Proceedings of the Federated Conference on Computer Science and Information Systems pp. 389–395 DOI: 10.15439/2017F432 ISSN 2300-5963 ACSIS, Vol. 11

Big Data Language Model of Contemporary Polish

Krzysztof Wołk
Polish-Japanese Academy of Information Technology,
ul. Koszykowa 86, 02-008
Warszawa, Poland
Email: kwolk@pja.edu.pl

Agnieszka Wołk
Polish-Japanese Academy of Information Technology,
ul. Koszykowa 86, 02-008
Warszawa, Poland
Email: awolk@pja.edu.pl

Krzysztof Marasek
Polish-Japanese Academy of Information Technology,
ul. Koszykowa 86, 02-008
Warszawa, Poland
Email: kmarasek@pja.edu.pl

Abstract - **Based on big data training we provide 5-gram language models of contemporary Polish which are based on the Common Crawl corpus (which is a compilation of more than 9,000,000,000 pages from across the web) and other resources. We prove that our model is better than the Google WEB1T n-gram counts and assures better quality in terms of perplexity and machine translation. The model includes lower-counting entries and also de-duplication in order to lessen boilerplate. We also provide POS tagged version of raw corpus and raw corpus itself. We also provide dictionary of contemporary Polish. By maintaining singletons, Kneser-Ney smoothing in SRILM toolkit was used in order to construct big data language models. In this research, it is detailed exactly how the corpus was obtained and pre-processed, with a prominence on issues which surface when working with information on this scale. We train the language model and finally present advances of BLEU score in MT and perplexity values, through the utilization of our model.**

I. Introduction

There are a large number of language processing tasks available that make web-scale corpora attractive and needed due in most, to the vast amount of information which exists in different languages. Language modelling is of great significance, where web-scale models for language have demonstrated their ability to enhance automated speech recognition performance and machine translation quality [1, 2, 3]. There are also other NLP tasks that depend greatly on language modelling e.g. language quantification. [4]

Contained within, are language models trained on the Common Crawl corpus and n-gram counts. Google has discharged n-gram counts which have been trained on 1,000,000,000,000 tokens of text [5]. N-grams which were present on fewer than forty occasions were pruned, and words which were present fewer than two hundred times were replaced with the unknown word. The counts are not suitable for judging a language model with the Kneser-Net smoothing algorithm due to this pruning as the algorithm needs unpruned counts, although pruning will happen on the last model anyway.

There is another challenge with the Google n-gram counts that are available publicly, [5] and this due to the fact that the training information was not de-duplicated, meaning that boilerplate, like copyright notices have got excessively high counts [6]. Despite Google sharing a version [7], in limited context [6], that has been de-duplicated, this was never officially released to the public [8]. Before adding up the n-grams, the training data was de-duplicated. There is a web service which is provided by Microsoft [9], you can query it for language model probabilities. However, this is limited to English language only, whereas our model preparation methodology is compatible with more languages outside of English. Additionally, there was an experiment conducted on the re-ranking of machine translated Polish, due to the number of queries from the output, the service crashed on several occasions, even with client-side caching. Utilization of the service from Microsoft, throughout machine translation decoding, would mean there is a requirement for a lower latency and there would be a greater volume of queries.

Summing up in our research we show how to build a contemporary language model from big data amounts of texts for any language supported in Common Crawl project (based on Polish). We compare its quality to Google WEB1T model and to set of freely available Polish corpora found in the web. We evaluate quality of our approach by measuring perplexity and showing higher quality of machine translation systems that use our model. Lastly, we share publicly results of our work as plain text data, trained 1-,2-,4- and 5-gram language model, RNN based language model and dictionary sorted by most frequent unigrams together with dictionary cleaned from numbers, names and less likely words. The data publicly available (https://goo.gl/hO1hTz).

II. Preparation of the Data

A crawl of the web which is in the available in the public domain is the CommonCrawl project. It contains petabytes of data collected over the last 7 years. It contains raw web page data, extracted metadata and text extractions.

