Towards Language Modeling for the Ukrainian Language*

Anastasiia Khaburska[1,2] and Igor Tytyk[3]

[1] Ukrainian Catholic University
[2] anastasykhab@gmail.com
[3] 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 is 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.

* Supported by Ciklum and Grammarly.

Table of Contents

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.

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

where S- sequences 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:

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

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 (Mikolov et al., 2010, Arisoy et al., 2012), machine translation (Schwenk, Rousseau, and Attik, 2012, Vaswani et al., 2017), sentiment analysis (Hu et al., 2007), text summarization (Rush, Chopra, and Weston, 2015, Filippova et al., 2015), grammatical error correction (Bryant and Briscoe, 2018), natural language generation (Edunov, Baevski, and Auli, 2019).

2 Motivation

Language Modeling is one of the central tasks to Natural Language Processing (NLP) 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. (2016): "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".

Also, to train and evaluate language models, it is required to have a well-composed corpus. In linguistics and natural language processing (NLP), corpus refers to a collection of texts. Such collections may be formed of a single language of texts or can 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.

Much work has been done for the English language in terms of developing both training and evaluation approaches. Count-based approaches (based on statistics of N-grams) typically add smoothing which accounts for unseen sequences. For example, Kneser-Ney smoothed 5-gram models (Kneser and Ney, 1995) traditionally were a fairly strong baseline. In recent years, much progress has been made by neural methods (Bengio et al., 2003; Mikolov et al., 2010), character-aware Neural Language Models (Kim et al., 2015), based on LSTMs (Jozefowicz et al., 2016), gated convolutional networks (N. Dauphin et al., 2017) and self-attentional networks (Al-Rfou et al., 2018).

At the same time, there hasn't 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) (Marcus, Marcinkiewicz, and Santorini, 2002), 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 (Chelba et al., 2013). 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 data-sets:

4.1.1 Ukrainian Brown Corpus [4] is a well balanced and redacted corpus of original Ukrainian texts published between 2010 and 2018 years, 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 and descriptive analysis of "Good" and "So-so" parts of this corpus. This consists of 924 texts, 600810 training words and 38728 unique lemmas (see Fig. 1).

Statistic by domain

Domain	Fraction	Description	length	RealFraction	NumSentences	NumTexts	UniqueLemmas
Преса	25%	Репортажі, огляди, редакційні статті, листи до редакції; національні й регіональні видання; тематично - політика, спорт, суспільство, економіка й фінанси, короткі новини, культура - театр, література, музика, танці	144276	0.240136	9333.0	363.0	18293.0
Релігійна література	3%	Книжки, періодика, брошури	19548	0.032536	1220.0	25.0	4710.0
Професійно-популярна література	7%	Книжки й періодика; домоводство, ремесла, «сад і город», хобі, ремонт і будівництво, конструювання, музика й танці, домашні тварини, спорт, їжа й вино, подорожі, фермерство, робочі професії тощо	32160	0.053528	2393.0	63.0	7356.0
«Естетичні інформативні» тексти	7%	Інформативні тексти, що не потрапляють в інші категорії, зокрема, біографії, мемуари, есеї, передмови, особисті листи, художня й мистецтвознавча критика, рекламні тексти	47491	0.079045	2820.0	57.0	10858.0
Адміністративні документи	3%	Закони, урядові акти, звіти організацій/фондів/компаній, офіційні листи	11964	0.019913	563.0	14.0	2272.0
Науково-популярна література	5%	Науково-популярні журнали, книги, медіа, статті	32545	0.054169	2137.0	44.0	7404.0
Наукова література	10%	Книжки й періодика; природничі й гуманітарні науки, техніка й інженерна справа	67690	0.112665	3114.0	82.0	9252.0
Навчальна література	15%	Підручники, посібники, гуманітарні та природничі науки тощо	81245	0.135226	4913.0	100.0	11010.0
Художні тексти	25%	Романи, повісті, оповідання, новели, за тематикою – загальна, детективи, фантастика, пригодницька, любовна, гумористична тощо	163891	0.272783	13527.0	176.0	21405.0

Fig. 1. Domain - category of the texts; Fraction - percentage stated by the Ukrainian brown corpus; Description - texts that fall within particular domain; length - sum of number of words for each text in the domain; RealFraction - length divided by the length of the whole data set; NumSentences - number of sentences in the domain; NumTexts - number of texts in the domain; UniqueLemmas - number of unique lemmas found by LemmatizeText.groovy (`https://github.com/brown-uk/nlp_uk`)

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

4.1.2 Uber-Text Corpus [5] More than 6 Gb amount of Ukrainian texts, but unfortunately, because of legal rules, split into sentences 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 from fiction writing and 631935 texts from Korrespondent news media data set. Of course, before using it, we should conduct some preprocessing.

