To Lemmatize or Not to Lemmatize: How Word Normalisation Affects ELMo Performance in Word Sense Disambiguation

Andrey Kutuzov*
University of Oslo
Oslo, Norway
andreku@ifi.uio.no

Elizaveta Kuzmenko
University of Trento
Trento, Italy
lizaku77@gmail.com

Abstract

In this paper, we critically evaluate the widespread assumption that deep learning NLP models do not require lemmatized input. To test this, we trained versions of contextualised word embedding *ELMo* models on raw tokenized corpora and on the corpora with word tokens replaced by their lemmas. Then, these models were evaluated on the word sense disambiguation task. This was done for the English and Russian languages.

The experiments showed that while lemmatization is indeed not necessary for English, the situation is different for Russian. It seems that for rich-morphology languages, using lemmatized training and testing data yields small but consistent improvements: at least for word sense disambiguation. This means that the decisions about text pre-processing before training *ELMo* should consider the linguistic nature of the language in question.

1 Introduction

Deep contextualised representations of linguistic entities (words and/or sentences) are used in many current state-of-the-art NLP systems. The most well-known examples of such models are arguably *ELMo* (Peters et al., 2018) and *BERT* (Devlin et al., 2019).

A long-standing tradition if the field of applying deep learning to NLP tasks can be summarised as follows: as minimal pre-processing as possible. It is widely believed that lemmatization or other text input normalisation is not necessary. Advanced neural architectures based on character input (CNNs, BPE, etc) are supposed to be able to learn how to handle spelling and morphology variations themselves, even for languages with rich morphology: 'just add more layers!'. Contextualised embedding models follow this tradition: as a rule, they are trained on raw text collections, with minimal linguistic pre-processing. Below, we show that this is not entirely true.

It is known that for the previous generation of word embedding models ('static' ones like *word2vec* (Mikolov et al., 2013), where a word always has the same representation regardless of the context in which it occurs), lemmatization of the training and testing data improves their performance. Fares et al. (2017) showed that this is true at least for semantic similarity and analogy tasks.

In this paper, we describe our experiments in finding out whether lemmatization helps modern contextualised embeddings (on the example of *ELMo*). We compare the performance of *ELMo* models trained on the same corpus before and after lemmatization. It is impossible to evaluate contextualised models on 'static' tasks like lexical semantic similarity or word analogies. Because of this, we turned to **word sense disambiguation in context** (WSD) as an evaluation task.

In brief, we use contextualised representations of ambiguous words from the top layer of an *ELMo* model to train word sense classifiers and find out whether using lemmas instead of tokens helps in this task (see Section 5). We experiment with the English and Russian languages and show that they differ significantly in the influence of lemmatization on the WSD performance of ELMo models.

Our findings and the contributions of this paper are:

1. Linguistic text pre-processing still matters in some tasks, even for contemporary deep representation learning algorithms.

2. For the Russian language, with its rich mor-

*Both authors contributed equally to the paper.

Proceedings of the First NLPL Workshop on Deep Learning for Natural Language Processing, pages 22–28

	English	Russian
Source	Wikipedia	Wikipedia + RNC
Size, tokens	2 174 mln	989 mln
Size, lemmas	1 977 mln	988 mln

Table 1: Training corpora

phology, lemmatizing the training and testing data for *ELMo* representations yields small but consistent improvements in the WSD task. This is unlike English, where the differences are negligible.

2 Related work

ELMo contextual word representations are learned in an unsupervised way through language modelling (Peters et al., 2018). The general architecture consists of a two-layer BiLSTM on top of a convolutional layer which takes character sequences as its input. Since the model uses fully character-based token representations, it avoids the problem of out-of-vocabulary words. Because of this, the authors explicitly recommend not to use any normalisation except tokenization for the input text. However, as we show below, while this is true for English, for other languages feeding *ELMo* with lemmas instead of raw tokens can improve WSD performance.

Word sense disambiguation or WSD (Navigli, 2009) is the NLP task consisting of choosing a word sense from a pre-defined sense inventory, given the context in which the word is used. WSD fits well into our aim to intrinsically evaluate *ELMo* models, since solving the problem of polysemy and homonymy was one of the original promises of contextualised embeddings: their primary difference from the previous generation of word embedding models is that contextualised approaches generate different representations for homographs depending on the context. We use two lexical sample WSD test sets, further described in Section 4.

