Syntax-based data augmentation for Hungarian-English machine translation

Attila Nagy[1], Patrick Nanys[1], Balázs Frey Konrád[1], Bence Bial[1], Judit Ács[2]

[1]Department of Automation and Applied Informatics
Budapest University of Technology and Economics
[2]Institute for Computer Science and Control
Eötvös Loránd Research Network

Abstract. We train Transformer-based neural machine translation models for Hungarian-English and English-Hungarian using the Hunglish2 corpus. Our best models achieve a BLEU score of 40.0 on Hungarian-English and 33.4 on English-Hungarian. Furthermore, we present results on an ongoing work about syntax-based augmentation for neural machine translation. Both our code and models are publicly available[1].

1 Introduction

Machine Translation (MT) is a subfield of natural language processing, where the task is to perform translation automatically from one language to another. To be able to translate corpora between arbitrary languages, one needs to develop a deep understanding of the underlying structure of language and for this reason MT has been considered one of the hardest problems in NLP.

Our contributions with this work are twofold. First, we train neural machine translation models for English-Hungarian and Hungarian-English using Transformer models. Second, we propose a new data augmentation technique for machine translation using syntactic parsing. To the best of our knowledge, no prior work has been published on Transformer-based machine translation for Hungarian-English.

2 Related work

Early approaches tried to model translation by deriving translation rules based on our knowledge of linguistics. A rule-based method, however, is insufficient for covering the countless edge cases in language. With the increasingly available parallel datasets, data-driven approaches gained dominance in the previous decades. Statistical machine translation (SMT) (Koehn et al., 2003; Brown et al., 1990) outperforms rule-based methods by learning latent structures in the data with the help of statistical methods. Although better than its predecessor, SMT struggles to capture long-term dependencies. Neural machine translation (NMT)

[1] https://github.com/attilanagy234/syntax-augmentation-nmt

(Bahdanau et al., 2014; Cho et al., 2014) tackles this problem by modeling translation as an end to end process using neural networks. In current machine translation research, the Transformer architecture (Vaswani et al., 2017) is almost exclusively used in supervised settings (Tran et al., 2021; Germann et al., 2021; Oravecz et al., 2020). For the Hungarian-English language-pair, published methods followed the same evolution of rule-based systems (Prószéky and Tihanyi, 2002), statistical methods (Laki et al., 2013) and neural models (Tihanyi and Oravecz, 2017).

Data augmentation is particularly important in machine translation research, because many language-pairs have insufficient resources to build complex models. Classical augmentation methods that are used in NLP are hard to apply to machine translation, because it is very hard to augment both the source and target sentence such that the parallelism of the sentence-pair holds. Wang and Yang (2015) stochastically select words for replacement based on a a distance metric in an embedding space. Kobayashi (2018) train a language model to predict new words based on its surrounding context and use this model to replace words. Xie et al. (2017) avoid overfitting to specific contexts by randomly replacing words in the training data with a blank token. Methods that use the dependency parse tree of a sentence for augmentation were proved useful in a number of tasks such as word relation classification (Xu et al., 2016), POS tagging (Şahin and Steedman, 2019) and dependency parsing (Vania et al., 2019; Dehouck and Gómez-Rodríguez, 2020). Duan et al. (2020) use the depth of words in a dependency parse tree as a clue of importance for selecting words for augmentation in machine translation. The syntax-aware data augmentation that we discuss in this work was first proposed by Nagy (2021). In machine translation, backtranslation is the most common data augmentation method, which creates pseudo-parallel sentences from monolingual data using a baseline translation model (Sennrich et al., 2015).

3 Methodology

We discuss two main experiments in this work. Firstly, we train competent neural machine translation models based on state-of-the-art architectures for HU-EN and EN-HU and provide a solid baseline for future NMT research in Hungarian by releasing the trained model. Secondly, we propose a novel data augmentation technique for machine translation using dependency parsing. As data augmentation is particularly useful when training data is insufficient, we perform these experiments in a simulated low-resource setting, using a subset of the Hunglish2 corpus.

