Записки научных
семинаров ПОМИ
Том 499, 2021 г.

A. M. Alekseev, S. I. Nikolenko

RECOVERING WORD FORMS BY CONTEXT FOR MORPHOLOGICALLY RICH LANGUAGES

Abstract. In this work, we focus on "sentence-level unlemmatization", the task of generating a grammatical sentence given a lemmatized one, which can usually be easily done by humans. We treat this setting as a machine translation problem and – as a first try – apply a sequence-to-sequence model to the texts of Russian Wikipedia articles, evaluate the effect of the different training sets sizes quantitatively and achieve the BLUE score of 67, 3 using the largest training set available. We discuss preliminary results and flaws of traditional machine translation evaluation methods for this task and suggest directions for future research.

§1. Introduction and related work

Different word forms in morphologically rich languages such as Russian may indicate grammatical and semantic relations between words in a sentence. However, given a lemmatized sentence (which usually does not sound "grammatical" or "comprehensible"), it is usually possible for a human to suggest word forms so that the sentence starts making sense. Sometimes there might be several possible ways to "unlemmatize" a sentence, but their number is almost always very limited compared to the exponential number of all possible combinations of all forms of all words the sentence consists of. We do not require a perfect match with the ground truth since that often requires a larger context; e.g., tenses of verbs can sometimes be derived from context but remain ambiguous in a single sentence, so in this task we dropped the requirement to reconstruct the tense.

This "decoding" problem of suggesting a set of word forms so that the sentence would become grammatical (and ultimately meaningful) could be modeled using recent developments in machine learning.

The two natural variations are:

Key words and phrases: deep learning, natural language processing, morphological agreement, machine translation.

This research was supported by the St. Petersburg State University, research project "Artificial Intelligence and Data Science: Theory, Technology, Industrial and Interdisciplinary Research and Applications".

(1) finding all possible sentence-level "unlemmatizations", probably in the form of a list of most likely paths through a word network, similar to the approach usually taken in speech recognition;
(2) generating at least one possible answer that human readers would consider acceptable.

In this work, we focus on the latter setting, as it seems to be a good start and a more natural formulation for applying the sequence-to-sequence approach directly. Treating various problems in the field of text processing as machine translation ones has proved successful in numerous cases; see, e.g., [3, 10].

While we view this problem more as a step towards a better understanding of morphology for morphology-rich languages, it can have direct applications as well. For instance, it can be used to automatically correct the results of third-party machine translation models, improve the results of dialog and conversational models (again as a post-processing improvement step), or make the results of models that operate with lemmatized text readable for humans. Methods developed in the pursuit of this problem may find applications in the field of native advertising, or as part of extractive summarization or machine translation systems. A model capable to generate comprehensible texts can also be used for data augmentation in statistical natural language processing for approaches dealing with different forms of words, e.g., at a character/subword level. E.g., as shown with the example of the *Concorde* model [12], a separate lemmatization-unlemmatization step can enhance the quality of a conversational model.

In this work, we consider the task of generating grammatical and/or comprehensible (we leave this subtle distinction open for debate) sentences given lemmatized sentences, show the initial approach where we treat it as a machine translation problem, report preliminary results with a sequence-to-sequence model, discuss evaluation challenges for such systems, list several ideas on possible approaches, and provide more ideas on the "guided" grammatical text generation.

The only similar work we are aware of is [12], where the authors show that operating with lemmatized sentences as an input to certain conversational models may improve conversational models and introduce their own model encoding the texts. The datasets in the work [12] consist of the short replies (e.g., replies in subtitles' dialogues), whereas we have decided to focus on longer sentences extracted from the Russian Wikipedia. To demonstrate the approach might prove useful whether the morphological

information is used explicitly (as in e.g. [1]) or not, in this paper we have worked on the words level only.