3 Training ELMo

For the experiments described below, we trained our own *ELMo* models from scratch. For English, the training corpus consisted of the English Wikipedia dump[1] from February 2017. For Russian, it was a concatenation of the Russian Wikipedia dump from December 2018 and the full Russian National Corpus[2] (RNC). The RNC texts were added to the Russian Wikipedia dump so as to make the Russian training corpus more comparable in size to the English one (Wikipedia texts would comprise only half of the size). As Table 1 shows, the English Wikipedia is still two times larger, but at least the order is the same.

The texts were tokenized and lemmatized with the *UDPipe* models for the respective languages trained on the Universal Dependencies 2.3 treebanks (Straka and Straková, 2017). *UDPipe* yields lemmatization accuracy about 96% for English and 97% for Russian[3]; thus for the task at hand, we considered it to be gold and did not try to further improve the quality of normalisation itself (although it is not entirely error-free, see Section 4).

ELMo models were trained on these corpora using the original TensorFlow implementation[4], for 3 epochs with batch size 192, on two GPUs. To train faster, we decreased the dimensionality of the LSTM layers from the default 4096 to 2048 for all the models.

4 Word sense disambiguation test sets

We used two WSD datasets for evaluation:

- *Senseval-3* for English (Mihalcea et al., 2004)
- *RUSSE'18* for Russian (Panchenko et al., 2018)

The *Senseval-3* dataset consists of lexical samples for nouns, verbs and adjectives; we used only noun target words:

1. *argument*
2. *arm*
3. *atmosphere*
4. *audience*
5. *bank*
6. *degree*
7. *difference*
8. *difficulty*

[1] https://dumps.wikimedia.org/

[2] http://ruscorpora.ru/en/

[3] http://ufal.mff.cuni.cz/udpipe/models#universal_dependencies_23_models

[4] https://github.com/allenai/bilm-tf

9. *disc*
10. *image*
11. *interest*
12. *judgement*
13. *organization*
14. *paper*
15. *party*
16. *performance*
17. *plan*
18. *shelter*
19. *sort*
20. *source*

An example for the ambiguous word *argument* is given below:

In some situations Postscript can be faster than the escape sequence type of printer control file. It uses post fix notation, where ***arguments*** *come first and operators follow. This is basically the same as Reverse Polish Notation as used on certain calculators, and follows directly from the stack based approach.*

It this sentence, the word '*argument*' is used in the sense of a mathematical operator.

The *RUSSE'18* dataset was created in 2018 for the shared task in Russian word sense induction. This dataset contains only nouns; the list of words with their English translations is given in Table 2.

Originally, it includes also the words байка 'tale/fleece' and гвоздика 'clove/small nail', but their senses are ambiguous only in some inflectional forms (not in lemmas), therefore we decided to exclude these words from evaluation.

The Russian dataset is more homogeneous compared to the English one, as for all the target words there is approximately the same number of context words in the examples. This is achieved by applying the lexical window (25 words before and after the target word) and cropping everything that falls outside of that window. In the English dataset, on the contrary, the whole paragraph with the target word is taken into account. We have tried cropping the examples for English as well, but it did not result in any change in the quality of classification. In the end, we decided not to apply the lexical window to the English dataset so as not to alter it and rather use it in the original form.

Target word	**Translation**
акция	'stock/marketing event'
гипербола	'hyperbola/exaggeration'
град	'hail/city'
гусеница	'caterpillar/track'
домино	'dominoes/costume'
кабачок	'squash/restaurant'
капот	'hood (part of a car/clothing)'
карьер	'mine/fast pace of a horse'
кок	'cook/hairstyle'
крона	'crown (tree/coin)'
круп	'crupper (part of a horse/illness)'
мандарин	'fruit/a Chinese official'
рок	'rock (music/destiny)'
слог	'syllable/text style'
стопка	'stack/glass'
таз	'basin/human body part'
такса	'tariff/dog breed'
шах	'check/prince'

Table 2: Target ambiguous words for Russian (*RUSSE'18*)

Here is an example from the *RUSSE'18* for the ambiguous word мандарин 'mandarin' in the sense 'Chinese official title':

"...дипломатического корпуса останкам богдыхана и императрицы обставлено было с необычайной торжественностью. Тысячи мандаринов и других высокопоставленных лиц разместились шпалерами на трех мраморных террасах ведущих к..."

'...the diplomatic bodies of the Bogdikhan and the Empress was furnished with extraordinary solemnity. Thousands of mandarins and other dignitaries were placed on three marble terraces leading to...'.