4.1.3 Wiki dumps [6]

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:

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

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

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

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 $c(u_{i-n+1}^{i-1})$.

[5] Uber-Text Corpus: `http://lang.org.ua/en/corpora/`
[6] Ukrainian Wiki dumps: `https://dumps.wikimedia.org/ukwiki/20190920/`

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. Chen and Goodman (1998) carried out an extensive empirical comparison of the most widely used smoothing techniques, including those described by Jelinek and Mercer (1980), Katz (1987), Bell, Witten, and Cleary (1990), Ney, Essen, and Kneser (1994) and Kneser and Ney (1995). 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

$$p(u_i|u_1^{i-1}) = p(u_i|u_f^{i-1}) \prod_{n=1}^{f-1} c(u_n^{i-1}) \tag{5}$$

where the probability $p(u_i|u_f^{i1})$ and back-off penalties $b(u_n^{i1})$ are given by an already-estimated model.

Open-source KenLM library proposed by Heafield (2011) 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.

We trained[7] four n-gram language models using KenLM library on the Ukrainian Brown Corpus (length = 817699 units (words+punctuation marks)) and evaluated it with the *perplexity* measure (see Tab. 1).

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 con-

[7] Git-Hub : `https://github.com/Anastasiia-Khab/LMForTheUkrainianLanguage/blob/master/KenLM_Paragraph-base.ipynb`

n-gram model	perplexity
3-gram	16.565
4-gram	10.341
5-gram	8.004
6-gram	7.218

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

ditional independence assumptions are unrealistic. Using artificial neural networks in statistical language modeling has been proposed by Bengio et al. (2003), who used feed-forward neural networks with fixed-length context. This approach was exceptionally successful and further investigation by Goodman (2001). Later, Schwenk and Gauvain (2005) 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 Mikolov et al. (2010) 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 by (Chelba et al. (2013)).

In recent years, strong character-level language models (Mikolov et al., 2011; *LSTM Neural Networks for Language Modeling*) 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, Al-Rfou et al. (2018) 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 (Vaswani et al., 2017) 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 Methodology

5.1 Tokenization and lemmatization:

For tokenization and lemmatization, we use the nlp-uk library [8] from Andriy Rysin and the BrUk group.

5.2 Word embeddings:

For word embeddings we can use lang-uk embeddings [9] or fast-text embeddings [10] calculate embedding in parralel with training a model.

5.3 Evaluation:

As an evaluation metrics, firstly, we are going to consider perplexity (Chen, Beeferman, and Rosenfeld, 1998).

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 analyse the evaluation metric and the evaluation set

[8] LanguageTool API NLP UK : `https://github.com/brown-uk/nlp_uk`
[9] Lang-uk embeddings : `http://lang.org.ua/en/models/#anchor4`
[10] Fasttext embeddings : `https://fasttext.cc/docs/en/crawl-vectors.html`

- **19 October - 31 October**
  - ☐ Train a neural language model
  - ☐ Analyse the evaluation results and write conclusions
- **2 November - 14 November**
  - ☐ Experiment with pre-trained embeddings
  - ☐ Analyse 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
- **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 & 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 System (Johnson et al., 2016). 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 that 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 it 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 independent of the actual

languages involved. But this decision 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 Cotterell et al. (2018), most methods are portable in the following sense: Given appropriately annotated data, they should, 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. Cotterell et al. (2018) also conduct a study on 21 languages, demonstrating that in some languages, 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. Specifically, in their controlled comparison, language models perform worse on fine-grained inflectional languages. Furthermore, this performance difference essentially vanishes if to remove the inflectional markings.