§2. Task and dataset

The task is inverse to lemmatization: given a sentence (preferably in a morphologically rich language, e.g., Russian) which has previously been lemmatized or artificially comprised of lemmatized words on purpose, provide a text with the same words but word forms chosen so that the text becomes grammatical and/or comprehensible for human readers. For example, in Russian "в 1987 год с юношеский сборная ссср выигрывать чемпионат мир" should become "в 1987 году с юношеской сборной ссср выиграл чемпионат мира" (won the World Championship in 1987 with the USSR Youth national team).

One advantage of this task is that it is quite easy to construct large datasets for training: one can simply use an existing lemmatizer and not worry too much about a rare lemmatization error since this will only make the problem a bit deeper and more interesting.

For this preliminary study, we took a snapshot of the Russian language *Wikipedia*, tokenized it and split into sentences, for a total of $\sim$ 900K articles and $\sim$ 12.5M sentences. Then we used the *Yandex.Mystem* morphological analyzer [13] for lemmatization, specifically the Python interface [14]). The test set consisted of 1000 pairs of sentences, chosen randomly once. Train set sizes varied from 10K to 2M pairs of sentences (see the table with results). The larger train sets include the sentences present in the smaller ones. We also used the Moses SMT engine [6] data preparation scripts to prepare the corpora.

§3. Model and evaluation

To train a machine learning model that tries to recover the original sentence given a lemmatized one, we suggest the following basic approach: apply sequence-to-sequence machine translation methods treating normalized sentences as "source language" and original sentences as "target language". For this preliminary report, we used the default OpenNMT model [5] that employs stacked LSTM [4] with attention based on [2,9] (source and target word embeddings of dim. 500, 2 layers in encoder and 2 layers in decoder, etc.). The model was implemented using PyTorch [11]. Both source and target vocabularies consisted of 50 000 tokens.

Table 1. Examples For Different Training Set Sizes

lemmatized	дарио оставлять жена в панама и предпринимать путешествие в столица аргентина
true answer	дарио оставил жену в панаме и предпринял путешествие в столицу аргентины
wiki-10k	центральная тела премии в восточной и 1920 на сторону коми
wiki-100k	дарио оставил жене в панаме и предпринял путешествие в столицу аргентины
lemmatized	в тот же год занимать пост министр внутренний дело
true answer	sв том же году занимает пост министра внутренних дел
wiki-10k	в том же году занял принял франции письмо государства
wiki-100k	в том же году занял пост министра внутренних дел
lemmatized	лейтенант николай пас отличаться 26 июль 1944 год в бой у река прут
true answer	лейтенант николай пасов отличился 26 июля 1944 года в бою у реки прут
wiki-10k	появилась депутатом взрослого дивизии 26 июля 1944 года в бою у реки среднегодового
wiki-100k	лейтенант николай пас отличился 26 июля 1944 года в боях у реки прут
lemmatized	после обработка он принимать зеленоватый оттенок
true answer	после обработки он принимает зеленоватый оттенок
wiki-10k	после завершения он принял хозяйство львовуголь
wiki-100k	после обработки он принял зеленоватый оттенок

We performed experiments with varying training set sizes. Table 2 and Figure 1 shows the numerical results, which clearly and expectedly show that the larger the corpus, the higher the resulting BLEU scores are. On the qualitative side, Table 1 indicates that "translations" indeed get much better pretty fast with growing training size, and even very modest datasets (100K sentences) can lead to good results.

§4. Challenges and future work

The first challenge lies in evaluation metrics. BLEU expects the true answer and prediction to match word by word. However, if, for example, the predicted sentence is in past tense and the ground truth is in the present

Table 2. BLEU Scores And N-gram Precisions For Different Corpora Sizes

Dataset	BLEU	Precision			
		Unigram	2-gram	3-gram	4-gram
wiki-10k	10.86	44.2	17.0	8.2	4.0
wiki-30k	38.60	66.8	44.7	31.9	23.4
wiki-100k	49.20	73.9	55.3	42.9	33.8
wiki-300k	56.64	77.5	61.9	50.8	42.2
wiki-1M	64.72	82.3	69.3	59.7	51.5
wiki-2M	67.30	84.1	71.9	62.5	54.7

tense, the generated one may be perfectly grammatical while having some very different word forms, and evaluation should not treat this as an error.

