Shortening of the Results of Machine Translation using Paraphrasing Dataset

Andrej Perković[1], Jernej Vičič[1], Dávid Javorský[2] and Ondřej Bojar[2]

[1]University of Primorska, Faculty of Mathematics, Natural Science and Information Technologies, Glagoljaska 8, 6000 Koper, Slovenia

[2]Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics, Malostranské nám. 25, Prague, 118 00, Czech Republic

Abstract

As machine translation applications continue to expand into the realm of real-time events, the need for faster and more concise translation becomes increasingly important. One such application is simultaneous speech translation, an emission of subtitles in the target language given speech in the source language. In this work, we focus on easing reader's comprehension of subtitles by making the translation shorter while preserving its informativeness. For this, we use the S, M and L version of the Paraphrase Database (PPDB), and exploit their property that some of the paraphrasing rules differ in length of the left and right side. Selecting rules that make the output shorter, we fine-tune an MT model to naturally generate shorter translations. The results show that the model's conciseness improves by up to 0.61%, which leaves the space for improvements using bigger versions of PPDB in future work.

Keywords

language shortening, constrained machine translation, NMT, Serbian language

1. Introduction

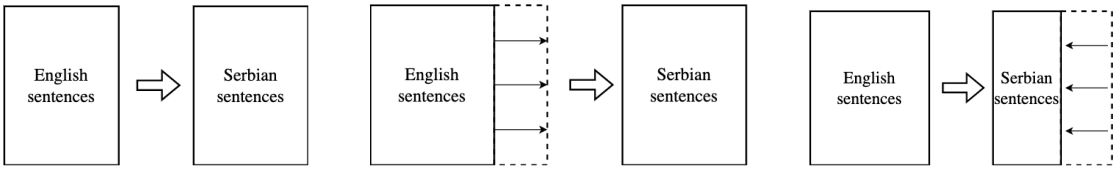
Machine translation (MT) has recently shown great improvements in both the translation quality and speed, allowing us to tackle more challenging tasks, e.g. simultaneous speech translation (SST). A typical approach to automatically deliver textual translation (i.e. subtitles) of input speech is a pipeline of several components: speech recognition, segmentation and translation. Preliminary experiments suggest that some users prefer low latency [1]. Furthermore, it is sometimes impossible to fit all the translated text in subtitle space given the high pace of the input speech. A possible way to make the user experience more pleasant is to reduce the amount of displayed text during subtitling whereas conveying the same amount of information. This work aims at addressing this challenge through the use of paraphrasing techniques to shorten translations.

To achieve this goal, we utilized the Paraphrasing Database (PPDB) [2, 3], which was developed by researchers at the University of Pennsylvania. Rules were created based on

ITAT'23: Information technologies – Applications and Theory, September 22–26, 2023, Tatranské Matliare, Slovakia

89201045@student.upr.si (A. Perković); jernej.vicic@upr.si (J. Vičič); javorsky@ufal.mff.cuni.cz (D. Javorský); ondrej.bojar@mff.cuni.cz (O. Bojar)

0000-0002-7876-5009 (J. Vičič)



(a) The standard form of corpora used in the training of the models.

(b) Intended modification of the corpora. Extension of English sentences using the paraphrasing rules.

(c) Expected results after fine-tuning.

Figure 1: Diagram representation of the research question

this database to make the English source sentences within the corpus longer. The rationale behind this approach is that training the model on "lengthened" source sentences paired to the "standard" target sentences (Figure 1b), would yield shorter target sentences when translating real-world "standard" source sentences (Figure 1c).

2. State of the Art

Text shortening is by itself a fairly explored phenomenon overlapping with text simplification, both are further explained in the next sections. The combination with machine translation adds another dimension.

2.1. Text simplification

Text simplification focuses on making the text more comprehensible for the target public (such as non-native speakers, children, and reading-impaired people). It disregards the length in the process of transformation, but text simplification usually produces shorter texts.

WordNet [4], the database of synonyms, was often used as a basis of the research for text simplifications. Shortcoming is that users report degraded experience. In the study done by Walker et al [5], the test groups report preference for less ambiguous words. Other research used Simplified English Wikipedia [6]. Championed by Coster and Kauchak [6], not only did it facilitate text simplification, but it also improved the BLEU [7] score quantifying the performance of a machine translation model. It keeps producing better results since it is richer with context, unlike WordNet [8]. Unfortunately, most other languages lack such a comprehensive simplified corpora. Wang, et al. [9] introduces a new approach to this task using NMT by applying the principles of MT which essentially maps source sentences to target sentences in a one-to-one relationship in most cases, while the simplification carries much more nuance in mapping source sentence to target sentences.