The data is accessible as text only files as well as raw HTML. The text only files contain all the RSS and HTML files that the tags were stripped from. The text is converted to UTF-8 and the HTML is in the original encoding. There is a distinct benefit to be gained when using the HTML files because the structure of the document can be used to choose paragraphs, and can tell actual content from boilerplate. Parsing vast amounts of HTML needs a lot of normalization step and it is non-trivial.

Throughout this work, the focus is on dealing with the text-only files that were downloaded and processed on a small cluster locally. The benefits of structured text are unable to cancel out the additional computing power that is needed for the processing.

There were many problems that needed to be solved as pre-processing step. First of all, the selection of data only in a specific language. CommonCrawl also has some mistakes with encoding when parsing to UTF-8 which resulted with spelling errors. What is more, some texts are repeated many times e.g. copyright, comment, data, etc. Many text structures were ungrammatical or contained strange insertions. There were also some language specific difficulties that must have been addressed as well for each language separately. In addition, data contained both samples of spoken texts like dialogs or written articles and literature. The text domain also was not defined.

A. Differences between Polish and English languages

In general, Polish and English differ in syntax and grammar. English is a positional language, which means that the syntactic order (the order of words in a sentence) plays a very important role, particularly due to the limited inflection of words (e.g., lack of declension endings). Sometimes, the position of a word in a sentence is the only indicator of the sentence's meaning. In a Polish sentence, a thought can be expressed using several different word orderings, which is not possible in English. For example, the sentence "I bought myself a new car." can be written in Polish as "Kupiłem sobie nowy samochód.", or "Nowy samochód sobie kupiłem.", or "Sobie kupiłem nowy samochód.", or "Samochód nowy sobie kupiłem." The only exception is when the subject and the object are in the same clause and the context is the only indication which is the object and which is subject. For example, "Mysz liże kość. (A mouse is licking a bone.)" and "Kość liże mysz. (A bone is licking a mouse).".

Differences in potential sentence word order make the translation process more complex, especially when using a phrase-model with no additional lexical information [10]. In addition, in Polish it is not necessary to use the operator, because the Polish form of a verb always contains information about the subject of a sentence. For example, the sentence "On jutro jedzie na wakacje." is equivalent to the Polish "Jutro jedzie na wakacje." and would be translated as "He is going on vacation tomorrow.". [11]

In the Polish language, the plural formation is not made by adding the letter "s" as a suffix to a word, but rather each word has its own plural variant (e.g., "pies - psy", "artysta - artyści", etc.). Additionally, prefixes before nouns like "a", "an", "the", do not exist in Polish (e.g., "a cat - kot", "an apple - jabłko", etc.) [10].

The Polish language has only three tenses (present, past, and future). However, it must be noted that the only indication whether an action has ended is an aspect. For example, "Robiłem pranie." Would be translated as "I have been doing laundry", but "Zrobiłem pranie." as "I have done laundry", or "płakać - wypłakać" as "cry - cry out" [10].

The gender of a noun in English does not have any effect on the form of a verb, but it does in Polish. For example, "Zrobił to. – He has done it.", "Zrobiła to. – She has done it.", "lekarz/lekarka - doctor", "uczeń/uczennica = student", etc. [10]

Because of this complexity, progress in the development of SMT systems for West-Slavic languages has been substantially slower than for other languages. On the other hand, excellent translation systems have been developed for many popular languages.

B. Spoken vs written language

The differences between speech and text within the context of the literature should also be clarified. Chong [11] pointed out that writing and speech differ considerably in both function and style. Writing tends towards greater precision and detail, whilst speech is often punctuated with repetition and includes prosody, which writing does not possess, to further convey intent and tone beyond the meaning of the words themselves.

According to William Bright [12], spoken language consists of two basic units: Phonemes, units of sound, (that are themselves meaningless) are combined into morphemes, which are meaningful (e.g., the phonemes /b/, /i/, and /t/ form the word "bit"). Contrary alphabetic scripts work in similar way. In a different type of script, the basic unit corresponds to a spoken syllable. In logographic script (e.g., Chinese), each character corresponds to an entire morpheme, which is usually a word [12].

It is possible to convey the same messages in either speech or writing, but spoken language typically conveys more explicit information than writing. The spoken and written forms of a given language tend to correspond to one or more levels and may influence each other (e.g., "through" is spoken as "thru").

