Toward Language Modeling for the Ukrainian Language

Anastasiia Khaburska[1] and Igor Tytyk[2]

[1] Ukrainian Catholic University, Faculty of Applied Science, Lviv, Ukraine
a.khaburska@ucu.edu.ua
[2] ProWritingAid, London, UK
igor.tytyk@gmail.com

Abstract. Language Modeling is one of the most important subfields of modern Natural Language Processing (NLP). The objective of language modeling is to learn a probability distribution over sequences of linguistic units pertaining to the language. As it produces a probability of the language unit that will follow, the language model can be viewed as a form of grammar for the language, and it plays a key role in traditional NLP tasks, such as speech recognition, machine translation, sentiment analysis, text summarization, grammatical error correction, natural language generation. Much work has been done for the English language in terms of developing both training and evaluation approaches. However, there has not been as much progress for the Ukrainian language. In this work, we are going to explore, extend, evaluate, and compare different language models for the Ukrainian language. The main objective is to provide a balanced evaluation data set and train a number of baseline models.

Keywords: Language Modeling · Natural Language Processing · Ukrainian Language · Language Corpus

1 Introduction

The objective of Language Modeling is to learn a probability distribution over sequences of linguistic units pertaining to a language. As **linguistic units**, we can consider any natural units into which linguistic messages can be divided, for example, characters, words, or phrases. These linguistic units, seen by the model, compose model's dictionary U:

$$P(S) = P(u_1, u_2, \dots, u_n), \tag{1}$$

where S – a sequence of linguistic units and u_i - i-th unit.

Typically, this is achieved by providing conditional probabilities $p(u|c)$, where c is the context of linguistic unit u. For example, the probability of a particular unit in the sequence:

$$P(u_i | u_{i-k_i}, u_{i-k_1+1}, \dots, u_{i+k_2-1}, u_{i+k_2}) \tag{2}$$

In: Proceedings of the 1st Masters Symposium on Advances in Data Mining, Machine Learning, and Computer Vision (MS-AMLV 2019), Lviv, Ukraine, November 15-16, 2019, pp. 71–80

Most fixed-vocabulary language models employ a distinguished symbol $< unk >$ that represents all units not present in vocabulary U. These units are termed out-of-vocabulary (OOV).

As it produces a probability of the following language unit, the language model (LM) can be viewed as a grammar of the language and it plays a key role in traditional NLP tasks, such as automatic speech recognition [1, 2], machine translation [3, 4], sentiment analysis [5], text summarization [6, 7], grammatical error correction [8], natural language generation [9].

2 Motivation

Language Modeling is one of the central tasks to Natural Language Processing and Natural Language Understanding. Thus, in order to elaborate upon an NLP task for the language, this language needs to have a well-designed high-quality language model.

As pointed out by Jozefowicz et al. in [10], ”Models which can accurately place distributions over sentences encode not only complexities of language such as grammatical structure, but also distil a fair amount of information about the knowledge that a corpora may contain”.

Furthermore, to train and evaluate language models, it is required to have a well-composed corpus. In linguistics and NLP, corpus refers to a collection of texts. Such collections may be formed of texts in a single language or span multiple languages and domains. In our case, it is very important to evaluate and benchmark the models on the data with balanced genres and topics.

Overall, building a baseline language model and a gold standard corpus for the Ukrainian language is a crucial step in the evolution of Ukrainian NLP.

For the English Language, language modelling went through multiple stages of evolvement. Much work has been done for the English language in terms of developing both training and evaluation approaches. Firstly, count-based approaches (based on statistics of N-grams), such as Kneser-Ney smoothed 5-gram models [11], were used as a fairly strong baseline. In recent years, much progress has been made by neural methods [1, 12], character-aware Neural Language Models [13], based on LSTMs [10], gated convolutional networks [14] and self-attentional networks [15].

At the same time, there has not been as much progress for the Ukrainian language in terms of language modeling. In this master’s thesis, we want to explore, extend, (or maybe develop), evaluate and compare a set of language models for the Ukrainian language. The main objective is to offer an evaluation corpus and set a number of baselines.

3 Goal

The main objective is to offer an evaluation corpus and set a number of baselines:

1. Which data corpus will be sufficient to train language models for the Ukrainian language? How do we need to preprocess available data sets?

2. Which linguistic units represent sequential information from Ukrainian texts more accurately?
3. What approaches and models perform better for the Ukrainian language? (classical probabilistic, n-gram based, neural networks)
4. How to evaluate language models trained for the Ukrainian language? Intrinsic and extrinsic evaluation metrics.