3.1 Formulation

We formulate machine translation on the sentence level. Given a dataset $\mathcal{D}$ that contains parallel sentences from the source and target language $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{D}$, we

define the goal of neural machine translation as estimating the unknown conditional probability $P(\boldsymbol{y}|\boldsymbol{x})$. This is a classical sequence-to-sequence problem: an encoder can be used to create a representation of the source sentence, which is fed into a decoder with the purpose of extracting relevant information from this representation. The decoder then generates the output symbols from left to right. This way the decoder can be thought of as a language model conditioned on the output of the encoder and the already generated symbols.

3.2 Dataset

We use the Hunglish2 corpus for our experiments, which is a sentence-aligned corpus consisting of 4.1M Hungarian-English bisentences (Varga et al., 2007). The dataset was constructed by scraping and aligning bilingual data in several domains from the internet. The distribution of each domain that make up Hunglish2 is shown in Table 1.

Subcorpus	**Tokens**	**Bisentences**
Modern literature	37.1M	1.67M
Classical literature	17.2M	652k
Movie subtitles	3.2M	343k
Software docs	1.2M	135k
Legal text	56.6M	1.351M
Total	115.3M	4.151M

Table 1: Statistics of the Hunglish2 corpus.

We applied thorough preprocessing before training the models. First, we removed seemingly incorrect data points: sentence pairs, which contained HTML code or were outliers with respect to sentence length. We also remove sentence-pairs, where either side is an empty string. A large number of sentences were wrapped in quotation marks, so in order to avoid overfitting on this behaviour, we also remove leading and trailing quotation marks. Next, we filter the dataset with length-based heuristics. We drop sentences if either the source or target sentence contains more than 32 words. Furthermore, we filter with the relative word counts of the source and target sentences using the below rule:

$$(|\mathrm{WC(x)} - \mathrm{WC(y)}| < 7) \vee (WC(x)/WC(y) < 1.6)$$

where $WC(x)$ and $WC(y)$ are the word counts of the source and target sentences respectively. The threshold parameters were determined by exploratory data analysis and a series of experiments. See Figure 1 for the distribution of length difference and length ratio of bisentences in the raw Hunglish2 corpus. The post-processed dataset contains 3.4M bisentences. Finally, we split the data to train,

development and test sets, with a 99-0.5-0.5 ratio. We do this with stratified sampling with respect to the the subcorpora in Hunglish2, to ensure that all splits have a similar distribution.

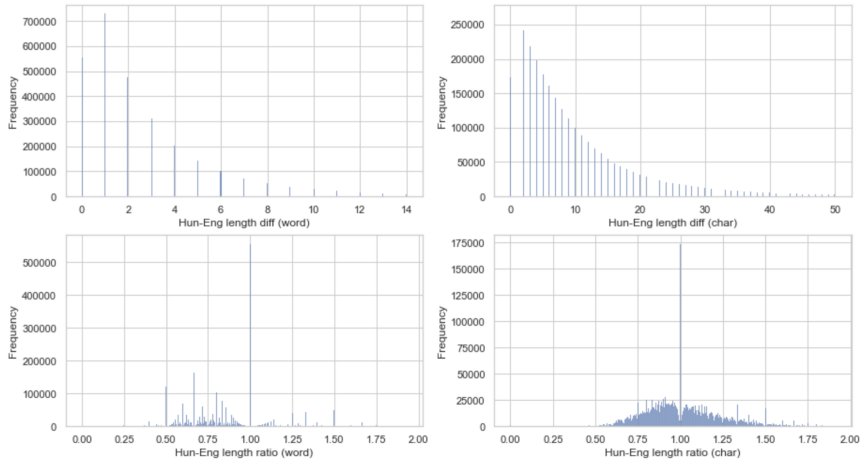

Fig. 1: Word- and character-level distributions of the length difference and ratio between source and target sentences in the Hunglish2 corpus.

3.3 Training

All of our models use identical architecture and hyperparameters: a vanilla Transformer-based encoder-decoder model. We tokenize the input with the unigram sentencepiece subword tokenizer (Kudo and Richardson, 2018), which works particularly well with morphologically rich languages, such as Hungarian. We found that using a shared vocabulary of size 32000 yields the best results. We perform gradient descent using the Adam optimizer (Kingma and Ba, 2014) with an initial learning rate of 2, 8000 warmup steps and Noam learning rate decay. To avoid overfitting, we perform early stopping based on the perplexity computed on the validation set. Due to computational constraints, we only perform a manual search in the hyperparameter space. The complete set of hyperparameters of our best model can be found in Table 2. All models were trained using the OpenNMT framework (Klein et al., 2017) with standalone Nvidia Tesla V100 GPUs.