In addition, writing can be perceived as colder, or more impersonal, than speech. Spoken languages have dialects varying across geographical areas and social groups. Communication may be formal or casual. In literate societies, writing may be associated with a formal style and speech with a more casual style. Using speech requires simplification, as the average adult can read around 300 words per minute, but the same person would be able to follow only 150-200 spoken words in the same amount of time [13]. That is why speech is usually clearer and more constrained.

The punctuation and layout of written text do not have any spoken equivalent. But it must be noted that some forms of written language (e.g., instant messages or emails) are closer to spoken language. On the other hand, spoken language tends to be rich in repetition, incomplete sentences, corrections, and interruptions [14].

When using written texts, it is not possible to receive immediate feedback from the readers. Therefore, it is not possible to rely on context to clarify things. There is more need to explain things clearly and unambiguously than in speech, which is usually a dynamic interaction between two or more people. Context, situation, and shared knowledge play a major role in their communication. It allows us to leave information either unsaid or indirectly implied [14].

C. Main types of errors found in textual data

Another problem was that the data contained many errors. This data set had spelling errors that artificially increased the dictionary size and made the statistics unreliable. Some of them were casual errors and most of them were because of wrong text encoding conversion. We extracted randomly 10,000 segments of text from different (also) random parts of the CommonCrawl corpus. Then, a dictionary consisting of 92,135 unique words forms was created from TED 2013 (iwslt.org) data. The intersection of those two dictionaries resulted in information that that about 12% of the whole test set were spelling errors.

What was found to be more problematic was that there were sentences with odd nesting, such as:

Part A, Part A, Part B, Part B., e.g.:

“Ale będę starał się udowodnić, że mimo złożoności, Ale będę starał się udowodnić, że mimo złożoności, istnieją pewne rzeczy pomagające w zrozumieniu. Istnieją pewne rzeczy pomagające w zrozumieniu.”

Some parts (words, full phrases, or even entire sentences) were duplicated. Furthermore, there are segments containing repetitions of whole sentences inside one segment. For instance:

Sentence A. Sentence A., e.g.:

“Zakumulują się u tych najbardziej pijanych i skąpych. Zakumulują się u tych najbardziej pijanych i skąpych.”

or: Part A, Part B, Part B, Part C, e.g.:

”Matka może się ponownie rozmnażać, ale jak wysoką cenę płaci, przez akumulację toksyn w swoim organizmie - przez akumulację toksyn w swoim organizmie - śmierć pierwszego młodego.”

The analysis identified that 4% of test data contained such mistakes.

In addition, there were numerous untranslated English names, words, and phrases mixed into the Polish texts. There are also some words that originate from other languages (e.g., German and French).

D. Language Detection

The initial stage in the data acquisition pipeline is to separate the information by language. We looked at the option of detecting the main language automatically for each page, however, we discovered the mixed language occurs frequently within one page, and is relatively common. We implemented python tool that worked in 3 phases. Firstly, we used Python LangDetect [15] library to discover entire pages that seemed to be in Polish language. In the second phase, we used plWordnet [16] in order to compare vocabulary of extracted articles with Polish vocabulary. We removed articles that contained less than 30% of Polish words. What is more before using the plWordnet the aspell tool was used in order to correct spelling errors that could be corrected automatically. In the last step, we divided text into sentences using automatic tool implemented within [17] research. When data was divided into sentences each sentence was checked by calculating its probability in Google WEB1T language model. We removed 20% of less likely sentences. This assured removal of grammatically incorrect sentences or sentences in different languages while maintaining data that included additional Polish data not calculated in Google WEB1T.

By facilitating this technique, we were able to gather 278GB of clean textual data in UTF-8 encoding, that was sentence spited. The text contained 1,962,047,863 sentences in total.

E. Deduplication and normalization

Because the CommonCrawl consists of web pages there are many fragments which are not content, but are artefacts of auto-page generation, copyright notices are just one example, it is essential to remove such data because it would alter wrongly the statistical model. It must also be noted that some texts are repeated over the internet many time e.g. press information. To lessen the volume of boilerplate, before further processing, we took out any lines which were duplicated. For the purpose of deduplication, we implemented a python tool. The comparison was done at the level of sentences. The following Table I contains details about quantity of data before and after deduplication.