4 Background and Results to Date

In this section, we describe the data sets we are going to train our language models on and explain the models, which we intend to train and evaluate at first. Also, we report our first results.

4.1 Data

Regarding the English language, despite much work being devoted to small data sets like the Penn Tree Bank (PTB) [16], research on larger tasks is very relevant as overfitting is not the main limitation in current language modeling, but is the main characteristic of the PTB task. Results on larger corpora usually show better. Further, given current hardware trends and vast amounts of text available on the Web, it is much more straightforward to tackle large-scale modeling than it used to be. Thus, it would be good for our research to train the language models on large-scale LM benchmark like the One Billion Word Benchmark data set [17]. This data set consists of one thousand fold, 800k word vocabulary and 1B words training data.

For the Ukrainian language, we do not have such a huge, well-redacted, tagged, and well-balanced corpora.

At this stage, we consider three datasets:

- **Ukrainian Brown Corpus**[1] is a well-balanced and redacted corpus of original Ukrainian texts published between 2010 and 2018, comprised of such domains as: 1) news media; 2) religious media; 3) professional literature; 4) aesthetic-informative literature; 5) administrative documents; 6) popular science; 7) science literature; 8) educational literature; 9) fiction writing. Unfortunately, it is comparatively small. We conduct a descriptive analysis of ”Good” and ”So-so” parts of this corpus. This consists of 924 texts, 600810 training words, and 38728 unique lemmas[2].
- **Uber-Text Corpus**[3] contains more than 6 Gb of Ukrainian texts, but unfortunately, because of legal rules, is split into sentences, deprived of punctuation and then shuffled randomly. Thus, only sentence-level sequences may be used to train and evaluate the language models. Dmitry Chaplinsky kindly shared with us 9971 full texts

[1] Ukrainian Brown Corpus: https://github.com/brown-uk/corpus

[2] Git-Hub: https://github.com/Anastasiia-Khab/LMForTheUkrainianLanguage/ blob/master/UkrBrownCorpusAnalysis_good%26soso.ipynb

[3] Uber-Text Corpus: http://lang.org.ua/en/corpora/

from fiction writing and 631935 texts from Korrespondent news media data set. Of course, before using it, we should conduct some preprocessing.

- **Wiki dumps**[4]

4.2 N-gram Language Models

N-gram models are a widely used type of language models. As a rule, they are very straightforward to construct except for the issue of smoothing, a technique used to better estimate probabilities when there is insufficient data to estimate probabilities accurately. Generalizing equation for n-gram model is:

$$p(s) = \prod_{i=1}^{i+1} p\left(u_i \middle| u_{i-n+1}^{i-1}\right), \tag{3}$$

where u_i^j denotes the units u_i ... u_j and where we take u_{-n+2} through u_0 to be <BOS> and u_{i+1} to be <EOS>. To estimate the probabilities:

$$p\left(u_i \middle| u_{i-n+1}^{i-1}\right) = \frac{c(u_{i-n+1}^{i})}{\sum_{u_i} c(u_{i-n+1}^{i})}, \tag{4}$$

where $c(u_{i-n+1}^{i})$ denotes the number of times the n-gram u_i ... u_{i-n+1} occurs in the given corpus. The units u_{i-n+1}^{i-1} preceding the current unit u_i are called the history. The sum $\sum_{u_i} c(u_{i-n+1}^{i})$ is equal to the count of the history $c(u_{i-n+1}^{i})$.

Smoothing is a technique used to adapt the maximum likelihood estimate of probabilities and to make distribution more uniform, by adjusting low probabilities such as zero probabilities upward and high probabilities downward. Smoothing methods generally prevent zero probabilities.

While sparse data is a central issue in n-gram language modeling, an enormous number of techniques have been proposed for smoothing n-gram models. In [18], the authors carried out an extensive empirical comparison of the most widely used smoothing techniques, including those described by [19–22, 11]. They introduced methodologies for analyzing smoothing algorithm performance in detail, and using these techniques they motivate a novel variation of Kneser-Ney smoothing that consistently outperforms all other algorithms evaluated.

This backoff-smoothed model estimates the probability based on the observed entry with longest matching:

$$p\left(u_i \middle| u_1^{i-1}\right) = p\left(u_i \middle| u_f^{i-1}\right) \prod_{n=1}^{f-1} b(u_n^{i-1}), \tag{5}$$

where the probability $p(u_i|u_f^{i-1})$ and back-off penalties $b(u_n^{i-1})$ are given by an already-estimated model.

Open-source KenLM library proposed by [23] efficiently uses two data structures (*PROBING* and *TRIE*) to query n-gram language model with modified Kneser-Ney smoothing, reducing both time and memory costs.