3.4 Syntax-aware data augmentation

One of the greatest challenges in data augmentation for machine translation is preserving parallelism. Taking the example from Duan et al. (2020) (see Table

3.) it is visible, that replacing words only on the source side of the bisentence can easily lead to noisy translation pairs. The success of backtranslation can be partially contributed to the fact that it is good at generating parallel bisentences. Backtranslation, however, might not always be an option, especially in low-resource scenarios, where there is not enough parallel data to train a model that can be used for backtranslation in the first place.

Parameter	**Value**	**Parameter**	**Value**
batch type	tokens	batch size	4096
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 our best model.

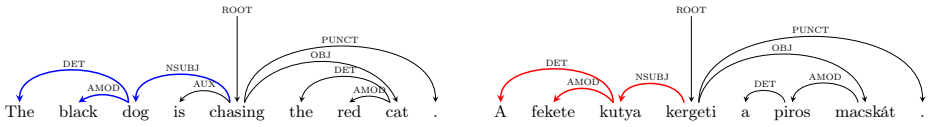

Fig. 2: An example of finding the same substructure in the dependency trees of a bisentence.

We propose a novel syntax-aware data augmentation technique, which is based on a hypothesis, that dependency parse trees of the source and target sentences contain subtrees, which carry the same meaning. If we can identify these subtrees simultaneously in the source and target language (see Figure 2 for an example), we have the possibility to generate new bisentences by swapping subtrees.

Original-EN	We shall fight on the beaches.
Original-HU	A tengerparton kellene küzdenünk.
Replacement-EN	We shall fight with the sandy.
Replacement-HU (Google Translate)	Harcolni fogunk a homokkal.

Table 3: Error analysis of a data augmentation method for NMT (Duan et al., 2020).

Finding any subtree-pair, which correspond to the same part of the source and target sentence is a very hard task, so we limit our work to finding two substructures that are common across languages: subjects and objects (see Figure 3 and Table 4). As the problem space of new sentence-pairs that we can generate explodes quickly with respect to the size of the dataset, we filter sentences based on their dependency parse trees with two conditions. First, the dependency trees must contain only one subject and object for both the source and target sentences. Second, the subtrees corresponding to objects and subjects must be a consecutive sequence of words with respect to the original word order. In our experiments, we found that about 5% of the sentence pairs satisfy the above two conditions and therefore are eligible for augmentation. It is convenient to have the same dependency relations for both English and Hungarian, so we use the Universal Dependencies[2] (Nivre et al., 2016) tag set. Implementation-wise, we chose the Stanza[3] dependency parser (Qi et al., 2020) for English and the parser in the Hungarian Spacy[4] model for Hungarian.

We perform a series of simulated low-resource experiments on a subcorpus of Hunglish2. We constrain our experiment to only one subcorpus, because this better simulates the lack of diversity in parallel corpora of low-resource languages. We subsample from the Modern Literature subcorpus of Hunglish2 from 5k to 500k. For each dataset, we perform three experiments: train a baseline Transformer model and train the same model with additional data from the two proposed augmentation methods. In every iteration, we extend the datasets with 50% augmented data. Every model is evaluated on the same held-out test set, that is a fraction of our original test set with samples only from the Modern Literature subcorpus.

We acknowledge that sampling from a medium- or high-resource corpus is not the same as working on a truly low-resource language pair. The latter is likely a much worse representation of the overall population, but for the course of this work we limit our experiments to Hungarian-English. Extending these experiments to actual low-resource language pairs is a promising direction for future work.