TABLE I.

DEDUPLICATION RESULTS

	Size in GB	**Number of sentences**	**Number of unique words**
Before	296,1	1,962,047,863	87,543,726
After	94,8	920,517,413	87,543,726

The step of de-duplication takes out around 75% of the Polish data. This is on par with the reductions reported by Bergsma et al. [18].

As well as de-duplicating the information, data was restricted to printable UTF-8 characters, we replaced all email addresses with the identical address, and removed the left-over HTML tags. Prior to the creation of the language models, punctuation was normalized utilizing the script which was supplied by the Workshop on Statistical Machine Translation [19], by using the Moses tokenizer [20] it was tokenized, and then the Moses true caser was applied.

III. Evaluation

In order to measure the performance of new language model we used the perplexity measure. Perplexity, developed for information theory, is a performance measurement for a language model. Specifically, it is the reciprocal of the average probability that the LM assigns to each word in a data set. Thus, when perplexity is minimized, the probability of the LM's prediction of an unknown test set is maximized. [21, 22, 23, 24] To be more precise, we chose 3 different test sets a corpus of TED lectures from IWSLT[1] conference, European Medicines Agency Leaflets (EMEA)[2] corpus and OpenSubtitles[3] corpus. From all 3 corpora, we randomly selected 1,000 sentences for the evaluation with perplexity. The details of used corpora are shown in Table II:

Table II.

Test Corpora Specification

	Number of sentences	Number of PL words	Number of EN words
TED	210,549	218,426	104,177
EMEA	1,046,764	148,230	109,361
OPEN	33,570,553	1,519,948	758,238

Secondly, using the same data sets, we trained 3 statistical machine translation models using Moses SMT toolkit. The translation took place from English to Polish. Translation systems were enriched with prepared language models and evaluated with BLEU metric.

BLEU was developed on a premise like that used for speech recognition, described in Papineni et al. [25] as: “The closer a machine translation is to a professional human translation, the better it is.” Hence, the BLEU metric is designed to measure how close SMT output is to that of human reference translations. It is important to note that translations, SMT or human, may differ significantly in word usage, word order, and phrase length [25]. To address these complexities, BLEU attempts to match phrases of variable length between SMT output and the reference translations. Weighted match averages are used to determine the translation score [26]. Several variations of the BLEU metric exist. The basic metric requires calculation of a brevity penalty PB as follows:

$$P_B = \begin{cases} 1, c > r \\ e^{(1-r/c)}, c \leq r \end{cases}$$

where r is the length of the reference corpus, and candidate (reference) translation length is given by c [27]. The basic BLEU metric is then determined as shown in [26]:

$$BLEU = P_B \exp(\sum_{n=0}^{N} w_n \log p_n)$$

where w_n are, positive weights summing to one, and the n-gram precision p_n is calculated using n-grams with a maximum length of N. There are several other important features of BLEU. Word and phrase positions in the text are not evaluated by this metric. To prevent SMT systems from artificially inflating their scores by overuse of words known with high confidence, each candidate word is constrained by the word count of the corresponding reference translation. The geometric mean of individual sentence scores, by considering the brevity penalty, is then calculated for the entire corpus [26].

The baseline results of SMT systems for each corpus are shown in Table III.

Table III.

Test Corpora Specification

Corpus Name	Baseline system score (BLEU)
TED	17,42
EMEA	36,74
OPEN	58,52

For language model training we used SRILM toolkit [28]. The fundamental challenge that language models handle is sparse data. It is possible that some possible translations were not present in the training data but occur in real life. There are some methods in SRILM, such as add-one smoothing, deleted estimation, and Good-Turing smoothing, that cope with this problem [23].

Interpolation and back-off are other methods of solving the sparse data problem in n-gram LMs. Interpolation is defined as a combination of various n-gram models with different orders. Back-off is responsible for choosing the highest-order n-gram model for predicted words from its history. It can also restore lower-order n-gram models that have shorter histories. There are many methods that determine the back-off costs and adapt n-gram models. The most popular method is known as Kneser-Ney smoothing. It analyses the diversity of predicted words and takes their histories into account [20]. We used this smoothing method and trained 5-gram language models.