[4] Ukrainian Wiki dumps: https://dumps.wikimedia.org/ukwiki/20190920/

We trained[5] four n-gram language models using KenLM library on the Ukrainian Brown Corpus (length = 817699 units (words + punctuation marks), split into sentences) and evaluated it with the *perplexity* measure (see Tab. 4.2).

Table 1. Perplexity of the KenLM n-gram models trained on Ukrainian Brown Corpus

n-gram model	perplexity
3-gram	18.68
4-gram	12.52
5-gram	11.60
6-gram	11.44

4.3 Neural Networks

Deep Learning has fueled language modeling research in the past years as it allowed researchers to explore many tasks for which the strong conditional independence assumptions are unrealistic. Using artificial neural networks in statistical language modeling has been proposed by [12], who used feedforward neural networks with fixed-length context. This approach was exceptionally successful and further investigation by [24]. Later, [25] has shown that neural network based models provide significant improvements in speech recognition for several tasks against good baseline systems.

If we want to build models that can really learn the language, then online learning is crucial - acquiring new information is definitely important. Simple Recurrent neural network introduced by [1] outperformed state of the art back-off models significantly.

We intend to train the state of the art architectures of Recurrent Neural Network Language Models (RNNLM) and Long-Short term memory Language models (LSTMLM) on Ukrainian Corpus. Also, we would like to combine RNNLM with N-gram models as proposed in [17].

In recent years, strong character-level language models [26], [27] typically follow a common template “truncated backpropagation through time” (TBTT). A recurrent neural net (RNN) is trained over mini-batches of text sequences, using a relatively short sequence length (e.g. 200 tokens). Also, [15] introduced and interesting approach. They show that a non-recurrent model can achieve strong results in character-level language modeling. Specifically, they use a deep network of transformer self-attention layers [4] with causal (backwards-looking) attention to process fixed-length inputs and predict upcoming characters.

We plan to train a character-level model on the Ukrainian Language data in order to test which models (Word-level vs Character-level) are more productive for the Ukrainian language and on what span of text.

[5] Git-Hub: https://github.com/Anastasiia-Khab/LMForTheUkrainianLanguage/
blob/master/KenLM_Sentence-base-tagged.ipynb

5 Methodology

- **Tokenization and lemmatization**: For tokenization and lemmatization, we use the nlp-uk library[6] from Andriy Rysin and the BrUk group.
- **Word embeddings**: For word embeddings we can use lang-uk embeddings[7] or fast-text embeddings[8] calculate embedding in parralel with training a model.
- **Evaluation**: As an evaluation metrics, firstly, we are going to consider perplexity [28].

6 Outline for Master Research and Thesis Completion

10 September - 19 September

- ✓ Write abstract
- ✓ Formulate a rough scope of research and the main objectives
- ✓ Start exploring the data

21 September - 3 October

- ✓ Explore the available data and search for more
- ✓ Run some initial experiments on limited amount of data
- ✓ Write a proposal for the symposium

5 October - 17 October

- ✓ Make sure all the necessary data is in place and preprocessed
- ✓ Formulate a list of experiments
- ✓ Start running experiments: train a baseline n-gram model
- ✓ Test and analyze the evaluation metric and the evaluation set

19 October - 31 October

- Train a neural language model
- Analyze the evaluation results and write conclusions

2 November - 14 November

- Experiment with pre-trained embeddings
- Analyze the evaluation results the results and write conclusions

14 November – 28 November

- Conduct experiments on some advanced ideas if time permits (e.g. language generation; e.g. testing language models on some downstream tasks)

[6] LanguageTool API NLP UK: https://github.com/brown-uk/nlp_uk

[7] Lang-uk embeddings: http://lang.org.ua/en/models/#anchor4

[8] Fasttext embeddings: https://fasttext.cc/docs/en/crawl-vectors.html

30 November - 12 December

- Decide on follow-up experiments and conduct them
- Start structuring the master thesis

14 December - 26 December

- Finalise the diagrams, plots, tables, and figures
- Write the master thesis

28 December - 8 January

- Proofread the thesis and polish the formatting

7 Discussion and Outlook

Modern natural language processing practitioners strive to create modeling techniques that work well on all of the world's languages. For example, Google's Multilingual Neural Machine Translation (NMT) System [29]. Rather than train a full sequence-to-sequence model for every pair of language that they support, which is a tremendous feat in terms of both data and compute time required – they built a single system that can translate between any two languages. This is a sequence-to-sequence model, which accepts as input a sequence of words and a token specifying what language to translate into and uses shared parameters to translate into any target language. The new multilingual model not only improved their translation performance, but also enabled ”zero-shot translation”. For instance, having examples of Norwegian-English and Ukrainian-English translations, Google's multilingual NMT system trained on this data could actually generate reasonable Norwegian-Ukrainian translations, if we lack in training data for those two languages. The powerful implication of this finding is that part of the decoding process is not language-specific, and the model is in fact maintaining an internal representation of the input/output sentences independently of the actual languages involved. This is a domain-specific finding, which is very useful in language translation and does not diminish the importance of having an evaluated language model trained for the Ukrainian language.