Table 3 compares both datasets. Before usage, they were pre-processed in the same way as the training corpora for *ELMo* (see Section 3), thus producing a lemmatized and a non-lemmatized versions of each.

As we can see from Table 3, for 20 target words in English there are 24 lemmas, and for 18 target words in Russian there are 36 different lemmas. These numbers are explained by occasional errors in the *UDPipe* lemmatization. Another interesting thing to observe is the number of distinct

Property	Senseval-3	RUSSE'18
Target words	20	18
Distinct target forms	39	132
Distinct target lemmas	24	36
Examples per target	171	126
Tokens per example	126	25
Senses per target	6	2

Table 3: Characteristics of the WSD datasets. The numbers in the lower part are average values.

Model	English	Russian
	Baselines	
Random	≈ 0.138	≈ 0.444
MFS	0.119	0.391
	Tokens	
SGNS (averaged)	0.299	0.851
ELMo (averaged)	0.362	0.885
ELMo (target)	0.463	0.875
	Lemmas	
SGNS (averaged)	0.300	0.854
ELMo (averaged)	0.365	0.888
ELMo (target)	0.452	**0.907**

Table 4: Averaged macro-F1 scores for WSD

word forms for every language. For English, there are 39 distinct forms for 20 target nouns: singular and plural for every noun, except '*atmosphere*' which is used only in the singular form. Thus, inflectional variability of English nouns is covered by the dataset almost completely. For Russian, we observe 132 distinct forms for 18 target nouns, giving more than 7 inflectional forms per each word. Note that this still covers only half of all the inflectional variability of Russian: this language features 12 distinct forms for each noun (6 cases and 2 numbers).

To sum up, the *RUSSE'18* dataset is morphologically far more complex than the *Senseval3*, reflecting the properties of the respective languages. In the next section we will see that this leads to substantial differences regarding comparisons between token-based and lemma-based *ELMo* models.

5 Experiments

Following Gorman and Bedrick (2019), we decided to avoid using any standard train-test splits for our WSD datasets. Instead, we rely on per-word random splits and 5-fold cross-validation. This means that for each target word we randomly generate 5 different divisions of its context sentences list into train and test sets, and then train and test 5 different classifier models on this data. The resulting performance score for each target word is the average of 5 macro-F1 scores produced by these classifiers.

ELMo models can be employed for the WSD task in two different ways: either by fine-tuning the model or by extracting word representations from it and then using them as features in a downstream classifier. We decided to stick to the second (feature extraction) approach, since it is conceptually and computationally simpler. Additionally, Peters et al. (2019) showed that for most NLP tasks (except those focused on sentence pairs) the performance of feature extraction and fine-tuning is nearly the same. Thus we extracted the single vector of the target word from the *ELMo* top layer ('target' rows in Table 4) or the averaged *ELMo* top layer vectors of all words in the context sentence ('averaged' rows in Table 4).

For comparison, we also report the scores of the 'averaged vectors' representations with Continuous Skipgram (Mikolov et al., 2013) embedding models trained on the English or Russian Wikipedia dumps ('SGNS' rows): before the advent of contextualised models, this was one of the most widely used ways to 'squeeze' the meaning of a sentence into a fixed-size vector. Of course it does not mean that the meaning of a sentence always determines the senses all its words are used in. However, averaging representations of words in contexts as a proxy to the sense of one particular word is a long established tradition in WSD, starting at least from Schütze (1998). Also, since SGNS is a 'static' embedding model, it is of course not possible to use only target word vectors as features: they would be identical whatever the context is.

Simple logistic regression was used as a classification algorithm. We also tested a multi-layer perceptron (MLP) classifier with 200-neurons hidden layer, which yielded essentially the same results. This leads us to believe that our findings are not classifier-dependent.

Table 4 shows the results, together with the ran-

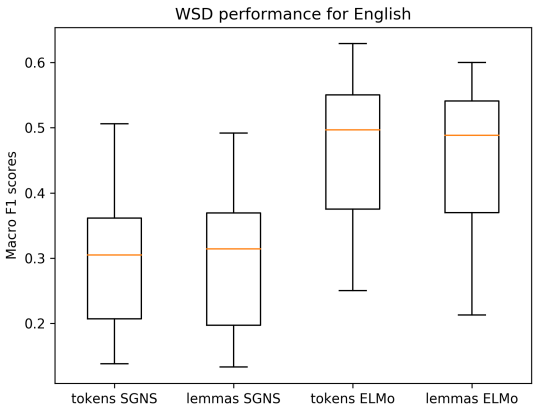

Figure 1: Word sense disambiguation performance on the **English** data across words (*ELMo* target models).

