Proceedings of the Federated Conference on
Computer Science and Information Systems pp. 193–197

DOI: 10.15439/2015F386
ACSIS, Vol. 5

Comparison Of Language Models Trained On Written Texts And Speech Transcripts In The Context Of Automatic Speech Recognition

Sebastian Dziadzio [1], Aleksandra Nabożny[1], Aleksander Smywiński-Pohl[1,2,3], Bartosz Ziółko[1,2]
[1] AGH University of Science and Technology,
Faculty of Computer Science, Electronics and Telecommunications, Krakow, Poland
[2] Techmo, Krakow, Poland, techmo.pl,
[3] Jagiellonian University, Department of Computational Linguistics, Krakow, Poland
dziadzio@student.agh.edu.pl, aleksander.pohl@uj.edu.pl, bziolko@agh.edu.pl

***Abstract*—We investigate whether language models used in automatic speech recognition (ASR) should be trained on speech transcripts rather than on written texts. By calculating log-likelihood statistic for part-of-speech (POS) n-grams, we show that there are significant differences between written texts and speech transcripts. We also test the performance of language models trained on speech transcripts and written texts in ASR and show that using the former results in greater word error reduction rates (WERR), even if the model is trained on much smaller corpora. For our experiments we used the manually labeled one million subcorpus of the National Corpus of Polish and an HTK acoustic model.**

***Index Terms*—automatic speech recognition, morphosyntactic language model, written and spoken language comparison**

I. Introduction

STATISTICAL language models (LM) are employed in various natural language processing applications, such as machine translation, information retrieval, ASR [21], or part-of-speech tagging [20]. Generally, they describe relations between words (or other tokens), thus enabling to choose most probable sequences. This proves to be especially useful in speech recognition, where acoustical models usually produce a number of hypotheses, and re-ranking them according to a language model can substantially improve recognition rates [20],[4],[6].

Despite extensive research into alternative techniques, n-gram models remain a technology of choice for most modern ASR systems. They are based on Markov assumption, which states that probability of a certain word is dependent only on its n-1 predecessors. It should be noted that efficiency of n-gram models is heavily language dependent. They correspond well to grammatical structure of positional languages (such as English), but in case of Polish and other highly inflected languages, words order is not a key indicator of relations between them [8]. The main difficulty in language modelling and learning problems in general is the curse of dimensionality. Higher-order models are usually more accurate, but with more dimensions the volume of space increases so fast that available data quickly become insufficient [2]. This problem is amplified in case of Polish due to complex inflectional rules resulting in a variety of word-forms.

Several techniques were proposed to account for long-span word dependencies and address the data sparsity problem. One of them are part-of-speech (POS) n-grams, which cluster words into categories based on grammatical classes [12], [14]. Such models are easy to build and allow the use of higher order n-grams, since there are far fewer grammatical categories than words. Furthermore, they can be trained on much smaller corpora, which is especially important for under-resourced languages. Written texts are usually easier to obtain than speech transcripts and consequently language models are commonly trained on the former [5] [18].

II. Motivation

There has been a lot of studies in the humanities and social sciences dealing with the comparison of speech and text. It is known that there are fundamental dissimilarities between oral and written language in terms of grammatical structures, sentence lengths, choice of words etc. [3]. Whether those differences can be captured by means of statistical analysis, remains an open question.

The main motivation behind our study was to investigate whether LM based on written texts are an appropriate source of information about spoken language for automatic speech recognition. We conducted a comparative analysis of two corpora. One of them consisted of speech transcripts, while the other contained only written texts. We were looking for general features allowing to distinguish between two channels of communication (speech vs. text) rather than stylistic differences resulting from distinct language domains. That is why traditional methods of corpus comparison based on word frequencies were not applicable [15]. We therefore decided to compare POS n-grams in order to find grammatical patterns typical of either spoken or written language. Our initial hypothesis holds that there are statistically significant differences between those two n-gram sets. If this assumption is correct, it would imply that training LM solely on speech transcripts could lead to greater WERR in ASR systems.

III. Related Work

The idea of comparing speech and text corpora in terms of POS tags was motivated by previous research concerning the use of morphosyntactic n-grams in speech recognition of Polish. Until recently, there was little interest in using POS tags in ASR. In [22] a POS tagger was tested as a possible improvement in speech recognition of Polish. The results were negative, because the tagger frequently produced ambiguous output. This issue was later addressed in [11] by reducing model specificity (only grammatical classes were taken into account). It was concluded that simplified POS tags can be very useful for building statistical models of Polish.

In [12] an optimal set of grammatical categories was experimentally derived. Thirteen trigram language models were built, each employing both grammatical classes and one selected grammatical category. Then they were compared to a model based only on grammatical classes (hereinafter called POS-only model) in terms of WERR. Only three categories (gender, number, and case) offered significant improvements over the POS-only model. Surprisingly, combining those categories resulted in a model performing insignificantly better than the POS-only model. For this reason, our research is mostly based on the POS-only model, although we also take into account three aforementioned categories.

IV. Data Preparation

The National Corpus of Polish (NKJP) is divided into two parts: manually annotated 1-million corpus (1MC) and automatically annotated 1-billion corpus (1BC). Texts are labeled on several lavels: word and sentence boundaries, morphosyntactic tags, named entities, and syntactic groups. Annotation in 1MC is conducted very strictly, as each element was labeled by two independent researchers and then corrected by a super-annotator in case of a tie. The corpus includes diverse materials: classic literature, daily newspapers, scientific journals, and a variety of short-lived and Internet texts. Most importantly, it also includes speech transcripts from parliament proceedings, real-life conversations, radio, and television [13]. The proportion of speech transcripts to text data in 1MC is 109 919 (speech) vs. 1 091 981 (text) tokens.

Each segment in NKJP belongs to one of 35 grammatical classes. They are far more detailed than traditional parts of speech (for example there are 14 distinctive verb classes and 4 adjective classes). Obtaining information about grammatical classes was straightforward and required parsing XML label files. Unfortunately each paragraph is described by several label files stored in a separate directory, so they had to be processed individually. Although rather inconvenient, this design prompted us to take advantage of parallel processing, which will later be useful in case of 1-billion corpus.

Extracting grammatical categories was a more demanding task, mainly because category tags take a form of a single, colon-delimited string. For example, the word objęcia has a following tagging: ger:sg:gen:n:perf:aff. The first element is the grammatical class (POS) tag, followed by a set of grammatical category tags. This notation is further complicated by the fact that each grammatical class has a different set of categories. For example, adjectives have gender, number, case, and degree, while verbs are described by their number, person, and aspect. As it has already been said, only gender, number, and case were taken into account, as they play primary role in agreement relation.

It should be noted that we ignored all non-lexical backchannels and other noise in the transcripts. We also discarded all utterances containing incomprehensible words, as we wanted to focus on grammatical properties of the spoken language.

V. Statistical Comparison

Selecting appropriate statistical tools was yet another challenge. We considered three methods: the Spearman's coefficient, $\chi 2$-test and log-likelihood statistic. We concluded that the first method is not applicable to POS n-grams because of its tendency to overestimate differences for rare units. We also rejected the $\chi 2$-test because its null hypothesis is that compared corpora comprise words drawn randomly from a larger population. Since words in texts are obviously not random, the null hypothesis is defeated for almost all common words [9]. It is especially problematic for POS n-grams, where there are typically several very common units (which can be expected to give high $\chi 2$ values) and a lot of rare units (for which the $\chi 2$ test is not applicable). We decided to use the third method, as it is applicable to corpora of different sizes and has been reported to work well with POS n-grams [15]. Given the frequency lists, we build a contingency table for each POS n-gram:

Table I.
Example contingency table.

	Corpus A	**Corpus B**
Count of unit:	n_A	n_B
Count of other units:	$N_A - n_A$	$N_B - n_B$
Total:	N_A	N_B

Values n_A and n_B are called observed values (O). We then calculate expected values (E) according to the formula:

$$E_i = \frac{N_i \sum_i O_i}{\sum_i N_i} \tag{10}$$

Using the data from Table 1, we obtain $E_A = \frac{N_A(n_A+n_B)}{N_A+N_B}$ and $E_B = \frac{N_B(n_A+n_B)}{N_A+N_B}$. The log-likelihood value is then calculated according to the following formula:

$$2\sum_i O_i \ln\left(\frac{O_i}{E_i}\right) \tag{2}$$

In our case this equals:

$$2n_A \ln\left(\frac{n_A}{E_A}\right)+2n_B \ln\left(\frac{n_B}{E_B}\right) \quad (3)$$

The higher this value, the more significant is the difference between two frequency scores. LL of 3.8 or higher is significant at the 5% level. For the purpose of comparison, we used five corpora of written texts and five corpora of speech transcripts (full corpus, two half-corpora and two smaller samples). We then performed a round robin comparison: for each pair of corpora we calculated the number of units for which the LL value was greater than 3.8. Averaged results are presented below. S-S and T-T denote intra-corpus comparisons (speech and text, respectively). S-T denotes a comparison between speech and text corpora.

TABLE II.
AVERAGE NUMBER OF N-GRAMS WITH DIFFERENCES IN FREQUENCY SIGNIFICANT AT 5% LEVEL. VALUES IN BRACKETS ARE STANDARD DEVIATIONS.

n	S-T	S-S	T-T
1	30.3 (2.0)	14.1 (5.1)	17.2 (3.6)
2	418.8 (42.6)	127.2 (28.2)	182.5 (49.0)
3	2281.4 (482.7)	1205.1 (215.8)	1628.4 (274.3)

The log-likelihood analysis reveals large differences in frequencies of POS-tags. The LL scores were significant at 5% level for more than 30 unigrams (out of 35). This number is much lower in case of intra-corpus comparisons. The same holds true for higher-order n-grams (bigrams and trigrams). As stated before, we used five corpora for speech and text (resulting in 10 intra-corpus comparisons and 25 inter-corpus comparisons), so observed differences are not an effect of differing corpus sizes. Qualitative analysis of POS tags with highest LL score could reveal usage patterns characteristic for written and spoken language.

Another test involved calculating the percentage of common n-grams in the set of k most popular units:

$$\frac{\left|K_1 \cap K_2\right|}{\left|K_1 \cup K_2\right|} \cdot 100 \quad (4)$$

In the above formula, K1 and K2 denote sets of k most popular n-grams in compared corpora. We considered unigrams, bigrams, and trigrams. We decided to set k in relation to the total number of units (5%, 10%, and 20% of all units). Table 3 presents calculated values. "S-T" denotes a comparison of full speech corpus vs. full text corpus. "S-S" and "T-T" denote a comparison between two halves of the same corpora (the split was made by randomly assigning each paragraph into one of two subcorpora).

The test reveals significant differences in POS n-gram distributions. The values in the first column (speech vs. text) are not only lower, but also decreasing with the model complexity. The values in the second and third column (speech vs. speech and text vs. text) are much higher and stay the same as the order of n-grams increases. This shows that grammatical patterns typical for spoken or written language can be captured with morphosyntactic n-gram models.

TABLE III.
PERCENTAGES OF COMMON UNITS AMONG K MOST POPULAR N-GRAMS.

Unigrams

k	S-T	S-S	T-T
2	100.0	100.0	100.0
5	100.0	100.0	100.0
10	90.0	100.0	100.0

Unigrams with categories

k	S-T	S-S	T-T
20	85.0	100.0	100.0
40	87.5	95.0	98.0
80	85.0	97.5	98.8

Bigrams

k	S-T	S-S	T-T
35	78.6	94.3	100.0
70	77.1	95.7	100.0
140	74.3	93.6	98.6

Bigrams with categories

k	S-T	S-S	T-T
400	70.6	88.8	97.8
800	72.8	87.4	95.8
1600	70.5	85.9	94.6

Trigrams

k	S-T	S-S	T-T
250	64.6	89.2	97.2
500	63.9	90.2	96.4
1000	64.6	89.2	95.8

VI. PERFORMANCE IN ASR

The results of statistical analysis indicated that language models trained on speech transcripts or written texts would have different properties and therefore give different results when applied to ASR. In order to test this hypothesis, we have built several language models and employed them in rescoring of the hypotheses produced by HTK (without any LM or grammar) for several hundred Polish sentences. For tagging we used Concraft-pl, a conditional random field tagger for Polish which had proved to be particularly effective in ASR applications [17],[12]. The rescoring was done by

combining the probabilities of the acoustic and morphosyntactic model

$$P(h_i) = P(h_i)_{LM}^{\alpha} \cdot P(h_i)_{AM}^{1-\alpha} \quad (5)$$

where

$P(h_i)$ – the probability of the i-th hypothesis,

$P(h_i)_{LM}$ – the probability of the i-th hypothesis according to the language model,

$P(h_i)_{AM}$ – the probability of the i-th hypothesis according to the acoustic model,

α – the weight of the LM component.

The models were tested on several audio corpora. The first one (K1) includes 107 sentences spoken by one male voice, without any added noise, but recorded in an office with working computers. It consists of political speeches and spoken fragments of political song lyrics. The second corpus (K2) includes 23 samples spoken by a young female professional speaker. The third corpus (K3) consists of 221 short utterances recorded during various tests of speech/speaker recognition systems at AGH University of Science and Technology with addition of recordings from meetings of the Department Council. This corpus includes many various voices (one speaker says no more than six sentences) and recording devices, often with a natural random noise due to bad acoustic conditions (reverberation, voices in the background, traffic from outside etc.) We also used some recordings from LUNA, a corpus of telephone conversations from a call center of Warsaw public transport [10]. 192 samples of various female voices (K4) and 226 of male voices (K5) were used. These are informal utterances with many questions. The corpus is full of grammar mistakes, very common in natural conversations. The last test corpus (K6) consists of 86 recordings randomly chosen from Polish Global Phone corpus [16]. It is a corpus of speech dictated from an everyday journal.

The union of the corpora was divided into two subsets: a tuning set containing 15% randomly chosen sentences, used to estimate the alpha parameter, and a testing set, containing the remaining sentences. The text and the speech corpora were used to build two language models (LMs): one containing only POS tags (POS-only) and the other containing POS tags together with gender, number and case tags (POS-gnc). In each case a trigram model was built, smoothed using Witten-Bell method [19], due to their small size.

The comparison of speech and text based LMs was conducted by measuring the Word Error Rate Reduction (WERR) obtained with a given model. The results of the test are given in Table 4. LMs with Speech prefix are based on the Speech sub-corpus of 1MC, with Text prefix – on the Text sub-corpus, and with Text-sample, on a text sub-corpus of the same size as the Speech sub-corpus. The best result is obtained for the LM based on the speech corpus using POS, gender, number and case tags. The difference between the best result and the second result (Text-sample-POS-gnc) is statistically significant (paired Student's t-test, n=724, $P < 0.028$). Interestingly, although the Speech-POS-only LM performs better than the Text-POS-only LM, the difference is not statistically significant.

TABLE 4.
PERFORMANCE OF DIFFERENT LMS IN ASR.

LM	WERR [percentage points]
Speech-POS-gnc	29.5
Text-sample-POS-gnc	28.0
Text-POS-gnc	27.8
Speech-POS-only	27.1
Text-POS-only	26.5
Text-sample-POS-only	25.9

VII. CONCLUSIONS

Building language models based on POS n-grams is a promising technique in ASR of highly inflected languages. Benefits include simple structure, substantial dimensionality reductions, and noticeable improvements in performance of ASR systems [12]. Our analysis shows that it is possible to discriminate between speech and text data using only POS n-grams. It implies that morphosyntactic models trained on written texts do not accurately reflect the grammatical structure of spoken language. This hypothesis was confirmed by the ASR experiments. The Speech-POS-gnc model outperformed all text-based models, even those trained on ten times more data. The experiment also show that grammatical categories (gender, number, and case) carry important information about the structure of inflectional languages. Including them improved recognition rates in all cases.

VIII. ACKNOWLEDGEMENTS

This work was supported by LIDER/37/69/L-3/11/NCBR/2012 and DOB-BIO6/22/133/2014 grants from the Polish National Center for Research and Development.

Narodowe Centrum Badań i Rozwoju

REFERENCES

[1] Bardoel, T. "Comparing n–gram frequency distributions". Tilburg University School of Humanities. Tilburg center for Cognition and Communication. 2012.

[2] Bengio, Yoshua, Ducharme, Réjean, Vincent, Pascal, Jauvin, Christian. "A neural probabilistic language model". Journal of Machine Learning Research. vol. 3. pp. 1137-1155. 2003.

[3] Biber, Douglas. "Variation across speech and writing". Cambridge University Press. 1991.

[4] Chelba Ciprian, Bikel Dan, Shugrina Maria, Nguyen Patrick, Kumar Shankar. "Large scale language modelling in automatic speech recognition.". Google Research. 2012.

[5] Hirsimaki, T., Pylkkonen, J., Kurimo, M., "Importance of high-order n-gram models in morph-based speech recognition". IEEE Trans.

Speech and Language Processing. 17(4):724-32. 2009. http://dx.doi.org/10.1109/TASL.2008.2012323

[6] Janicki, A., Wawer, D., "Automatic Speech Recognition of Polish in a Computer Game Interface", Proceedings of the Federated Conference on Computer Science and Information System 2011, pp. 711–716. 2011.

[7] Jurafsky, D., Martin, J. H. "Speech and language processing. 2nd edition". Prentice-Hall. Inc. New Jersey. 2008.

[8] Karpov, A., Ronzhin, A., Markov, K., Kipyatkova, I., Vazhenina, D. "Large vocabulary Russian speech recognition using syntactico-statistical language modelling". Speech Communication 56 (2014) 213-228. 2014. http://dx.doi.org/ 10.1016/j.specom.2013.07.004

[9] Kilgarriff, Adam. "Comparing Corpora". International Journal of Corpus Linguistics. 6:1. 97-133. 2001.

[10] Marciniak, M. "Anotowany korpus dialogów telefonicznych.". Akademicka Oficyna wydawnicza EXIT. 2011.

[11] Pohl, A., Ziółko, B. "Using part of speech n-grams for improving automatic speech recognition of Polish". 9th International Conference on Machine Learning and Data Mining MLDM. 2013. http://dx.doi.org/10.1007/978-3-642-39712-7_38