Indeed, as mentioned by [30], most methods are portable in the following sense: given appropriately annotated data, these could in principle be trainable in any language. However, despite this crude cross-linguistic compatibility, it is unlikely that all languages are equally easy, or that our methods are equally good at all languages. Furthermore, [30] presents a study on 21 languages, demonstrating that in languages with complex inflectional morphology, the textual expression of the information is harder to predict with both n-gram and LSTM language models. They show complex inflectional morphology to be a cause of performance differences among languages.

Ukrainian is an East Slavic language and is famous for its rich inflexions. It is noted by [31], that the number of inflexions in Ukrainian by far exceeds their number in English since every notional part of speech has a variety of endings. The latter express number, case and gender of nominal parts of speech (nouns, adjectives, numerals, pronouns) and tense, aspect, person, number, voice and mood forms of verbs. Additionally, in the Ukrainian language any part of speech may form diminutive forms of the word, while in English only nouns have this possibility.

We consider experimenting with multilingual language modeling or sharing model parameters from the models trained on structurally similar languages, for example, Polish, Russian, Slovak, or Belarusian languages. Then, we would compare this model with the other models using our evaluation techniques.

8 Acknowledgments

First of all, I would like to thank my supervisor Igor Tytyk who directed me throughout the research for this proposal and provided a lot of useful ideas and pieces of advice regarding contents, structure and possible future development of the master thesis. Special thanks to Artem Chernodub (Ukrainian Catholic University, Grammarly) and Grammarly itself for the motivation, computational resources, and consultation. This work would not be possible without Ukrainian Brown Corpus. Thanks to Brown-Uk enthusiastic group[9] for the meticulous work on constructing the corpus. Sincerely wish them success in future developing of their work. Many thanks to Dmitry Chaplinsky, who kindly agreed to share with me the data from Uber Text corpora[10]. Last, but not least, I am grateful to Ukrainian Catholic University and Oleksii Molchanovskyi personally for the first master program in Data Science in Ukraine and to Ciklum for covering my tuition fees.

References

1. Mikolov, T., Karafiát, M., Burget, L., Cernocký, J., Khudanpur, S.: Recurrent neural network based language model. In: Eleventh Annual Conference of the International Speech Communication Association (2010)
2. Arisoy, E., Sainath, T.N., Kingsbury, B., Ramabhadran, B.: Deep neural network language models. In: 2012 NAACL-HLT Workshop: Will We Ever Really Replace the N-gram Model? On the Future of Language Modeling for HLT, pp. 20–28. Association for Computational Linguistics (2012)
3. Schwenk, H., Rousseau, A., Attik, M.: Large, pruned or continuous space language models on a GPU for statistical machine translation. In: 2012 NAACL-HLT Workshop: Will We Ever Really Replace the N-gram Model? On the Future of Language Modeling for HLT, pp. 11-19. Association for Computational Linguistics (2012)
4. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: 30th Annual Conference on Neural Information Processing Systems, pp. 5998–6008 (2017)