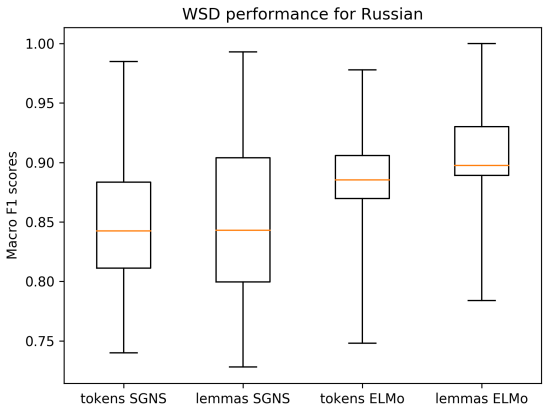

Figure 2: Word sense disambiguation performance on the **Russian** data across words (*ELMo* target models).

dom and most frequent sense (MFS) baselines for each dataset.

First, *ELMo* outperforms SGNS for both languages, which comes as no surprise. Second, the approach with averaging representations from all words in the sentence is not beneficial for WSD with *ELMo*: for English data, it clearly loses to a single target word representation, and for Russian there are no significant differences (and using a single target word is preferable from the computational point of view, since it does not require the averaging operation). Thus, below we discuss only the single target word usage mode of *ELMo*.

But the most important part is the comparison between using tokens or lemmas in the train and test data. For the 'static' SGNS embeddings, it does not significantly change the WSD scores for both languages. The same is true for English *ELMo* models, where differences are negligible and seem to be simple fluctuations. However, for Russian, *ELMo* (target) on lemmas outperforms *ELMo* on tokens, with small but significant[5] improvement. The most plausible explanation for this is that (despite of purely character-based input of *ELMo*) the model does not have to learn idiosyncrasies of a particular language morphology. Instead, it can use its (limited) capacity to better learn lexical semantic structures, leading to better WSD performance. The box plots 1 and 2 illustrate the scores dispersion across words in the test sets for English and Russian correspondingly (orange lines are medians). In the next section 6 we analyse the results qualitatively.

[5] At p value of 0.1, according to the Welch's t-test.

Word	Tokens	Lemmas	STD
акция	0.876	**0.978**	0.050
крона	0.978	**1.000**	0.018
круп	0.927	**1.000**	0.070
домино	**0.910**	0.874	0.057

Table 5: F1 scores for target words from *RUSSE'18* with significant differences between lemma-based and token-based models

6 Qualitative analysis

In this section we focus on the comparison of scores for the Russian dataset. The classifier for Russian had to choose between fewer classes (two or three), which made the scores higher and more consistent than for the English dataset. Overall, we see improvements in the scores for the majority of words, which proves that lemmatization for morphologically rich languages is beneficial.

We decided to analyse more closely those words for which the difference in the scores between lemma-based and token-based models was statistically significant. By 'significant' we mean that the scores differ by more that one standard deviation (the largest standard deviation value in the two sets was taken). The resulting list of targets words with significant difference in scores is given in Table 5.

We can see that among 18 words in the dataset only 3 exhibit significant improvement in their scores when moving from tokens to lemmas in the input data. It shows that even though the over-

all F1 scores for the Russian data have shown the plausibility of lemmatization, this improvement is mostly driven by a few words. It should be noted that these words' scores feature very low standard deviation values (for other words, standard deviation values were above 0.1, making F1 differences insignificant). Such a behaviour can be caused by more consistent differentiation of context for various senses of these 3 words. For example, with the word кабачок 'squash / small restaurant', the contexts for both senses can be similar, since they are all related to food. This makes the WSD scores unstable. On the other hand, for акция 'stock, share / event', крона 'crown (tree / coin)' or круп 'croup (horse body part / illness)', their senses are not related, which resulted in more stable results and significant difference in the scores (see Table 5).

There is only one word in the RUSSE'18 dataset for which the score has strongly decreased when moving to lemma-based models: домино 'domino (game / costume)'. In fact, the score difference here lies on the border of one standard deviation, so strictly speaking it is not really significant. However, the word still presents an interesting phenomenon.

Домино is the only target noun in the RUSSE'18 that has no inflected forms, since it is a borrowed word. This leaves no room for improvement when using lemma-based *ELMo* models: all tokens of this word are already identical. At the same time, some information about inflected word forms in the context can be useful, but it is lost during lemmatization, and this leads to the decreased score. Arguably, this means that lemmatization brings along both advantages and disadvantages for WSD with *ELMo*. For inflected words (which constitute the majority of Russian vocabulary) profits outweigh the losses, but for atypical non-changeable words it can be the opposite.