[2] https://universaldependencies.org/
[3] https://stanfordnlp.github.io/stanza/
[4] https://github.com/huspacy/huspacy

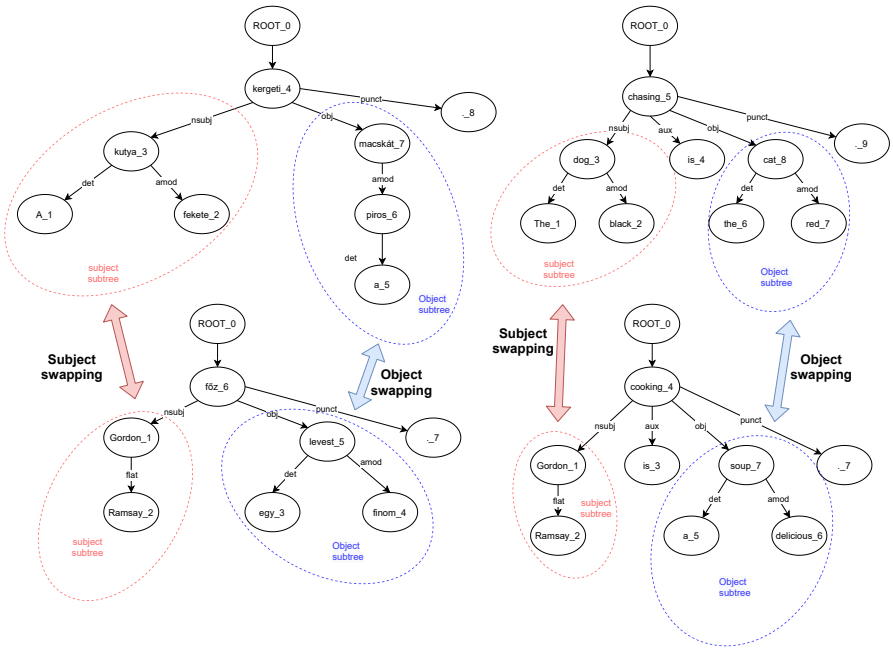

Fig. 3: The process of swapping subject and object dependency subtrees between bisentences. The augmented sentences can be seen in Table 4.

Sentence1-EN	The black dog is chasing the red cat.
Sentence1-HU	A fekete kutya kergeti a piros macskát.
Sentence2-EN	Gordon Ramsay is cooking a delicious soup.
Sentence2-HU	Gordon Ramsay egy finom levest főz.
EN-OBJ-AUG-1	The black dog is chasing a delicious soup.
HU-OBJ-AUG-1	A fekete kutya kergeti egy finom levest.
EN-OBJ-AUG-2	Gordon Ramsay is cooking the red cat.
HU-OBJ-AUG-2	Gordon Ramsay a piros macskát főz.

Table 4: Augmentation via subtree swapping of objects.

4 Results

In this section, we provide a detailed evaluation of both the Transformer-based machine translation models and the proposed augmentation method. For quantitative evaluation, we use the BLEU score.

NMT models for EN-HU and HU-EN Our baseline model without augmentation achieved a BLEU score of **33.4** for EN-HU and **40.0** for HU-EN on the held-out test set. In Table 5, we collected a few example translations from the test set provided by our best models.

Syntax-aware data augmentation The scores of our simulated low-resource experiments can be found in Table 7 for both HU-EN and EN-HU. The absolute performance gain due to the augmentation is visualized in Figures 4 and 5 for each sample size. With a smaller sample size (5k, 10k, 25k), the models had a close to 0 BLEU score. This can likely be contributed to the fact that the model is fixed for all experiments and it is probably too complex for datasets of this size regardless of augmentation. In the 50k-100k range, we observe visible improvements in the BLEU score of models trained with augmentation compared to the baseline models. With a 75k sample size, the baseline BLEU scores of 0.9 and 1.4 are significantly outperformed by the model trained with object swapping augmentation with BLEU scores of 6.1 and 8.8 for EN-HU and HU-EN respectively. In the 100k-500k range, we do not see improvement in BLEU score with the augmentation methods. With two exceptions, the models using augmented data perform slightly worse than the baseline model. As the augmentation ratio is also fixed to 0.5 during all experiments, it is possible that at this scale we inject too many and too noisy new data points into the training set. We also examined the reason behind the noisiness of the augmentation by manual analysis. We collected a few common error types, which are listed in Table 6. Apart from the ones listed below, we found that most of the errors propagate from an incorrect dependency parsing, especially for Hungarian. We observed many falsely identified subjects, especially in cases, where the subject of the sentence was dropped (pronoun-dropping).