As one possible way to approximate the "grammaticality" of the resulting text, we propose to use proxies from other NLP models. For example, "grammaticality" could be estimated using the

(1) probabilities of generated sentences computed by some gold standard language model, or
(2) probabilities of generated sentences computed by a probabilistic POS-tagging model.

We have analyzed the errors produced by the proposed model. Some characteristic examples are shown in Table 1. We see that the `wiki-100k` model produces excellent results, but they still often do not match the ground truth exactly due to multiple possible reconstructions. Note that the model trained on a small dataset often simply cannot reproduce the input and works more like a language model.

As for further work on the models, we put forth the following suggestions:

(1) apply statistical machine translation methods – they are cheaper and faster, and they may prove to be as effective as NMT for the task;
(2) do a thorough analysis of errors and find their possible causes;
(3) apply character-level machine translation models, e.g., [8];
(4) apply state-of-the-art models in machine translation, e.g., [15];
(5) construct a large training corpus based on Russian fiction texts or user-generated texts from social networks, as "Wikipedia language" is rather specific;

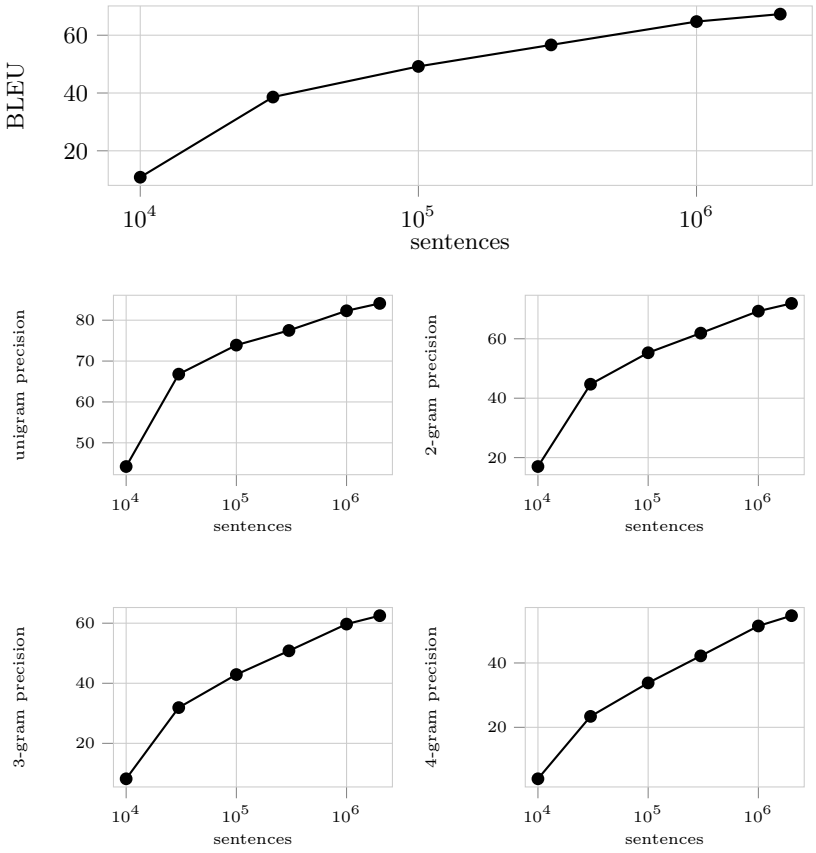

Figure 1. Quality scores: BLEU and n-gram precisions for different corpora sizes.

(6) experiment with other morphology-rich languages (German, French, Turkish etc.);
(7) compare model results with what is possible for humans through a crowdsourced evaluation; this may alleviate the problem of multiple correct answers by providing a golden yardstick to measure the models against.