Rule-based automatic approach to text simplification consists of applying predefined alphabetical, lexical, syntactical and phrasal rules in an algorithmic way to achieve simpler target sentences. In addition to English leading the way in the scope of the development of

these methodologies, other languages also enjoy benefits of robust automatic simplification systems. Since much of the progress is achieved for the English language, we are interested in trickling those discoveries on other ones as well, especially the South Slavic. One example is German. Suter et al. [10] have pioneered rule-based automatic text simplification for German. Their system is able to reduce the complexity by a level on the LIX scale [11], while the human simplification was able to reduce it by two levels. Besides the still limited advancements in this field, they are all constrained in the regard that for each languages, researches have to develop a new set of rules and interconnection of steps. Our method would use the existing advantage English has over other languages, namely being the most researched one, in achieving the desired results. The same method applicable to all translation model regardless of the target languages.

2.2. Controlled length and text shortening

There are several approaches to controlling the length of outputs of natural language processing tasks. One such approach is that of text summarizing. Research was made where the models are trained to create outputs of fixed length through rule-based approach [12] or by using statistical methods [13]. Text shortening concerns only with reducing the length of the input, with varying degree of worry about grammaticality and meaning preservation. Research into this topic has been accomplished in many European languages due to the many applications of shortening. For English and French, Yousfi-Monod and Prince [14] were able to achieve substantial 40% reduction in length on average with a slight decrease in readers' satisfaction. The basis of their work consists in representing a sentence as a tree of constituents and then pruning the tree accordingly.

All of these approaches are focusing on the only task of shortening the input text. The approach presented in this paper focuses on translating from source language to target language in a shortened way with a single model.

2.2.1. Machine translation with text shortening

Machine translation with the focus on length constraints as a single job presents some obvious advantages over a split job of first translating and then compressing (in some cases expanding) the final text. Such operation is essential if the translation should be displayed in a given format. Jan Niehues [15] reports a significant improvement of the translation quality under constraints using coder-decoder architecture. Nguyen et. al [16] present a rule-based for text shortening in Vietnamese sign language translation. A large-scale MT project for TV titles is presented in [17].

3. Methodology and work

The reported research project includes modifying the English sentences in the training dataset `OPUS-100` [18, 19, 20] by the means of the selected paraphrasing pairs, referenced as "rules" hereinafter, from `PPDB`[1], fine-tuning the translation model from `Helsinki-NLP` [21, 22]

[1] http://paraphrase.org/#/download

Version	Total	Filtered	Applied
S	231k	10331	4268
M	463k	12 836	5105
L	926k	15 838	6956

Table 1
Total number of rules, number of rules after filtering and, finally, the number of applied rules from each packages. All filtered rules were searched for in the English part of the test set, but in the end only certain number of them was actually applied

```
[RB] ||| consequently ||| hence ||| PPDB2.0Score=3.73421 PPDB1.0Score=7.837540 …WordLogCR=0
||| 0-0 ||| Equivalence
```

(a) Typical example

```
[CD] ||| 79 ||| seventy-nine ||| PPDB2.0Score=4.07594 PPDB1.0Score=8.260030 …WordLogCR=0 |||
0-0 ||| Equivalence
```

(b) Numerical rule

```
[VBN] ||| fluctuated ||| oscillated ||| PPDB2.0Score=3.80730 PPDB1.0Score=10.341820
…WordLogCR=0 ||| 0-0 ||| Equivalence
```

(c) Equal length rule

Figure 2: Examples of rules in the PPDB

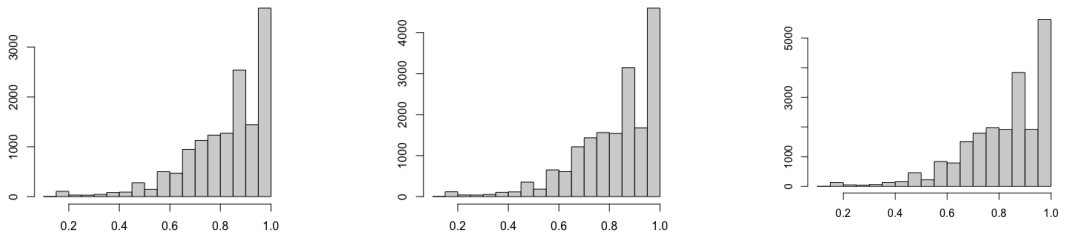
and finally evaluating and comparing the performance.