For machine translation, we used the Experiment Management System [20] from the open source Moses SMT toolkit to conduct the experiments. Binarization of 5-gram language model was accomplished in our resulting systems using the KenLM Modeling Toolkit and language modelling itself, as mentioned, with SRILM [28] with an interpolated version of Kneser-Key discounting (interpolate – unk – kndiscount) that was used in our baseline systems. Word and phrase alignment was performed using SyMGIZA++ [29]

[1] iwslt.org
[2] opus.lingfil.uu.se
[3] opensubtitles.org

instead of standard The OOV's were handled by using Unsupervised Transliteration Model [30].

Summing up in this research we used big data CommonCrawl based corpus (COMMON), Google Corpus (WEB1T) and corpus gathered from available resources and crawled sources (OTHER). All but WEB1T that was already trained by Google in 5-gram order. The details about those corpora and number of ngrams are showed in following Table IV.

TABLE IV.

NUMBER OF N-GRAMS IN LANGUAGE MODELS

	COMMON	WEB1T	OTHER
1-grams	102,742,823	9,749,397	18,953,166
2-grams	1,227,434,111	72,096,704	248,705,481
3-grams	1,208,818,561	128,491,454	350,220,758
4-grams	1,513,980,357	128,789,635	468,203,863
5-grams	1,433,864,427	113,097,133	431,451,627

IV. EXPERIMENTS

The new data were:

- A Polish – English dictionary (bilingual parallel)
- Additional (newer) TED Talks data sets not included in the original train data (we crawled bilingual data and created a corpus from it) (bilingual parallel)
- E-books
- Subtitles for movies and TV series
- Parliament and senate proceedings
- Wikipedia Comparable Corpus (bilingual parallel)
- Euronews Comparable Corpus (bilingual parallel)
- Repository of PJIIT's diplomas
- Many PL monolingual data web crawled from main web portals like blogs, chip.pl, Focus news archive, interia.pl, wp.pl, onet.pl, money.pl, Usenet, Termedia, Wordpress web pages, Wprost news archive, Wyborcza news archive, Newsweek news archive, etc.

"Other" in the table below stands for many very small models merged together. EMEA are texts from the European Medicines Agency, KDE4 is a localization file of that GUI, ECB stands for European Central Bank corpus, OpenSubtitles [31] are movies and TV series subtitles, EUNEWS is a web crawl of the euronews.com web page and EUBOOKSHOP comes from bookshop.europa.eu. Lastly bilingual TEDDL is additional TED data.

TABLE V.

CRAWLED CORPORA SPECIFICATION

Data set	Dictionary	Sentences
EMEA	148,230	1,046,764
KDE4	131,477	185,282
ECB	62,147	73,198
OpenSubtitles	2,446,006	33,570,553
EBOOKS	1,283,060	17,256,305
EUNEWS	33.591	43,534
NEWS COMM	85,380	1,209,608
EUBOOKSHOP	599,405	593,818
UN TEXTS	606,989	5,312,280
DICTIONARY	92,121	n/a
OTHER	51,056	61,384
WIKIPEDIA	887,999	172,663
WEB PORTALS	4,797,497	26,578,683
BLOGS	1,645,106	2,735,568
USENET	1,583,413	3,768,719
DIPLOMAS	490.616	666,576
TEDDL	129,436	54,142

Data perplexity was examined by experiments with the TED lectures, OPEN and EMEA corpora. Perplexities for the test sets are shown in Table VI. The perplexity (PPL) values are with Kneser-Ney smoothing of the data.

TABLE VI.