5 Conclusion

We presented Transformer-based NMT models for Hungarian-English and English-Hungarian. Our best models achieve a BLEU score of **40.0** and **33.4** for HU-EN and EN-HU respectively. We also shared results of an ongoing work on a potential data augmentation method alternative to back-translation in lower-resource scenarios. We briefly discussed this syntax-aware method, which creates new data points by swapping specific subtrees of dependency parse trees in parallel for both the source and target sentences. Regarding our future work, we plan to fix some of the common errors listed in Table 6 and therefore enhance the

Example #		Sentence
1	**Source**	Villefort ezeket az utolsó szavakat olyan lázas dühvel ejtette ki, ami egészen vadul ékesen szólóvá tette.
	Reference	Villefort pronounced these last words with a feverish rage, which gave a ferocious eloquence to his words.
	Predicted	Villefort pronounced these last words with a feverish rage which rendered him passionately eloquent.
2	**Source**	Amióta Merytonba szállásolták az ezredet, csak a szerelem, az udvarlás, a tisztek jártak az eszében.
	Reference	Since the shire were first quartered in Meryton, nothing but love, flirtation, and officers have been in her head.
	Predicted	He had been thinking of love, of courting, of officers, ever since the regiment came to Meryton.
3	**Source**	Malfoy, Crak és Monstro Csikócsőrrel próbálkoztak.
	Reference	Malfoy, Crabbe, and Goyle had taken over Buckbeak.
	Predicted	Malfoy, Crabbe, and Goyle had tried Buckbeak.
4	**Source**	His remembrance shall be sweet as honey in every mouth, and as music at a banquet of wine.
	Reference	Mint ínyünknek a méz, édes az emléke, vagy mint a nótaszó borozgatás közben.
	Predicted	Emlékezete édes lesz, mint a méz minden szájban, és mint a zene a bor lakomáján.
5	**Source**	His eyes moved toward the hunting knife that had been slung over the mosquito-net bar by the dead man the day he arrived.
	Reference	Szeme a vadászkésre siklott, amelyet a halott ember a moszkitóháló keretére dobott azon a napon, amikor megérkezett.
	Predicted	Szeme a vadászkés felé fordult, melyet a halott férfi a moszkitóháló rácsára dobott az érkezése napján.
6	**Source**	He was frozen stiff in the weeds beside the track.
	Reference	Csonttá fagyva feküdt a vágány mellett a gazos földön.
	Predicted	Merevre fagyott a vágány melletti gazban.

Table 5: Example translations produced by the best EN-HU and HU-EN translator models.

Error type	Sentences
Article definiteness	**Source (aug)** Outside the apothecary, Hagrid checked the weapons again. **Target (aug)** Mikor végeztek a patikában, Hagrid még egyszer ellenőrizte a fegyvert.
Coreference	**Source (aug)** Two other companies claimed only it. **Target (aug)** Két másik vállalat csak egyéni elbánást kérelmezett.
Conjugation	**Source (aug)** Member states had also played an ironic role here. **Target (aug)** A tagállamok ismét ironikus játékot űzött vele.
Different subject	**Source (aug)** Captain, how many men did the wind leave on mars? **Target (aug)** A szél, hány embert hagytál a Marson?
Pronoun dropping	**Source (orig)** It shall submit the reports to the european parliament and to the council. **Target (orig)** Ő A jelentéseket az Európai Parlamenthez és a Tanácshoz nyújtja be.

Table 6: Common error types of the syntax-aware data augmentation.

Sample size	Method	EN-HU BLEU	HU-EN BLEU
5k	base	0.1	0.0
	object swapping	0.0	0.0
	subject swapping	0.0	0.0
10k	base	0.2	0.1
	object swapping	**0.3**	0.1
	subject swapping	**0.3**	0.0
25k	base	0.4	0.1
	object swapping	0.3	0.1
	subject swapping	0.3	**0.3**
50k	base	0.4	0.8
	object swapping	**1.6**	**2.4**
	subject swapping	**1.7**	**2.6**
75k	base	0.9	1.4
	object swapping	**6.1**	**8.8**
	subject swapping	**5.9**	**8.2**
100k	base	3.0	5.4
	object swapping	**7.4**	**9.7**
	subject swapping	**7.8**	**10.0**
200k	base	12.7	14.5
	object swapping	12.1	14.5
	subject swapping	12.2	**14.9**
300k	base	14.4	16.5
	object swapping	14.1	16.4
	subject swapping	14.1	16.1
400k	base	15.5	17.0
	object swapping	15.3	16.7
	subject swapping	15.3	16.9
500k	base	15.5	17.7
	object swapping	15.5	17.6
	subject swapping	**15.6**	17.2