PPDB is divided into six increasingly large sizes – S, M, L, XL, XXL, and XXXL based on how closely related the rules are. Larger sizes contain increasingly less related paraphrasing rules [3]. The number of paraphrases doubles with each increase in size, and larger sizes subsume smaller sizes. Additionally, there are three types of paraphrases – lexical, phrasal and syntactic. The researches focuses on lexical ones. Table 1 shows the exact number of rules for each package used in the research project. We have taken three different versions. Each was taken through the same steps. First thing accomplished was filtering the given version package. It is worth noting that some rules are duplicates, i.e. the same pair of words can appear in two entries with reversed positions. Entries in the database are illustrated in Figure 2.

All the paraphrasing pairs were rearranged so that the shorter phrase was on the left and the longer on the right. When filtering, all the feature-value pairs (4th column in the examples in Figure 2) were removed. The length ratio between the phrases was incorporated. It was calculated as the quotient between the longer and the shorter phrase. Additionally, there are different types of entailment of pairs in PPDB - Equivalent (e.g. look at/watch), Entailment (e.g. tower/building), Exclusion (e.g. close/open), Other relation (e.g. swim/water) and Unrelated (e.g. girl/play). Only those pairs labeled as Equivalence are retained, as they have the potential to shorten a word without distorting the meaning. The following pairs were also removed:

- those where one entry is just a number,
- pairs that are of equal length.

You can see examples in Figures 2b and 2c. After this, rules were applied to the English



(a) S version (b) M version (c) L version

Figure 3: Frequency distribution of the ratios of the rule pairs extracted from PPDB. Width of the interval is 0.05

	Baseline	S	M	L
Characters	26'789'153	28'147'979	28'572'397	29'171'997
Increase	/	5.07	6.66	8.89

Table 2
Relative differences in the number of alphanumeric characters for different versions of the training set of OPUS-100

source sentences of OPUS-100. This modified training set was then used to fine-tune the models.

In Figure 3, the distribution of the ratios in the three packages is shown. A notable percentage of pairs with equal-length phrases was observed. Other than that, we can see that most of ratios have the value around 0.9.

Additionally, the part of the OPUS-100 dataset with the Serbian translation was modified as well. There were issues with certain letters and scripts, the former having a qualitative and the latter quantitative influence on the translation. The first issue relates to errors in encoding of the letters in the Latin script specific to the Serbian language. Namely, letters č, ć and đ were encoded as è, æ and ð, respectively. The concern with the scripts is that the dataset contained sentences in both interchangeable writing systems of the Serbian language - Latin and Cyrillic. To avoid training the MT model to relate a word or a context with one of the scripts, all the sentences in the Cyrillic script were changed to their Latin equivalents. Additionally, sentences in Cyrillic are shorter on average, making length comparisons between sentences in different scripts nonsensical. This part of the detest with the encoding of Serbian translation corrected in the manner described above is referred to as "corrected Serbian" hereinafter. Despite this, OPUS-100 was a great dataset, since it had one million sentences in the training set and a separate test set.

In Table 2, you can see the effects on the increase in alphanumeric characters in the training corpora the application of PPDB had. This corresponds to the diagram in Figure 1b.

	Default	S	M	L
Number of characters	59869	59722	59505	59372
Decrease	/	0.25	0.61	0.33
Number of words	13637	13632	13601	13617

Table 3
Absolute translation lengths in characters and words and decreas in translation length in percentage terms

Default	21.45
S	23.54
M	23.20
L	23.42

Table 4
BLEU scores of the different versions of the model

4. Results

To measure possible effects fine-tuning might have had on the model, the model was deployed for the translations of the test set sentences before and after fine-tuning it with the datasets modified using the PPDB. We also measured output length in characters for the original and fine-tuned models. The degree of shortening was calculated as the ratio between these two numbers, presented in Table 3. What is surprising is that the number of words has decreased, even though we did not change the word count when modifying the training set. The majority of word-compressed sentences are the result of shortened forms of verb tenses in Serbian language.