[9] Brown-Uk: https://r2u.org.ua/corpus

[10] Uber Text: http://lang.org.ua/en/corpora/

5. Hu, Y., Lu, R., Li, X., Chen, Y., Duan, J.: A language modeling approach to sentiment analysis. In: Shi, Y., van Albada, G.D., Dongarra, J., Sloot, P.M.A. (eds.) ICCS 2007. LNCS, vol. 4488. pp. 1186–1193. Springer, Berlin, Heidelberg (2007)
6. Rush, A.M., Chopra, S., Weston, J.: A neural attention model for abstractive sentence summarization. arXiv preprint arXiv:1509.00685 (2015)
7. Filippova, K., Alfonseca, E., Colmenares, C.A., Kaiser, L., Vinyals, O.: Sentence compression by deletion with lstms. In: 2015 Conference on Empirical Methods in Natural Language Processing, pp. 360–368 (2015)
8. Bryant, C., Briscoe, T.: Language model based grammatical error correction without annotated training data. In: 13th Workshop on Innovative Use of NLP for Building Educational Applications, pp. 247–253 (2018)
9. Edunov, S., Baevski, A., Auli, M.: Pre-trained language model representations for language generation. arXiv preprint arXiv:1903.09722 (2019)
10. Jozefowicz, R., Vinyals, O., Schuster, M., Shazeer, N., Wu, Y.: Exploring the limits of language modeling. arXiv preprint arXiv:1602.02410 (2016)
11. Kneser, R., Ney, H.: Improved backing-off for m-gram language modeling. In: 1995 International Conference on Acoustics, Speech, and Signal Processing. Vol. 1, pp. 181–184. IEEE Press, New York (1995)
12. Bengio, Y., Ducharme, R., Vincent, P., Jauvin, C.: A neural probabilistic language model. Journal of Machine Learning Research **3**(Feb), 1137–1155 (2003)
13. Kim, Y., Jernite, Y., Sontag, D., Rush, A.M.: Character-aware neural language models. In: 13th AAAI Conference on Artificial Intelligence, pp. 2741–2749. AAAI (2016)
14. Dauphin, Y.N., Fan, A., Auli, M., and Grangier, D.: Language modeling with gated convolutional networks. In: 34th International Conference on Machine Learning, pp. 933–941 (2017)
15. Al-Rfou, R., Choe, D., Constant, N., Guo, M., Jones, L.: Character-level language modeling with deeper self-attention. In: 16th AAAI Conference on Artificial Intelligence, pp. 3159–3166. AAAI (2019)
16. Marcus, M., Santorini, B., Marcinkiewicz, M.A.: Building a large annotated corpus of English: The Penn Treebank. Computational Linguistics **19**(2), 313–330 (1993)
17. Chelba, C., Mikolov, T., Schuster, M., Ge, Q., Brants, T., Koehn, P., Robinson, T.: One billion word benchmark for measuring progress in statistical language modeling. arXiv preprint arXiv:1312.3005 (2013)
18. Chen, S.F., Goodman, J.: An empirical study of smoothing techniques for language modeling. Computer Speech Language **13**(4), 359–394 (1999)
19. Jelinek, F.: Interpolated estimation of Markov source parameters from sparse data. In: Workshop on Pattern Recognition in Practice (1980)
20. Katz, S.: Estimation of probabilities from sparse data for the language model component of a speech recognizer. IEEE Transactions on Acoustics, Speech, and Signal Processing **35**(3), 400–401 (1987)
21. Bell, T., Witten, I.H., Cleary, J.G.: Modeling for text compression. ACM Computing Surveys **21**(4), 557–591 (1989)
22. Ney, H., Essen, U., Kneser, R.: On structuring probabilistic dependences in stochastic language modelling. Computer Speech Language **8**(1), 1–38 (1994)
23. Heafield, K.: KenLM: Faster and smaller language model queries. In: 6th Workshop on Statistical Machine Translation, pp. 187–197. Association for Computational Linguistics (2011)
24. Goodman, J.T.: A bit of progress in language modeling. Computer Speech Language **15**(4), 403–434 (2001)

25. Schwenk, H., Gauvain, J.L.: Training neural network language models on very large corpora. In: Conference on Human Language Technology and Empirical Methods in Natural Language Processing, pp. 201–208. Association for Computational Linguistics (2005)
26. Mikolov, T., Kombrink, S., Burget, L., Cernocký, J., Khudanpur, S.: Extensions of recurrent neural network language model. In: 2011 IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 5528–5531. IEEE Press, New York (2011)
27. Sundermeyer, M., Schlüter, R., Ney, H.: LSTM neural networks for language modeling. In: 13th Annual Conference of the International Speech Communication Association (2012)
28. Chen, S.F., Beeferman, D., Rosenfeld, R.: Evaluation metrics for language models. In: DARPA Broadcast News Transcription and Understanding Workshop, pp. 275–280 (1998)
29. Johnson, M., Schuster, M., Le, Q.V., Krikun, M., Wu, Y., Chen, Z., Thorat, N., Vi´egas, F., Wattenberg, M., Corrado, G., Hughes, M.: Google’s multilingual neural machine translation system: Enabling zero-shot translation. Transactions of the Association for Computational Linguistics **5**, 339–351 (2017)
30. Cotterell, R., Mielke, S.J., Eisner, J., Roark, B.: Are all languages equally hard to language-model? arXiv preprint arXiv:1806.03743 (2018)
31. Pavlyuk, N.: Contrastive Grammar of English and Ukrainian. DonNU, Donetsk (2010)