4 hours using the openNMT framework (Klein et al., 2017). We used early stopping to avoid overfitting on the training set based on the perplexity of the validation set. For tokenization, part-of-speech tagging and dependency parsing, we use HuSpacy (Orosz et al., 2022) for Hungarian and Stanza (Qi et al., 2020) for the rest of the languages.

For each experiment, we generate augmented data and shuffle it with the training sets before training. We explore data generation with different augmentation configurations, including the graph similarity metrics (GED and EM),

² <https://github.com/alvations/sacremoses>

Parameter	Value	Parameter	Value
batch type	tokens	batch size	12288
accumulation count	4	average decay	0.0005
train steps	150000	valid steps	5000
early stopping	4	early stopping criteria	ppl
optimizer	adam	learning rate	2
warmup steps	8000	decay method	noam
adam beta2	0.998	max grad norm	2
label smoothing	0.1	param init	0
param init glorot	true	normalization	tokens
max generator batches	32	encoder layers	8
decoder layers	8	heads	16
RNN size	1024	word vector size	1024
Transformer FF	2096	dropout steps	0
dropout	0.1	attention dropout	0.1
share embeddings	true	position encoding	true

Table 2: Hyperparameters of the models.

the subtree type that is swapped (Subj, Obj) and different augmentation ratios (0.5, 2, 3). For the graph similarity filtering, we use 0.4 as a generic threshold for all language pairs. We found during qualitative analysis of augmented data points that this value yields high quality and diverse results. Selecting a higher threshold value does not necessarily mean better quality augmentations, because in practice, a very high similarity often means that the subtrees are small and this way we cannot guarantee with high confidence that the subtrees correspond to the same meaning. The number of sentence pairs with similarity above a given threshold between the source and target sentences are shown for all language-pairs in Figure 5.

5 Results

We measure the effectiveness of the proposed method with BLEU scores of NMT models trained with and without augmented data. Figure 6 shows the BLEU scores for each dataset and each augmentation setup. It is clear that there is no *one size fits all*, but with the exception of Vietnamese, at least one configuration is always better than the baseline with a non-negligible margin. Unfortunately the Vietnamese dependency parser is not very high quality, which explains the subpar performance after our augmentation. We find an upward trend in the other three language pairs, in other words, more augmented data seems to help, although the exact augmentation ratio needs to be determined separately for

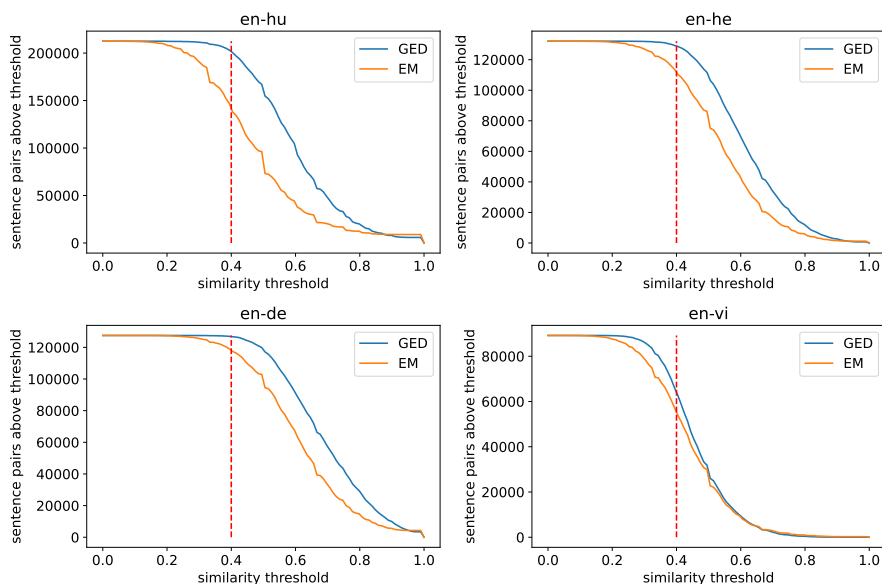


Fig. 5: The number of sentence pairs with similarity above a given threshold for both the GED and EM methods computed between the dependency parse trees of the source and target sentences.

each dataset. Table 3 reports the BLEU scores of the baseline models and the best performing models’ hyperparameters and results.

Language pair	Baseline	Graph sampling	Aug type	Aug ratio	Aug BLEU
En-De	20.1	GED	Subj	3	21.3
De-En	23.0	GED	Obj	3	24.5
En-He	17.1	EM	Obj	3	17.5
He-En	25.2	GED	Obj	2	27.0
En-Vi	21.8	EM	Subj	3	21.6
Vi-En	22.1	EM	Obj	0.5	21.7
En-Hu	18.1	EM	Subj	2	20.0
Hu-En	21.7	EM	Subj	2	23.6

Table 3: Augmentation parameters of the best performing models for each language pair.