We also calculated the BLEU scores, using the OPUS-100 test set with the corrected Serbian translation as the reference. The results are presented in Table 4, with a score of 21.45 points for the original model, 23.54 points for the model fine-tuned with S version of PPDB, 23.20 for the M and 23.42 for the L version one. Based on this automatic metric, translation quality was not harmed by our method.

Lastly, the quality of the translation was manually evaluated for the S version. Fifty-five sentences were selected at random and labeled according to how well they represent the given source English sentence. This was performed in a way that anonymized the systems producing the translation in order to remove human bias as much as possible when evaluating. We used the quickjudge[2] program, which allowed us to see a block of four lines for each sentence. It included:

1. the source English sentence in the first line (labeled with "in.txt" in Figure 4)
2. suggested translation into Serbian extracted from the OPUS-100 dataset as a reference translation in the next line (labeled with "ref.txt" in the aforementioned Figure)
3. default output, i.e. translation by the default Helsinki-NLP system trained on the English-to-Serbian OPUS-100 data, and

[2]https://github.com/ufal/quickjudge

```
in.txt I know my dad loves socket wrenches.
ref.txt On obožava ključeve.
- Znam da moj tata voli utičnice.
** Znam da moj tata voli ključeve.
```

(a) An example of a good and bad translation, where the last word is wrongly translated in the first translation

```
in.txt Of dying?
ref.txt Od smrti?
* missT Umiranje?
** Umiranja?
```

(b) An example of a good and better translation, where the good translation is missing the more appropriate grammatical case of the noun "dying" for the given context. It uses nominative, while genitive would be more reasonable, which is also the case utilized in the reference translation.

Figure 4: Examples of sentences in the annotation file

4. output of the model re-trained using data modified by the paraphrases in the S version of `PPDB` data

The last two lines were not distinguished. You can see two examples in the Figure 4.

We then labeled both translations. Exactly one of the following labels were given to each sentence:

- `**` - better of two good translations
- `*` - good translation
- `-` - wrong translation

Furthermore, translated sentences that were good, but not perfect were marked with additional labels:

- `missT` - mistranslated a word or used a wrong case for a noun
- `missW` - lacking translation of a word or a part of source English sentence, i.e. having correct but partial translation that does not drastically affect the meaning

Counting the number of repetition of labels gives us the rough idea of the performance of the two models. We conclude that the model without shortening has a higher percentage of better translations, as was expected. What is interesting is that the number of bad translations is approximately the same, which goes in line with what the BLEU score is suggesting. The shortened translation has a higher rate of missing words or translations without significantly deteriorating the quality of translation, consistent with the expectations. Exact results are visible in Table 5.

		**	*	-	Total	missT	missW	Total
Normal	Absolute	32	15	8	55	1	2	3
	Relative (%)	58.2	27.3	14.6	100	1.8	3.6	5.5
Short	Absolute	19	27	9	55	3	4	7
	Relative (%)	34.6	49.1	16.4	100	5.5	7.3	12.7

Table 5
Results of the annotation of the translations. All relative values are calculated as the percentage of the 55 sentences

5. Conclusion

For this research project, we modified the OPUS-100 dataset in two ways. One modification included the correction of the Serbian sentences while the other had expanded English sentences using paraphrasing rules from S, M and L versions of PPDB. These tweaked datasets were then used to fine-tune the Helsinki-NLP MT model. Lastly, we compare their performance qualitatively and quantitatively.

Using the three versions of PPDB to lengthen the English source sentences in training sets to shorten the translation from English to Serbian has minimal results at this scale. It is noteworthy that the BLEU score has not degraded after this modification. On the contrary, it slightly increased. Manual inspection of the quality of translation confirmed the BLEU results.

For the next step, we could experiment by expand the entailment types of the paraphrases used beyond just Equivalence. Such enlargement would require careful consideration of the trade-offs involved, namely the relationship between greater conciseness and translation variety. For instance, some form of generalization (via the Entailment relation) may be desirable.

Acknowledgments

Computational resources were provided by the e-INFRA CZ project (ID:90254), supported by the Ministry of Education, Youth and Sports of the Czech Republic. This work was partially supported by the sabbatical grant of the University of Primorska. This work was partially supported by the grant 19-26934X (NEUREM3) of the Czech Science Foundation.

References

[1] D. Javorskỳ, D. Macháček, O. Bojar, Continuous rating as reliable human evaluation of simultaneous speech translation, in: Proceedings of the Seventh Conference on Machine Translation (WMT), 2022, pp. 154–164.