PERPLEXITY-BASED LANGUAGE MODEL EVALUATION

CORPUS	MODEL	PERPLEXITY (PPL)
TED	Common Crawl	1471
TED	WEB1T	1523
TED	OTHER	1628
OPEN	Common Crawl	480
OPEN	WEB1T	671
OPEN	OTHER	823
EMEA	Common Crawl	1163
EMEA	WEB1T	1253
EMEA	OTHER	1417

The following Table VII provides results of our language model evaluation using SMT systems. We trained 3 baseline systems (Baseline BLEU) and then augmented them with our CommonCrawl-based language model (Augmented BLEU). The same was done using WEB1T and OTHER language models. The translation was conducted into Polish direction. The Delta column contains difference between baseline and augmented systems. It must be noted that we did not conduct any in-domain adaptation of language models.

TABLE VII.

SMT-BASED LANGUAGE MODEL EVALUATION

CORPUS	LANGUAGE MODEL	Baseline BLEU	Augumented BLEU	Delta
TED	Common Crawl	17.42	18.33	0.91
TED	WEB1T	17.42	17.97	0.55
TED	OTHER	17.42	17.76	0.34
OPEN	Common Crawl	58.52	59.23	0.71
OPEN	WEB1T	58.52	59.01	0.49
OPEN	OTHER	58.52	58.79	0.27
EMEA	Common Crawl	36.74	38.34	1.6
EMEA	WEB1T	36.74	37.93	1.19
EMEA	OTHER	36.74	37.26	0.52

V. RESULTS AND CONCLUSIONS

Summing up, we successfully released n-gram counts and language models built using big data textual corpora which overcomes limitations of other smaller, publicly available resources. In addition, we were able to show that after some basic pre-processing of the data we were able to obtain BLEU and perplexity results that outperform state-of-the-art language models like WEB1T and other smaller corpora even after merging them together. We proved that improvements in perplexity and also in machine translation lead to better language knowledge utilisation. The results of our work are free and publicly available . The resources we share are the raw data after pre-processing, raw data tagged with POS using Morfeusz tagger [32], trained 5-gram language model with pruned 20% of less likely n-grams, dictionary with count of most frequent words in Polish based on CommonCrawl corpus and lastly a similar dictionary without counts but manually cleaned from noisy data by native Polish translators.

VI. ACKNOWLEDGEMENTS

Work financed as part of the investment in the CLARIN-PL research infrastructure funded by the Polish Ministry of Science and Higher Education and was backed by the PJATK legal resources.

VII. REFERENCES

[1] Brants, T., Popat, A. C., Xu, P., Och, F. J., & Dean, J. (2007). Large language models in machine translation. In Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning.

[2] Guthrie, D., & Hepple, M. (2010, October). Storing the web in memory: Space efficient language models with constant time retrieval. In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (pp. 262-272). Association for Computational Linguistics.

[3] Chelba, C., & Schalkwyk, J. (2013). Empirical exploration of language modeling for the google. com query stream as applied to mobile voice search. In Mobile Speech and Advanced Natural Language Solutions (pp. 197-229). Springer New York, DOI: 10.1007/978-1-4614-6018-3_8

[4] Lenko-Szymanska, A., (2016). A corpus-based analysis of the development of phraseological competence in EFL learners using the CollGram profile. Paper presented at the 7 th Conference of the Formulaic Language Research Network (FLaRN), Vilnius, 28-30 June.

[5] Brants, T., & Franz, A. (2006). Web 1T 5-gram corpus version 1.1. Google Inc.

[6] Lin, D., Church, K., Ji, H., Sekine, S., Yarowsky, D., Bergsma, S., Patil, K., Pitler, E., Lathbury, R., Rao, V., Dalwani, K., & Narsale, S. (2010). Final report of the 2009 JHU CLSP workshop.

[7] Bergsma, S., Pitler, E., & Lin, D. (2010, July). Creating robust supervised classifiers via web-scale N-gram data. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (pp. 865-874). Association for Computational Linguistics.

[8] Lin, D., (2013). Personal communication, October

[9] Wang, K., Thrasher, C., Viegas, E., Li, X., & Hsu, B. J. P. (2010, June). An overview of Microsoft Web N-gram corpus and applications. In Proceedings of the NAACL HLT 2010 Demonstration Session (pp. 45-48). Association for Computational Linguistics.

[10] Swan, O. E. (2003). Polish Grammar in a Nutshell. University of Pittsburgh.