Table 7: Results of our low-resource experiments. The BLEU scores in bold indicate an experiment, where the model with augmentation outperformed the baseline for that particular sample.

augmentation technique by making the generated samples less noisy. We also plan to extend our experiments to other languages.

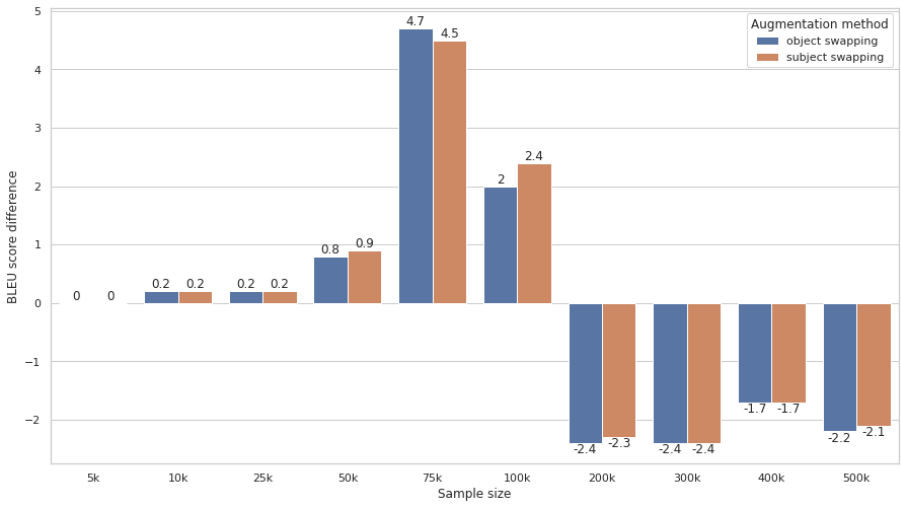

Fig. 4: Absolute BLEU score differences compared to the baseline model for each sample size. (EN-HU)

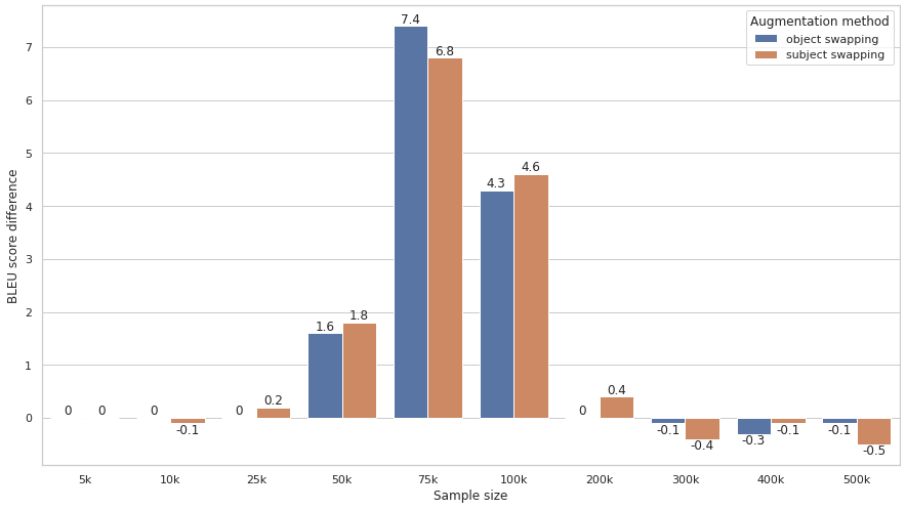

Fig. 5: Absolute BLEU score differences compared to the baseline model for each sample size. (HU-EN)

Acknowledgements

The authors would like to thank András Kornai for discussions on the syntax-aware data augmentation.