In general we observed that the augmentation method starts to improve the performance of the models considerably if the augmentation ratio is 2 or higher.

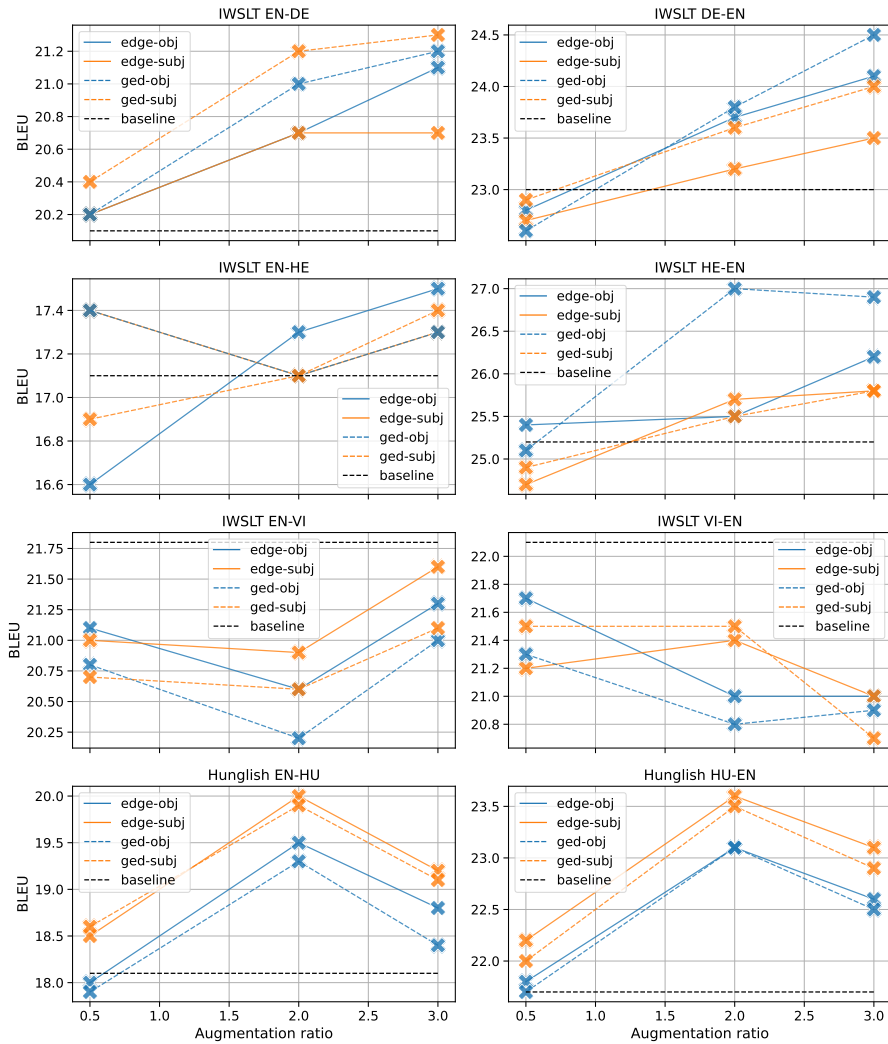


Fig. 6: Test BLEU scores for each augmentation method and augmentation ratio. We indicate the baseline results with a horizontal black dashed line.

This means augmenting a dataset of size X with at least $2X$ amount of new samples. The choice of augmentation type (Object or Subject) seems to depend on the language. Subject is usually better, particularly for Hungarian, but it is less obvious in other languages. The same can be said about the filter type (GED or EM). One trend, we see is that the same augmentation type and ratio usually yields similar results regardless of the filter type. This suggests that the filter methods we use are somewhat robust.

The augmentation methods did not improve the scores for English-Vietnamese in neither directions. We suspect that this is because Stanza has relatively poor performance on tokenization and dependency parsing for Vietnamese³. It would be interesting to reproduce the experiments with a better NLP pipeline for Vietnamese and compare the results.

6 Conclusion

We presented an augmentation method for machine translation based on dependency subtree swapping. We applied two types of graph distance based filtering to discard low quality augmentation. We demonstrated the effectiveness of our method by performing resource-constrained experiments on 4 language pairs in both directions. With the exception of English–Vietnamese, we see consistent improvement over the non-augmented baselines in English–German, English–Hebrew and English–Hungarian.

Acknowledgements

The authors would like to thank András Kornai and Csaba Oravecz for discussions on the augmentation methodology.