[11] Choong, C., & Power, M. S. The Difference between Written and Spoken English. Assignment Unit, 1.

[12] Daniels, P. T., & Bright, W. (1996). The world's writing systems. Oxford University Press on Demand.

[13] Coleman, J. (2014). A speech is not an essay. Harvard Business Review.

[14] Ager, S. (2013). Differences between writing and speech, Omniglot—the online encyclopedia of writing systems and languages.

[15] Language detection library ported from Google's language-detection. https://pypi.python.org/pypi/langdetect?

[16] Maziarz, M., Piasecki, M., & Szpakowicz, S. (2012). Approaching plWordNet 2.0. In Proceedings of 6th International Global Wordnet Conference, The Global WordNet Association (pp. 189-196).

[17] Wołk, K., Marasek, K. (2014) Polish – English Speech Statistical Machine Translation Systems for the IWSLT 2014, Proceedings of the 11th International Workshop on Spoken Language Translation, Tahoe Lake, USA, p. 143-149

[18] Bergsma, S., Pitler, E., & Lin, D. (2010, July). Creating robust supervised classifiers via web-scale N-gram data. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (pp. 865-874). Association for Computational Linguistics.

[19] Bojar, O., Buck, C., Callison-Burch, C., Federmann, C., Haddow, B., Koehn, P., Monz, C., Post, M., Soricut, R. & Specia, L. (2013). Findings of the 2013 Workshop on Statistical Machine Translation. In Proceedings of the Eighth

Workshop on Statistical Machine Translation, pages 1–44, Sofia, Bulgaria. Association for Computational Linguistics

[20] Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., ... & Dyer, C. (2007, June). Moses: Open source toolkit for statistical machine translation. In Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions (pp. 177-180). Association for Computational Linguistics.

[21] Chen, S. F., & Goodman, J. (1996, June). An empirical study of smoothing techniques for language modeling. In Proceedings of the 34th annual meeting on Association for Computational Linguistics (pp. 310-318). Association for Computational Linguistics, DOI: 10.3115/981863.981904

[22] Perplexity [Online]. Hidden Markov Model Toolkit website. Cambridge University Engineering Dept. Available: http://www1.icsi.berkeley.edu/Speech/docs/HTKBook3.2/node188_mn.html, retrieved on November 29, 2015.

[23] Koehn, P., (2010) Moses, statistical machine translation system, user manual and code guide.

[24] Jurafsky, D., [Online] Language modeling: Introduction to n-grams [Online]. Stanford University. Available: https://web.stanford.edu/class/cs124/lec/languagemodeling.pdf, retrieved on November 29, 2015.

[25] Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002, July). BLEU: a method for automatic evaluation of machine translation. In Proceedings of the 40th annual meeting on association for computational linguistics (pp. 311-318). Association for Computational Linguistics.

[26] Axelrod, A. (2006). Factored language models for statistical machine translation. DOI 10.1007/s10590-010-9082-5

[27] Graves, A., & Schmidhuber, J. (2005). Framewise phoneme classification with bidirectional LSTM and other neural network architectures. Neural Networks, 18(5), 602-610.

[28] Stolcke, A. (2002, September). SRILM-an extensible language modeling toolkit. In Interspeech (Vol. 2002, p. 2002).

[29] Junczys-Dowmunt, M., & Szał, A. (2012). Symgiza++: symmetrized word alignment models for statistical machine translation. In Security and Intelligent Information Systems (pp. 379-390). Springer Berlin Heidelberg, DOI: 10.1007/978-3-642-25261-7_30

[30] Durrani, N., Sajjad, H., Hoang, H., & Koehn, P. (2014, April). Integrating an Unsupervised Transliteration Model into Statistical Machine Translation. In EACL (Vol. 14, pp. 148-153), DOI: 10.3115/v1/E14-4029

[31] Wołk, K., & Marasek, K. (2014). Real-time statistical speech translation. In New Perspectives in Information Systems and Technologies, Volume 1 (pp. 107-113). Springer International Publishing, DOI: 10.1007/978-3-319-05951-8_11

[32] Morfeusz Tagger, Available: http://sgjp.pl/morfeusz/morfeusz.html, retrieved on March 23, 2017