Bibliography

Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473 (2014)

Brown, P.F., Cocke, J., Della Pietra, S.A., Della Pietra, V.J., Jelinek, F., Lafferty, J., Mercer, R.L., Roossin, P.S.: A statistical approach to machine translation. Computational linguistics 16(2), 79–85 (1990)

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078 (2014)

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 (2020)

Duan, S., Zhao, H., Zhang, D., Wang, R.: Syntax-aware data augmentation for neural machine translation. arXiv preprint arXiv:2004.14200 (2020)

Germann, P.C.J.H.U., Bogoychev, L.B.N., Waldendorf, A.V.M.B.J., Heafield, A.B.K.: The University of Edinburgh's English-German and English-Hausa submissions to the WMT21 news translation task (2021)

Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014), https://arxiv.org/abs/1412.6980

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

Kobayashi, S.: Contextual augmentation: Data augmentation by words with paradigmatic relations. arXiv preprint arXiv:1805.06201 (2018)

Koehn, P., Och, F.J., Marcu, D.: Statistical phrase-based translation. Tech. rep., UNIVERSITY OF SOUTHERN CALIFORNIA MARINA DEL REY INFORMATION SCIENCES INST (2003)

Kudo, T., Richardson, J.: Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. arXiv preprint arXiv:1808.06226 (2018)

Laki, L., Novák, A., Siklósi, B.: English to Hungarian morpheme-based statistical machine translation system with reordering rules. In: Proceedings of the Second Workshop on Hybrid Approaches to Translation. pp. 42–50 (2013)

Nagy, A.: Developing neural machine translation models for Hungarian-English. arXiv preprint arXiv:2111.04099 (2021)

Nivre, J., De Marneffe, M.C., Ginter, F., Goldberg, Y., Hajic, J., Manning, C.D., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., et al.: Universal dependencies v1: A multilingual treebank collection. In: Proceedings of the Tenth

International Conference on Language Resources and Evaluation (LREC'16). pp. 1659–1666 (2016)
Oravecz, C., Bontcheva, K., Tihanyi, L., Kolovratnik, D., Bhaskar, B., Lardilleux, A., Klocek, S., Eisele, A.: etranslation's submissions to the WMT 2020 news translation task. In: Proceedings of the Fifth Conference on Machine Translation. pp. 254–261 (2020)
Prószéky, G., Tihanyi, L.: Metamorpho: A pattern-based machine translation system. In: Proceedings of the 24th'Translating and the Computer'Conference. pp. 19–24. Citeseer (2002)
Qi, P., Zhang, Y., Zhang, Y., Bolton, J., Manning, C.D.: Stanza: A Python natural language processing toolkit for many human languages. arXiv preprint arXiv:2003.07082 (2020)
Şahin, G.G., Steedman, M.: Data augmentation via dependency tree morphing for low-resource languages. arXiv preprint arXiv:1903.09460 (2019)
Sennrich, R., Haddow, B., Birch, A.: Improving neural machine translation models with monolingual data. arXiv preprint arXiv:1511.06709 (2015)
Tihanyi, L., Oravecz, C.: First experiments and results in English-Hungarian neural machine translation. XIII. Magyar Számítógépes Nyelvészeti Konferencia. Szeged, Hungary (2017)
Tran, C., Bhosale, S., Cross, J., Koehn, P., Edunov, S., Fan, A.: Facebook AI WMT21 news translation task submission. arXiv preprint arXiv:2108.03265 (2021)
Vania, C., Kementchedjhieva, Y., Søgaard, A., Lopez, A.: A systematic comparison of methods for low-resource dependency parsing on genuinely low-resource languages. arXiv preprint arXiv:1909.02857 (2019)
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, W.Y., Yang, D.: That's so annoying!!!: A lexical and frame-semantic embedding based data augmentation approach to automatic categorization of annoying behaviors using# petpeeve tweets. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. pp. 2557–2563 (2015)
Xie, Z., Wang, S.I., Li, J., Lévy, D., Nie, A., Jurafsky, D., Ng, A.Y.: Data noising as smoothing in neural network language models. arXiv preprint arXiv:1703.02573 (2017)
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. arXiv preprint arXiv:1601.03651 (2016)