[2] J. Ganitkevitch, C. Callison-Burch, The multilingual paraphrase database, in: The 9th edition of the Language Resources and Evaluation Conference, European

Language Resources Association, Reykjavik, Iceland, 2014, pp. 1–8. URL: http://cis.upenn.edu/~ccb/publications/ppdb-multilingual.pdf.

[3] E. Pavlick, P. Rastogi, J. Ganitkevitch, B. Van Durme, C. Callison-Burch, Ppdb 2.0: Better paraphrase ranking, fine-grained entailment relations, word embeddings, and style classification, in: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), 2015, pp. 425–430.

[4] G. A. Miller, Wordnet: a lexical database for english, Communications of the ACM 38 (1995) 39–41.

[5] A. Walker, A. Siddharthan, A. Starkey, Investigation into human preference between common and unambiguous lexical substitutions, in: Proceedings of the 13th European Workshop on Natural Language Generation, 2011, pp. 176–180.

[6] W. Coster, D. Kauchak, Simple english wikipedia: a new text simplification task, in: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, 2011, pp. 665–669.

[7] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, Bleu: a method for automatic evaluation of machine translation, in: Proceedings of the 40th annual meeting of the Association for Computational Linguistics, 2002, pp. 311–318.

[8] O. Biran, S. Brody, N. Elhadad, Putting it simply: a context-aware approach to lexical simplification, in: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, 2011, pp. 496–501.

[9] T. Wang, P. Chen, J. Rochford, J. Qiang, Text simplification using neural machine translation, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 30, 2016, pp. 4270–4271.

[10] J. Suter, S. Ebling, M. Volk, Rule-based automatic text simplification for german, Zurich Open Repository and Archive (2016).

[11] J. Anderson, Lix and rix: Variations on a little-known readability index, Journal of Reading 26 (1983) 490–496. URL: http://www.jstor.org/stable/40031755.

[12] B. Dorr, D. Zajic, R. Schwartz, Hedge trimmer: A parse-and-trim approach to headline generation, Technical Report, Maryland university college park inst for advanced computer studies, 2003.

[13] D. Galanis, I. Androutsopoulos, An extractive supervised two-stage method for sentence compression, in: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, 2010, pp. 885–893.

[14] M. Yousfi-Monod, V. Prince, Sentence compression as a step in summarization or an alternative path in text shortening, in: Coling 2008: Companion volume: Posters, 2008, pp. 139–142.

[15] J. Niehues, Machine translation with unsupervised length-constraints, 2020. arXiv:2004.03176.

[16] T. B. D. Nguyen, P. Trung-Nghia, V. Vu Tat-Thang", editor="Bhateja, N. B. Le, N. N. Gia, S. S. Chandra, L. Dac-Nhuong, A rule-based method for text shortening in vietnamese sign language translation, in: Information Systems Design and Intelligent Applications, Springer Singapore, Singapore, 2018, pp. 655–662.

[17] M. Volk, R. Sennrich, C. Hardmeier, F. Tidström, Machine translation of tv subtitles for large scale production, in: JEC 2010; November 4th, 2010; Denver, CO, USA, Association for Machine Translation in the Americas, 2010, pp. 53–62.

[18] J. Tiedemann, Parallel data, tools and interfaces in OPUS., in: Lrec, volume 2012, Citeseer, 2012, pp. 2214–2218.

[19] B. Zhang, P. Williams, I. Titov, R. Sennrich, Improving massively multilingual neural machine translation and zero-shot translation, 2020. arXiv:2004.11867.

[20] R. Aharoni, M. Johnson, O. Firat, Massively multilingual neural machine translation, arXiv preprint arXiv:1903.00089 (2019).

[21] J. Tiedemann, The tatoeba translation challenge – realistic data sets for low resource and multilingual MT, in: Proceedings of the Fifth Conference on Machine Translation, Association for Computational Linguistics, Online, 2020, pp. 1174–1182. URL: https://aclanthology.org/2020.wmt-1.139.

[22] J. Tiedemann, S. Thottingal, OPUS-MT – building open translation services for the world, in: Proceedings of the 22nd Annual Conference of the European Association for Machine Translation, European Association for Machine Translation, Lisboa, Portugal, 2020, pp. 479–480. URL: https://aclanthology.org/2020.eamt-1.61.