Ukrainian is an East Slavic language and is famous for it's rich inflexions. It is noted by Pavliuk (2018), 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. For example: Петро - Петра - Петрові - Петром, він - йому - його - ним, всі - всіх - всім - всіма; червоний - червоного - червоному - червоним, двоє - двох - двом - двома; сонний – сонного - сонному – сонним; читав - читала - читали, читатиму - читатимеш - читатимете, etc. Whereas in English, which exposes an analytical structure, these word-classes are utterly devoid of any grammatical markers with the exception of a few pronouns. In Old English, the noun paradigm included 9 different inflectional forms, the paradigm of the weak verb had 10 forms, and the paradigm of adjectives - 13 synthetic (inflected) forms. Also, as pointed out by Prystai and Prystai (2017), 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 [11] 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. [12]

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.

Reference & Literature

Al-Rfou, Rami, Dokook Choe, Noah Constant, Mandy Guo, and Llion Jones (2018). "Character-Level Language Modeling with Deeper Self-Attention". In: *CoRR* abs/1808.04444. arXiv: 1808.04444. URL: http://arxiv.org/abs/1808.04444.

[11] Brown-Uk: https://r2u.org.ua/corpus
[12] Uber Text: http://lang.org.ua/en/corpora/

Arisoy, Ebru, Tara N. Sainath, Brian Kingsbury, and Bhuvana Ramabhadran (2012). "Deep Neural Network Language Models". In: *Proceedings of the NAACL-HLT 2012 Workshop: Will We Ever Really Replace the N-gram Model? On the Future of Language Modeling for HLT*. WLM '12. Montreal, Canada: Association for Computational Linguistics, pp. 20–28. URL: http://dl.acm.org/citation.cfm?id=2390940.2390943.

Bell, Timothy C., I. H. Witten, and John G. Cleary (1990). *Text compression / Timothy C. Bell, John G. Cleary, Ian H. Witten*. English. Prentice Hall Englewood Cliffs, N.J, xviii, 318 p. : ISBN: 0139119914.

Bengio, Yoshua, Réjean Ducharme, Pascal Vincent, and Christian Jauvin (2003). "A Neural Probabilistic Language Model". In: *Journal of Machine Learning Research* 3, pp. 1137–1155.

Bryant, Christopher and Ted Briscoe (2018). "Language Model Based Grammatical Error Correction without Annotated Training Data". In: *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*. New Orleans, Louisiana: Association for Computational Linguistics, pp. 247–253. DOI: 10.18653/v1/W18-0529. URL: https://www.aclweb.org/anthology/W18-0529.

Chelba, Ciprian, Tomas Mikolov, Mike Schuster, Qi Ge, Thorsten Brants, and Phillipp Koehn (2013). "One Billion Word Benchmark for Measuring Progress in Statistical Language Modeling". In: *CoRR* abs/1312.3005. arXiv: 1312.3005. URL: http://arxiv.org/abs/1312.3005.

Chen, Stanley, Douglas Beeferman, and Ronald Rosenfeld (1998). *Evaluation Metrics For Language Models*.

Chen, Stanley F. and Joshua Goodman (1998). *An Empirical Study of Smoothing Techniques for Language Modeling*. Tech. rep.

Cotterell, Ryan, Sebastian J. Mielke, Jason Eisner, and Brian Roark (2018). "Are All Languages Equally Hard to Language-Model?" In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, pp. 536–541. DOI: 10.18653/v1/N18-2085. URL: https://www.aclweb.org/anthology/N18-2085.

Edunov, Sergey, Alexei Baevski, and Michael Auli (2019). "Pre-trained Language Model Representations for Language Generation". In: *CoRR* abs/1903.09722. arXiv: 1903.09722. URL: http://arxiv.org/abs/1903.09722.

Filippova, Katja, Enrique Alfonseca, Carlos A. Colmenares, Lukasz Kaiser, and Oriol Vinyals (2015). "Sentence Compression by Deletion with

LSTMs". In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Lisbon, Portugal: Association for Computational Linguistics, pp. 360–368. DOI: 10.18653/v1/D15-1042. URL: https://www.aclweb.org/anthology/D15-1042.

Goodman, Joshua (2001). "A Bit of Progress in Language Modeling". In: *CoRR* cs.CL/0108005. URL: http://arxiv.org/abs/cs.CL/0108005.

Heafield, Kenneth (2011). "KenLM: Faster and Smaller Language Model Queries". In: *Proceedings of the EMNLP 2011 Sixth Workshop on Statistical Machine Translation*. Edinburgh, Scotland, United Kingdom, pp. 187–197. URL: https://kheafield.com/papers/avenue/kenlm.pdf.