³ <https://stanfordnlp.github.io/stanza/performance.html>

References

- Cettolo, M., Niehues, J., Stüker, S., Bentivogli, L., Cattoni, R., Federico, M.: The IWSLT 2015 evaluation campaign. In: Proceedings of the 12th International Workshop on Spoken Language Translation: Evaluation Campaign. pp. 2–14. Da Nang, Vietnam (Dec 3-4 2015), <https://aclanthology.org/2015.iwslt-evaluation.1>
- Cettolo, M., Niehues, J., Stüker, S., Bentivogli, L., Federico, M.: Report on the 11th IWSLT evaluation campaign. In: Proceedings of the 11th International Workshop on Spoken Language Translation: Evaluation Campaign. pp. 2–17. Lake Tahoe, California (Dec 4-5 2014), <https://aclanthology.org/2014.iwslt-evaluation.1>
- Champin, P.A., Solnon, C.: Measuring the similarity of labeled graphs. vol. 2689 (07 2003)
- Chen, G., Chen, Y., Wang, Y., Li, V.O.: Lexical-constraint-aware neural machine translation via data augmentation. In: Bessiere, C. (ed.) Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20. pp. 3587–3593. International Joint Conferences on Artificial Intelligence Organization (7 2020), <https://doi.org/10.24963/ijcai.2020/496>, main track
- Dai, X., Adel, H.: An analysis of simple data augmentation for named entity recognition. In: Proceedings of the 28th International Conference on Computational Linguistics. pp. 3861–3867. International Committee on Computational Linguistics, Barcelona, Spain (Online) (Dec 2020), <https://aclanthology.org/2020.coling-main.343>
- Dehouck, M., Gómez-Rodríguez, C.: Data augmentation via subtree swapping for dependency parsing of low-resource languages. In: Proceedings of the 28th International Conference on Computational Linguistics. pp. 3818–3830. International Committee on Computational Linguistics, Barcelona, Spain (Online) (Dec 2020), <https://aclanthology.org/2020.coling-main.339>
- Duan, S., Zhao, H., Zhang, D., Wang, R.: Syntax-aware data augmentation for neural machine translation. arXiv preprint arXiv:2004.14200 (2020)
- Fadaee, M., Bisazza, A., Monz, C.: Data augmentation for low-resource neural machine translation. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). pp. 567–573. Association for Computational Linguistics, Vancouver, Canada (Jul 2017), <https://aclanthology.org/P17-2090>
- Feng, S.Y., Gangal, V., Wei, J., Chandar, S., Vosoughi, S., Mitamura, T., Hovy, E.: A survey of data augmentation approaches for NLP. In: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. pp. 968–988. Association for Computational Linguistics, Online (Aug 2021), <https://aclanthology.org/2021.findings-acl.84>
- Gao, F., Zhu, J., Wu, L., Xia, Y., Qin, T., Cheng, X., Zhou, W., Liu, T.Y.: Soft contextual data augmentation for neural machine translation. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 5539–5544. Association for Computational Linguistics, Florence, Italy (Jul 2019), <https://aclanthology.org/P19-1555>

- Guo, D., Kim, Y., Rush, A.: Sequence-level mixed sample data augmentation. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 5547–5552. Association for Computational Linguistics, Online (Nov 2020), <https://aclanthology.org/2020.emnlp-main.447>
- Johnson, M., Schuster, M., Le, Q.V., Krikun, M., Wu, Y., Chen, Z., Thorat, N., Viégas, F., Wattenberg, M., Corrado, G., et al.: Google’s multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics* 5, 339–351 (2017)
- Joulin, A., Grave, E., Bojanowski, P., Douze, M., Jégou, H., Mikolov, T.: Fasttext. zip: Compressing text classification models. arXiv preprint arXiv:1612.03651 (2016)
- Klein, G., Kim, Y., Deng, Y., Senellart, J., Rush, A.: OpenNMT: Open-source toolkit for neural machine translation. In: Proceedings of ACL 2017, System Demonstrations. pp. 67–72. Association for Computational Linguistics, Vancouver, Canada (Jul 2017), <https://www.aclweb.org/anthology/P17-4012>
- Kocmi, T., Bojar, O.: Trivial transfer learning for low-resource neural machine translation. In: Proceedings of the Third Conference on Machine Translation: Research Papers. pp. 244–252. Association for Computational Linguistics, Brussels, Belgium (Oct 2018), <https://aclanthology.org/W18-6325>
- Levenshtein, V.I., et al.: Binary codes capable of correcting deletions, insertions, and reversals. In: *Soviet physics doklady*. vol. 10, pp. 707–710. Soviet Union (1966)
- Moussallem, D., Arčan, M., Ngomo, A.C.N., Buitelaar, P.: Augmenting neural machine translation with knowledge graphs. arXiv preprint arXiv:1902.08816 (2019)
- Nagy, A., Nanys, P., Konrád, B.F., Bial, B., Ács, J.: Syntax-based data augmentation for hungarian-english machine translation. arXiv preprint arXiv:2201.06876 (2022)
- Nguyen, X.P., Joty, S., Wu, K., Aw, A.T.: Data diversification: A simple strategy for neural machine translation. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (eds.) *Advances in Neural Information Processing Systems*. vol. 33, pp. 10018–10029. Curran Associates, Inc. (2020)
- Nivre, J., de Marneffe, M.C., Ginter, F., Hajič, J., Manning, C.D., Pyysalo, S., Schuster, S., Tyers, F., Zeman, D.: Universal Dependencies v2: An evergrowing multilingual treebank collection. In: Proceedings of the Twelfth Language Resources and Evaluation Conference. pp. 4034–4043. European Language Resources Association, Marseille, France (May 2020), <https://aclanthology.org/2020.lrec-1.497>
- Orosz, G., Szántó, Z., Berkecz, P., Szabó, G., Farkas, R.: HuSpaCy: an industrial-strength Hungarian natural language processing toolkit (2022)
- Qi, P., Zhang, Y., Zhang, Y., Bolton, J., Manning, C.D.: Stanza: A Python natural language processing toolkit for many human languages. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations (2020), <https://nlp.stanford.edu/pubs/qi2020stanza.pdf>

- Şahin, G.G., Steedman, M.: Data augmentation via dependency tree morphing for low-resource languages. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. pp. 5004–5009. Association for Computational Linguistics, Brussels, Belgium (Oct-Nov 2018), <https://aclanthology.org/D18-1545>
- Sánchez-Cartagena, V.M., Esplà-Gomis, M., Pérez-Ortiz, J.A., Sánchez-Martínez, F.: Rethinking data augmentation for low-resource neural machine translation: A multi-task learning approach. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. pp. 8502–8516. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic (Nov 2021), <https://aclanthology.org/2021.emnlp-main.669>
- Sanfeliu, A., Fu, K.S.: A distance measure between attributed relational graphs for pattern recognition. *IEEE Transactions on Systems, Man, and Cybernetics SMC-13(3)*, 353–362 (1983)
- Sennrich, R., Haddow, B., Birch, A.: Improving neural machine translation models with monolingual data. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 86–96. Association for Computational Linguistics, Berlin, Germany (Aug 2016), <https://aclanthology.org/P16-1009>
- Sharaf, A., Hassan, H., Daumé III, H.: Meta-learning for few-shot nmt adaptation. arXiv preprint arXiv:2004.02745 (2020)
- Shi, H., Livescu, K., Gimpel, K.: On the role of supervision in unsupervised constituency parsing. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 7611–7621. Association for Computational Linguistics, Online (Nov 2020), <https://aclanthology.org/2020.emnlp-main.614>
- Shi, H., Livescu, K., Gimpel, K.: Substructure substitution: Structured data augmentation for NLP. In: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. pp. 3494–3508. Association for Computational Linguistics, Online (Aug 2021), <https://aclanthology.org/2021.findings-acl.307>
- Vania, C., Kementchedjheva, Y., Søgaard, A., Lopez, A.: A systematic comparison of methods for low-resource dependency parsing on genuinely low-resource languages. 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). pp. 1105–1116. Association for Computational Linguistics, Hong Kong, China (Nov 2019), <https://aclanthology.org/D19-1102>
- Varga, D., Halácsy, P., Kornai, A., Nagy, V., Németh, L., Trón, V.: Parallel corpora for medium density languages. *Amsterdam Studies In The Theory And History Of Linguistic Science Series 4* 292, 247 (2007)
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in neural information processing systems. pp. 5998–6008 (2017)
- Wang, X., Pham, H., Dai, Z., Neubig, G.: SwitchOut: an efficient data augmentation algorithm for neural machine translation. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. pp. 856–

861. Association for Computational Linguistics, Brussels, Belgium (Oct-Nov 2018), <https://aclanthology.org/D18-1100>
- Wei, X., Yu, H., Hu, Y., Weng, R., Luo, W., Jin, R.: Learning to generalize to more: Continuous semantic augmentation for neural machine translation. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, ACL 2022 (2022)
- Wei, X., Yu, H., Hu, Y., Weng, R., Xing, L., Luo, W.: Uncertainty-aware semantic augmentation for neural machine translation. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 2724–2735. Association for Computational Linguistics, Online (Nov 2020), <https://aclanthology.org/2020.emnlp-main.216>
- Xu, Y., Jia, R., Mou, L., Li, G., Chen, Y., Lu, Y., Jin, Z.: Improved relation classification by deep recurrent neural networks with data augmentation. In: Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers. pp. 1461–1470. The COLING 2016 Organizing Committee, Osaka, Japan (Dec 2016), <https://aclanthology.org/C16-1138>
- Zhang, J., Zong, C.: Exploiting source-side monolingual data in neural machine translation. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. pp. 1535–1545. Association for Computational Linguistics, Austin, Texas (Nov 2016), <https://aclanthology.org/D16-1160>