The scores for the excluded target words байка 'tale / fleece' and гвоздика 'clove / small nail' are given in Table 6 (recall that they were excluded because of being ambiguous only in some inflectional forms). For these words we can see a great improvement with lemma-based models. This, of course stems from the fact that these words in different senses have different lemmas. Therefore, the results are heavily dependent on the quality of lemmatization.

Word	Tokens	Lemmas	STD
байка	0.421	**0.627**	0.099
гвоздика	0.553	**0.619**	0.038

Table 6: F1 scores for the excluded target words from *RUSSE'18*.

7 Conclusion

We evaluated how the ability of *ELMo* contextualised word embedding models to disambiguate word senses depends on the nature of the training data. In particular, we compared the models trained on raw tokenized corpora and those trained on the corpora with word tokens replaced by their normal forms (lemmas). The models we trained are publicly available via the NLPL word embeddings repository[6] (Fares et al., 2017).

In the majority of research papers on deep learning approaches to NLP, it is assumed that lemmatization is not necessary, especially when using powerful contextualised embeddings. Our experiments show that this is indeed true for languages with simple morphology (like English). However, for rich-morphology languages (like Russian), using lemmatized training data yields small but consistent improvements in the word sense disambiguation task. These improvements are not observed for rare words which lack inflected forms; this further supports our hypothesis that better WSD scores of lemma-based models are related to them better handling multiple word forms in morphology-rich languages.

Of course, lemmatization is by all means not a silver bullet. In other tasks, where inflectional properties of words are important, it can even hurt the performance. But this is true for any NLP systems, not only deep learning based ones.

The take-home message here is twofold: first, text pre-processing still matters for contemporary deep learning algorithms. Their impressive learning abilities do not always allow them to infer normalisation rules themselves, from simply optimising the language modelling task. Second, the nature of language at hand matters as well, and differences in this nature can result in different decisions being optimal or sub-optimal at the stage of deep learning models training. The simple truth 'English is not representative of all languages on Earth' still holds here.

[6] http://vectors.nlpl.eu/repository/

In the future, we plan to extend our work by including more languages into the analysis. Using Russian and English allowed us to hypothesise about the importance of morphological character of a language. But we only scratched the surface of the linguistic diversity. To verify this claim, it is necessary to analyse more strongly inflected languages like Russian as well as more weakly inflected (analytical) languages similar to English. This will help to find out if the inflection differences are important for training deep learning models across human languages in general.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Murhaf Fares, Andrey Kutuzov, Stephan Oepen, and Erik Velldal. 2017. Word vectors, reuse, and replicability: Towards a community repository of large-text resources. In *Proceedings of the 21st Nordic Conference on Computational Linguistics*, pages 271–276, Gothenburg, Sweden. Association for Computational Linguistics.

Kyle Gorman and Steven Bedrick. 2019. We need to talk about standard splits. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2786–2791, Florence, Italy. Association for Computational Linguistics.

Rada Mihalcea, Timothy Chklovski, and Adam Kilgarriff. 2004. The Senseval-3 English lexical sample task. In *Proceedings of SENSEVAL-3, the Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, pages 25–28, Barcelona, Spain. Association for Computational Linguistics.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems 26*, pages 3111–3119.

Roberto Navigli. 2009. Word sense disambiguation: A survey. *ACM computing surveys (CSUR)*, 41(2):10.

Alexander Panchenko, Anastasia Lopukhina, Dmitry Ustalov, Konstantin Lopukhin, Nikolay Arefyev, Alexey Leontyev, and Natalia Loukachevitch. 2018. RUSSE'2018: A Shared Task on Word Sense Induction for the Russian Language. In *Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference "Dialogue"*, pages 547–564, Moscow, Russia. RSUH.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Matthew E. Peters, Sebastian Ruder, and Noah A. Smith. 2019. To tune or not to tune? Adapting pretrained representations to diverse tasks. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pages 7–14, Florence, Italy. Association for Computational Linguistics.

Hinrich Schütze. 1998. Automatic word sense discrimination. *Computational Linguistics*, 24(1):97–123.

Milan Straka and Jana Straková. 2017. Tokenizing, POS tagging, lemmatizing and parsing UD 2.0 with UDPipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99, Vancouver, Canada. Association for Computational Linguistics.