Hu, Yi, Ruzhan Lu, Xuening Li, Yuquan Chen, and Jianyong Duan (2007). "A Language Modeling Approach to Sentiment Analysis". In: *International Conference on Computational Science*.

Jelinek, Fred and Robert L. Mercer (1980). "Interpolated estimation of Markov source parameters from sparse data". In: *Proceedings, Workshop on Pattern Recognition in Practice*. Ed. by Edzard S. Gelsema and Laveen N. Kanal. Amsterdam: North Holland, pp. 381–397.

Johnson, Melvin, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda B. Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean (2016). "Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation". In: *CoRR* abs/1611.04558. arXiv: 1611.04558. URL: http://arxiv.org/abs/1611.04558.

Jozefowicz, Rafal, Oriol Vinyals, Mike Schuster, Noam Shazeer, and Yonghui Wu (2016). "Exploring the limits of language modeling". In: *arXiv*, 1602.02410v2 [cs.CL]. URL: https://arxiv.org/abs/1602.02410.

Katz, Slava M. (1987). "Estimation of probabilities from sparse data for the language model component of a speech recognizer". In: *IEEE Trans. Acoustics, Speech, and Signal Processing* 35, pp. 400–401.

Kim, Yoon, Yacine Jernite, David Sontag, and Alexander M. Rush (2015). "Character-Aware Neural Language Models". In: *CoRR* abs/1508.06615. arXiv: 1508.06615. URL: http://arxiv.org/abs/1508.06615.

Kneser, Reinhard and Hermann Ney (1995). "Improved backing-off for M-gram language modeling". In: *1995 International Conference on Acoustics, Speech, and Signal Processing* 1, 181–184 vol.1.

Marcus, Mitchell, Mary Marcinkiewicz, and Beatrice Santorini (2002). "Building a Large Annotated Corpus of English: The Penn Treebank". In: *Computational Linguistics* 19, pp. 313–330.

Mikolov, Tomas, Stefan Kombrink, Lukás Burget, Jan ernocký, and Sanjeev Khudanpur (2011). "Extensions of recurrent neural network language model". In: *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5528–5531.

Mikolov, Tomáš, Martin Karafiát, Lukáš Burget, Jan Černocký, and Sanjeev Khudanpur (2010). "Recurrent neural network based language model". In: *INTERSPEECH* 2, pp. 1045–1048. URL: https://www.isca-speech.org/archive/interspeech_2010/i10_1045.html.

N. Dauphin, Yann, Angela Fan, Michael Auli, and David Grangier (2017). "Language Modeling with Gated Convolutional Networks". In: *arXiv*, 1612.08083v3 [cs.CL]. URL: https://arxiv.org/abs/1612.08083.

Ney, Hermann, Ute Essen, and Reinhard Kneser (1994). "On structuring probabilistic dependences in stochastic language modelling". In: *Computer Speech Language* 8, pp. 1–38.

Pavliuk, Nataliia (2018). *Contrastive Grammar of English and Ukrainian.*

Prystai, Galyna and Bogdan Prystai (2017). "Функціональні моделі творення демінутивів у англійській та українській мовах". In: *Молодий вчений*, pp. 212–215.

Rush, Alexander M., Sumit Chopra, and Jason Weston (2015). "A Neural Attention Model for Abstractive Sentence Summarization". In: *CoRR* abs/1509.00685. arXiv: 1509.00685. URL: http://arxiv.org/abs/1509.00685.

Schwenk, Holger and Jean-Luc Gauvain (2005). "Training Neural Network Language Models on Very Large Corpora". In: *HLT/EMNLP.*

Schwenk, Holger, Anthony Rousseau, and Mohammed Attik (2012). "Large, Pruned or Continuous Space Language Models on a GPU for Statistical Machine Translation". In: *Proceedings of the NAACL-HLT 2012 Workshop: Will We Ever Really Replace the N-gram Model? On the Future of Language Modeling for HLT*. Montréal, Canada: Association for Computational Linguistics, pp. 11–19. URL: https://www.aclweb.org/anthology/W12-2702.

Sundermeyer, Martin, Ralf Schlüter, and Hermann Ney. *LSTM Neural Networks for Language Modeling.*

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin (2017). "Attention is All you Need". In: *Advances in Neural Information Processing Systems 30*. Ed. by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. Curran Associates, Inc., pp. 5998–6008. URL: http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf.