Proceedings of the 10th IASTED International Conference
Artifical Intelligence and Applications (AIA 2010)
February 15 - 17, 2010 Innsbruck, Austria

WORD N-GRAMS FOR POLISH

Bartosz Ziółko, Dawid Skurzok, Mariusz Ziółko
Department of Electronics
AGH University of Science and Technology
Al. Mickiewicza 30, 30-059 Kraków, Poland
{bziolko,ziolko}@agh.edu.pl

ABSTRACT
The large collection of word n-gram statistics for Polish is described. Some details of the text analysis algorithm supporting processing data on computer clusters is presented as well. The corpora of total size of 267 030 267 words were used. The encountered problems due to the special Polish characters are described as well as the impact of rich morphology in Polish on this type of statistics. The most common n-grams are presented and commented. This is the first publication of such statistics of Polish.

KEY WORDS
Polish, n-grams, speech recognition, language modelling

1 Introduction

Usually language modelling is based on stochastic process approaches. Let us assume the existence of a probabilistic space which consist of sequences of random variables. E is the space of process states, and T stands for the domain of a stochastic process, which is defined as a set $S(t) = \{A(t), t \in T\}$ of random variables $S(t)$, where T is a set of time indexes. A sequence of spoken words can be treated as a realisation of a stochastic process.

The language properties have been very often modelled by *n*-grams [3], [12], [7], [4], [5], [10], [2]. Let us assume the word string $w \in W$ consisting of n words $w_1, w_2, w_3, ..., w_n$. Let $P(W)$ be a set of probability distributions over word strings W that reflect how often $w \in W$ occurs. It can be decomposed as

$$P(w) = P(w_1)P(w_2|w_1)P(w_3|w_1, w_2)...P(w_n|w_1, ..w_{n-1}). \quad (1)$$

It is theoreticaly justified and practically useful assumption that, $P(w)$ dependence is limited to n words backwards. Probably the most popular are trigram models where $P(w_i|w_{i-2}, w_{i-1})$, as a dependence on the previous two words is the most important, while model complication is not very high. Such models need statistics collected over a vast amount of text. As a result many dependencies can be averaged.

N-grams are very popular in automatic speech recognition (ASR) systems [12], [6], [10], [2]. They have been found as the most effective models for several languages. Our attempt was to build such models for Polish language. N-grams calculated by us will be used for the major layer of language model of a large vocabulary ASR system. The large number of analysed texts will allow us to predict words being recognised and improve recognition highly in this way.

Polish is highly inflective in contrast to English. The rich morphology causes difficulties in training language models due to data sparsity. Much more text data must be used for inflective languages than for positional ones to achieve the model of the same efficiency [10]. This is why our algorithm was considered from time-efficiency point of view and computer clusters were used for calculations.

N-gram statistics should improve the ASR systems considerably. The standard ASR scheme is as follows. The acoustic signal is framed into 23 ms long segments, which are represented by mel frequency cepstral coefficients (MFCC) [1]. The hypotheses of words are found applying hidden Markov model (HMM) [8] on the MFCCs. Then the word hypotheses are analysed using *n*-grams. There are no records of applying the whole scheme on Polish because of the lack of the last element, *n*-gram models. Our research are conducted to make it possible.

There are several existing tools for creating n-grams like SRLIM, IRSTLM or RandLM [9]. We did not use them for two reasons. First, Polish has several characters which exist only in Polish. There are a few different standards to code them. It happens that in some texts there are different coding standards used in different parts of texts. The existing solutions were designed mainly for English and it would be difficult to adapt them to Polish for the mentioned reason. Another issue is that the system designed by us is dedicated to be used by Police and other homeland security forces. For these reasons the deliverables of our project can be declared as classified so we cannot use solutions with code freely available on Internet.

2 Collecting *n*-grams on a cluster from large corpora

Around 9 gigabytes of data was analysed. However, we did not combine all statistics yet, so the presented results are for part of the corpora only, which are above 2 GB. Text data for Polish are still being collected and will be included in the future works. The system is designed and tested to be able to process unlimited source of texts.

The calculations were conducted on cluster Mars in

Cyfronet Centre, Krakow. It is a cluster for calculations with following specification: IBM Blade Center HS21 - 112 Intel Dual-core processors, 8GB RAM/core, 5 TB disk storage and 1192 Gflops. It operates under Red Hat Linux.

N-gram collection was build on SQLite database. A file or part of file is loaded to 1 MB buffer. Words to build *n*-grams are taken one after another from the buffer. While loading words from the buffer, every character is checked if it is an ASCII or UTF-8 letter. If it is an end of word character (like space, tabulator, end of line, or dot), the length of the word is stored for later use. Text is arranged in sequence of three words, then saved as 1, 2 and 3-grams. Only the first word is saved as 1-gram, the first and the second as 2-gram and the whole sequence is saved as 3-gram. After that, the second word is moved at the begining and the third one is moved to the second position. The new third word is loaded from a buffer, and again the sequence is saved. 2 and 3-grams are stored as one string with each word seperated by a space.

The access to files on Mars is the bottleneck of time efficiency. Everytime while adding or updating some *n*-gram, a database engine reads and writes some data from a file or cache. To improve speed of input/output operations, we have increased cache size, compiled SQL statments before using them and disabled synchronisation between data in memory and on hard drive. This is why every word is read only once from a corpus. The process of checking if the new word already exists in the statistics is conducted on the database cache rather than on the file for the same reason.

It was checked that applying multithread algorithm did not improve efficiency. The problems of the access to files and the necessity of using calculation power on controlling threads are possible reasons. Instead of multithreading, big corpora were split to smaller parts and processed separately, at the same time. Then results were joined together.

We faced another problem which was caused by format of special Polish characters like *ó, ł, ę, ż*. The same letter is kept in different formats using different bytes. Eventhough Gżegżółka software was used to change text files from one standard into another and to unify them into UTF-8, some parts of files contained unexpected values which looked like they belong to a different standard. It is one of the reasons why we did not use any of the ready solutions for English. The percentage of 1-grams with the unrecognised symbols is from 6.5% in the literature corpus to 0.4% in the transcript corpus, which was not yet included in the general statistics due to memory problems on the cluster described later. All punctuations were removed and all symbols which were not letters of the Polish alphabet were replaced with asteriks using stream editor SED. STL library was replaced by our own function to manage strings in aim of improving speed of basic string operations.

The algorithm was created in a way that lenghts of words are calculated once. Every symbol is checked once if it is a recognised ASCII letter or a special UTF-8 symbol.

3 Corpora

Newspaper articles in Polish were used as our first corpus. They are Rzeczpospolita newspaper articles taken from years 1993-2002. They cover mainly political and economic issues, so they contain quite many proper names. In total, 879 megabytes of text (103 655 666 words) were included in the process.

Several millions of Wikipedia articles in Polish made another corpus. They are of encyclopedia type, so they also contain many names including foreign ones. In total, it has 754 megabytes of text (96 679 304 words). Table 1 clearly shows that Wikipedia has much more types of words, incuding basic forms, than other corpora, eventhough they are of similar size.

The third corpus consists of several literature books in Polish from different centuries. Some of them are translations from other languages, so they also contain some foreign words. The corpus includes 490 megabytes (68 144 446 words) of text.

We collected a few more corpora, being around 9 GB in total, however, they were not combained yet. The results are presented for the corpora described above. The number of *n*-grams in Table 2 for the combination of all corpora is not a sum of *n*-grams for particular corpora because many words are repeated between different corpora.

4 Results and discussion

Table 1 summarises the corpora we used to calculate the statistics. Polish is a highly inflective language what can be seen by comparing the number of 1-grams and the number of basic morphological forms. In average, there are less than two times fewer basic forms than *n*-grams. To calculate this, all corpora were analysed using morphological analyser for Polish - Morfeusz [11]. It has to be mentioned here that around 40 % of 1-grams appeared just once for most of the corpora and 54 % for Wikipedia (see Table 3). This is why the ratio of a number of 1-grams and basic forms is quite low. It should change after adding more text data into the statistics. The higher percentage of words which appeared only once in Wikipedia is probably connected to its encyclopedia nature. It contains many names and rare words. We plan to add more data as long as the percentage of single appearences will not drop. Then the rare words can be removed from the statistics.

The most popular 1-grams in Polish are often pronouns, what is not surprising. They are presented in Table 4, where their English translations were provided. However, it is difficult to translate pronouns without a context. This is why there are often several translations. One of the commonly used words is *się*. It is a reflexive pronoun. It could be translated as *oneself*, but it is much more common in Polish than in English. It is used always, if a subject activity is conducted on herself or himself.

Some English words appeared in the statistics as well as single letters. These were removed as errors. Such words

Table 1. Analysed text corpora with their sizes, perplexity.

Corpus	MBytes	Mwords	Basic forms	Perplexity
Rzeczpospolita journal	879	104	832 732	8 918
Wikipedia	754	97	2 084 524	16 436
Literature	490	68	610 174	9 031
Combination of all	2123	267		9 199

Table 2. The number of different *n*-grams in the analysed corpora.

Corpus	1-grams	2-grams	3-grams
Rzeczpospolita journal	1 275 475	26 390 703	62 440 894
Wikipedia	2 623 358	31 139 080	61 865 543
Literature	1 151 043	23 830 490	50 794 854
Combination of all	3 161 748	59 590 565	143 502 429

Table 3. Analysed text corpora with their sizes, perplexity and number of n-grams.

Corpus	single 1-grams	%	1-grams with errors	%
Rzeczpospolita journal	560 549	44	26 786	2
Wikipedia	1 426 958	54	108 338	4
Literature	467 376	41	75 204	6.5
Combination of all	1 645 474	52	187 382	5.9

could be effects of including some English citations in articles or not removed tags. All files were preprocessed using SED as was described earlier, eventhough some tags were not removed for unknown reasons. They were especially strongly represented in 3-grams. Some *n*-grams are also results of sentences repeated in all articles like *external links*.

Collected statistics show that the amount of text we used was enough to create representative statistics for 1-grams and 2-grams but not for 3-grams. The most common triples are very strongly connected to the corpora, especially Wikipedia, where all towns and villages are described with the same pattern. These and some other patterns constitute most of the top of 3-gram statistics. We plan to remove all small Wikipedia sites from our corpus.

5 Conclusion

It is more difficult to construct practically useful *n*-grams for Polish than for English. There are some problems with existence of several standards of special Polish characters. Eventhough there is software to change all texts into one format (UTF-8), some unexpected characters may remain. Polish is morphologically rich what enlarges the number of possible words in *n*-grams. This is why more text data is necessary than in English to create *n*-grams which could be applied in ASR. Unfortunetlly much less linguistic data is available for Polish than English, so the sources of data for training are limited.

Table 4. The most popular 1-grams in the analysed Polish language corpora (r.p. - reflexive pronoun) presented with the number of times they occured in the analysed corpora and a percentage regarding to the whole text.

Polish	English	occureances	%
.	.	17 245 057	6.066
w	in	8 937 331	3.143
i	and	5 117 887	1.800
na	on, at	4 261 092	1.499
z	with	3 951 484	1.390
się	r.p.	3 904 232	1.373
do	to, till	2 873 022	1.010
nie	no, not	2 863 107	1.007
to	it, this	1 768 183	0.622
że	that	1 656 240	0.583
jest	is	1 489 305	0.524
o	about, at	1 412 878	0.497
a	and	1 386 398	0.488
1	1	1 076 986	0.379
od	from, since	1 019 478	0.359
po	after	946 097	0.333
przez	through	883 156	0.311
2	2	882 907	0.310
0	0	865 825	0.304

Acknowledgement

This work was supported by MNISW grant OR00001905.

References

[1] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. ASSP-28, no. 4, pp. 357–366, 1980.

[2] T. Hirsimaki, J. Pylkkonen, and M. Kurimo, "Importance of high-order n-gram models in morph-based speech recognition," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 17(4), pp. 724–32, 2009.

[3] W. Huang and R. Lippman, "Neural net and traditional classifiers," *Neural Information Processing Systems, D. Anderson, ed.*, pp. 387–396, 1988.

[4] D. Jurafsky and J. H. Martin, *Speech and Language Processing*. New Jersey: Prentice-Hall, Inc., 2000.

[5] S. Khudanpur and J. Wu, "A maximum entropy language model integrating n-grams and topic dependencies for conversational speech recognition," *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Phoenix, AZ*, 1999.

[6] P. Lamere, P. Kwok, E. Gouvea, B. Raj, R. Singh, W. Walker, and P. Wolf, "The CMU Sphinx-4 speech recognition system," *Sun Microsystems*, 2004.

[7] C. D. Manning, *Foundations of Statistical Natural Language Processing*. MIT Press. Cambridge, MA, 1999.

[8] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[9] D. Talbot and M. Osborne, "Randomised language modelling for statistical machine translation," *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, p. 512519, 2007.

[10] E. Whittaker and P. Woodland, "Language modelling for Russian and English using words and classes," *Computer Speech and Language*, vol. 17, pp. 87–104, 2003.

[11] M. Woliński, "System znaczników morfosyntaktycznych w korpusie ipi pan (Eng. The system of morphological tags used in IPI PAN corpus)," *POLONICA*, vol. XII, pp. 39–54, 2004.

[12] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *HTK Book*. UK: Cambridge University Engineering Department, 2005.

Polish	English	occureances	%
procent	percent	782 809	0.275
za	behind, by, for	781 177	0.275
3	3	766 422	0.269
jak	how, like	752 865	0.265
roku	year (genitive and locative)	657 982	0.231
co	what	639 291	0.225
ale	but	629 278	0.221
5	5	590 454	0.207
tym	this	572 733	0.201
dla	for	571 678	0.201
jego	his	550 170	0.193
4	4	533 390	0.188
tak	yes	483 853	0.170
6	6	473 800	0.167
oraz	and	468 375	0.165
są	are	467 789	0.164
był	was	449 357	0.158
tego	that, hereof	437 488	0.154
już	already	424 368	0.149
czy	if	415 564	0.146
ma	has	410 683	0.144
ze	of, by, about, with	409 919	0.144
tylko	only	409 134	0.144
też	also	403 824	0.142
pod	under	388 418	0.137
jako	as	379 231	0.133
może	maybe	375 574	0.132
jej	her	375 023	0.132
jednak	however	374 738	0.132
ich	their	359 955	0.127
7	7	359 066	0.126
10	10	336 073	0.118
go	him	334 166	0.117
8	8	328 647	0.116
który	which	321 723	0.113
0	0	315 358	0.111
zł	PLN	312 103	0.110
było	was	306 472	0.108
20	20	305 885	0.107
także	also	296 440	0.104
lub	or	294 174	0.103
które	which (pl., fem.)	292 527	0.103
przy	next to	287 959	0.101
być	to be	286 527	0.101
będzie	will be	281 776	0.099
przed	before, in front of	280 890	0.099
9	9	280 038	0.098
ten	this	268 160	0.094
jeszcze	still	264 517	0.093
lat	years	263 891	0.092
tej	this	262 056	0.092
by		244 951	0.086
12	12	243190	0.085
była	was	240 347	0.084
15	15	237 998	0.084

Polish	English	occureances	%
bardzo	very	232 988	0.082
gdy	when	229 259	0.081
50	50	227 809	0.080
został	became	225 959	0.079
mu	him	224 764	0.079
sobie	ourself	212 314	0.075
również	also	211 971	0.074
kiedy	when	211 621	0.074
we	in	208 094	0.073
nad	over	207 888	0.073
latach	years	206 613	0.073
nawet	even	205 643	0.072
można	may	204 603	0.072
11	11	202 950	0.071
30	30	202 934	0.071
2006	2006	202 512	0.071
mnie	me	200 252	0.070
2007	2007	197 029	0.069
niż	then	195 468	0.069
21	21	192 822	0.068
22	22	192 224	0.068
bez	without	191 954	0.067
jeśli	if	191 209	0.067
18	18	190 322	0.067
linki	links	188 501	0.066
25	25	186 608	0.066
polski	Polish	185 336	0.065
tys	thousand(s)	184 389	0.065
14	14	183 077	0.064
mi	me	182 398	0.064
między	between	181 646	0.064
13	13	180 847	0.063
on	he, him	179 275	0.063
więc	so	179 225	0.063
16	16	177 938	0.062
osób	people	177 535	0.062
zewnętrzne	external	176 813	0.062
gdzie	where	176 330	0.062
polsce	Poland (locative)	175 419	0.062
19	19	173 087	0.061
miejscowość	town	170 155	0.060
tu	here	166 608	0.059
która	which (fem.)	166 421	0.058
u	in, at	166 328	0.058
mln	mln	166 291	0.058
tych	these	163 628	0.057
innymi	others	162 391	0.057
17	17	162 305	0.057

Table 5. The most popular 2-grams in the analysed Polish language corpora (r.p. - reflexive pronoun)

Polish	English	occur.	%
się w	r.p. in	330 439	0.1237
się na	r.p. {on, at}	260 365	0.0975
w tym	in this	224 702	0.0842
się z	r.p. with	191 529	0.0717
się do	r.p. to	183 245	0.0699
linki zewnętrzne	external links	174 269	0.0653
w latach	in years	169 156	0.0633
w polsce	in Poland	165 278	0.0619
nie ma	does not have	129 060	0.0483
zobacz też	look also	119 127	0.0446
nie jest	is not	114 487	0.0429
się że	r.p. that	111 635	0.0418
roku .	year.	108 944	0.0408
0 0	0 0	102 788	0.0385
jest to	is this	102 244	0.0383
że nie	that no	92 798	0.0347
na przykład	in example	90 209	0.0338
i w	and in	88 003	0.0329
w gminie	in municipality	83 990	0.0314
w stanie	in state	82 088	0.0307
pod względem	regarding	81 790	0.0306
między innymi	including	79 331	0.0297
o tym	about this	76 952	0.0288
że w	that in	76 626	0.0287
zł .	PLN.	76 496	0.0286
w tej	in this	76 207	0.0285
się .	r.p. .	74 951	0.0281
i nie	and no	74 838	0.0280
a w	and in	74 098	0.0277
wieś w	village in	71 086	0.0266
w województwie	in a voivodship	70 847	0.0265
w ciągu	during	69 108	0.0259
2007 cest	?	68 887	0.0258
tys zł	thousands of PLN	68 882	0.0258
w roku	in a year	68 292	0.0256
na to	for this	67 497	0.0253
mln zł	mln PLN	67 372	0.0252
to nie	this {not,no}	66 479	0.0249
w powiecie	in a county	66 171	0.0248
a także	and also	65 725	0.0246
w czasie	in time	63 878	0.0239
wraz z	toegether with	63 685	0.0238
znajduje się	located r.p.	62 306	0.0233
kilometrów kwadratowych	km^2	62 049	0.0232