Moreover, since "sentence-level unlemmatization" does not require the model to change the word order in "translated sentences", neural machine translation with attention might be an overkill. Hence, another option is to try decoding approaches that do not do any alignment, e.g., sequence learning methods with grammatical information as a hidden state provided we can put any word into a chosen form with external tools such as [7] for the Russian language.

References

1. I. Anisimov, V. Polyakov, E. Makarova, and V. Solovyev, *Spelling correction in english: Joint use of bi-grams and chunking*, 2017 Intelligent Systems Conference (IntelliSys), IEEE, 2017, pp. 886–892.
2. D. Bahdanau, K. Cho, and Y. Bengio, *Neural machine translation by jointly learning to align and translate*, arXiv preprint arXiv:1409.0473 (2014).
3. D. Gavrilov, P. Kalaidin, and V. Malykh, *Self-attentive model for headline generation*, CoRR **abs/1901.07786** (2019).
4. S. Hochreiter and J. Schmidhuber, *Long short-term memory*, Neural Comput. **9** (1997), no. 8, 1735–1780.
5. G. Klein, Y. Kim, Y. Deng, J. Senellart, and A. M. Rush, *OpenNMT: Open-Source Toolkit for Neural Machine Translation*, ArXiv e-prints.
6. P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, Chr. Moran, Zens R., et al., *Moses: Open source toolkit for statistical machine translation*, Proc. the 45th annual meeting of the ACL on interactive poster and demonstration sessions, Association for Computational Linguistics, 2007, pp. 177–180.
7. M. Korobov, *Morphological analyzer and generator for russian and ukrainian languages*, Analysis of Images, Social Networks and Texts (M. Yu. Khachay, N. Konstantinova, A. Panchenko, D. I. Ignatov, V. G. Labunets, eds.), Communications in Computer and Information Science, vol. 542, Springer International Publishing, 2015, pp. 320–332 (English).
8. J. Lee, K. Cho, T. Hofmann, *Fully character-level neural machine translation without explicit segmentation*, Transactions of the Association for Computational Linguistics **5** (2017), 365–378.
9. M.-T. Luong, H. Pham, C. D. Manning, *Effective approaches to attention-based neural machine translation*, arXiv preprint arXiv:1508.04025 (2015).
10. Z. Miftahutdinov, E. Tutubalina, *Deep learning for ICD coding: Looking for medical concepts in clinical documents in english and in French*, Experimental IR Meets Multilinguality, Multimodality, and Interaction (Cham) (P. Bellot, Ch. Trabelsi, J. Mothe, F. Murtagh, J. Y. Nie, L. Soulier, E. SanJuan, Linda Cappellato, and Nicola Ferro, eds.), Springer International Publishing, 2018, pp. 203–215.
11. A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z DeVito, Z. Lin, A. Desmaison, L. Antiga, A. Lerer, *Automatic differentiation in pytorch*, NIPS-W, 2017.
12. D. Polykovskiy, D. Soloviev, S. Nikolenko, *Concorde: Morphological agreement in conversational models*, Proc. The 10th Asian Conference on Machine Learning (Jun

Zhu and Ichiro Takeuchi, eds.), Proceedings of Machine Learning Research, vol. 95, PMLR, pp. 407–421.
13. I. Segalovich, *A fast morphological algorithm with unknown word guessing induced by a dictionary for a web search engine.*, MLMTA, Citeseer, 2003, pp. 273–280.
14. D. Sukhonin, A. Panchenko, *A python wrapper of the yandex mystem 3.1 morphological analyzer*, `https://github.com/nlpub/pymystem3`, 2013.
15. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, *Attention is all you need*, arXiv (2017).

Поступило 2 октября 2020 г.

St. Petersburg Department of
Steklov Institute of Mathematics,
St. Petersburg, Russia
E-mail: `anton.m.alexeyev@gmail.com`

St. Petersburg State University
7/9 Universitetskaya nab.,
St. Petersburg, 199034 Russia;
St. Petersburg Department of
Steklov Institute of Mathematics,
St. Petersburg, Russia
E-mail: `s.nikolenko@spbu.ru, sergey@logic.pdmi.ras.ru`