Polish n-grams and their correction process

Bartosz Ziółko, Dawid Skurzok, Małgorzata Michalska
Department of Electronics
AGH University of Science and Technology
Kraków, Poland
bziolko@agh.edu.pl
www.dsp.agh.edu.pl

***Abstract*—Word n-gram statistics collected from over 1 300 000 000 words are presented. Eventhough they were collected from various good sources, they contain several types of errors. The paper focuses on the process of partly supervised correction of the n-grams. Types of errors are described as well as our software allowing efficient and fast corrections.**

***Index Terms*—n-grams, Polish, language modelling, speech recognition**

I. Introduction

The language properties have been very often modelled by *n*-grams [1], [2], [3], [4], [5], [6], [7]. 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 reflects 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 theoretically 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 still 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 [2], [8], [6], [7]. They have been found as the most effective models for several languages. Our attempt is to build such model for Polish language. N-grams calculated by us will be used for the 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.

Polish is very 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 [6].

We faced a 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 (www.gzegzolka.com) was used to change text files from various standards into UTF-8, some parts of files contained unexpected values which looked like they belong to a different standard.

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. All punctuations were removed and all words with symbols which were not letters of the Polish alphabet were moved to another database with errors. Standard Template Library (STL) was replaced by our own function to manage strings in aim to improve speed of basic string operations.

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

Another problem was strongly related to specific of Polish language. There are pairs of double orthographic notation for the same phoneme like *ó* and *u* or *rz* and *ż*. The pair *h* and *ch* from traditional point of view has slightly different pronunciation, however, the very most of the native speakers do not use these variations any more. Only one of the orthographic notation is correct for a particular word. There are rare cases, where there are words with different senses for different orthographical notations. However, the number of words with both notations which appeared in our statistics was worrying. This is why we investigated this topic as well.

The last issue is that our statistics show that there are more words in Polish than expected. This is why, a check against *myspell* dictionary is also necessary. However, the large number of words was probably caused mainly by the large number of proper names which are not included in dictionaries rather then errors.

There are 280 000 words in Polish *myspell* dictionary. It does not contain proper names and includes only basic forms. With all inflections over 1 000 000 words can be expected, even without proper names.

II. 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.

978-1-4244-7566-7/10/$26.00 ©2010 IEEE

TABLE II
THE NUMBER OF DIFFERENT n-GRAMS IN THE ANALYSED CORPORA.

Corpus	Basic forms	1-grams	2-grams	3-grams	single 1-grams	%	1-grams with errors	%
Rzeczpospolita journal	832 732	856 349	18 115 373	43 414 592	363 391	42.4	7435	0.86
Wikipedia	2 084 524	2 623 358	31 139 080	61 865 543	379 147	46.5	108 338	4
Literature	610 174	1 151 043	23 830 490	50 794 854	467 376	41	75 204	6.5
Transcripts	183 363	381 166	6 848 729	16 283 781	147 440	39	1 373	0.4
Literature 2	?	6 162 530	153 152 158	441 284 743	3 552 379	57.6	343 211	5.27
Literature 3	?	1 229 331	36 297 382	93 751 340	485 713	39.5	6040	0.48

TABLE I
ANALYSED TEXT CORPORA WITH THEIR SIZES, PERPLEXITY.

Corpus	MBytes	Mwords	Perplexity
Rzeczpospolita journal	879	104	8 918
Wikipedia	754	97	16 436
Literature	490	68	9 031
Transcripts	325	32	4 374
Literature 2	6500	949	6181
Literature 3	285	181	4258

Several millions of Wikipedia articles in Polish made another corpus. They are of encyclopaedia type, so they also contain many proper names including foreign ones. In total, it has 805 megabytes of text. All small articles were removed from the corpus. In this way we avoided some Wikipedia patterns like *Zawada wieś w Polsce położona w województwie łódzkim, w powiecie tomaszowskim, w gminie Tomaszów Mazowiecki. W latach 1975-1998 miejscowość należała administracyjnie do województwa piotrkowskiego. (Eng. Zawada - a village in Poland, located in Łódzki voivodeship, in Tomaszewski powiat, in Tomaszów Mazowiecki parish. During years 1975-1998, the village belonged to piotrkowskie voivodeship.)* There are over 50 000 villages described using exactly same pattern. As a result before we removed them this pattern provided the list of 5 most common 3-grams, even after combining Wikipedia with two other corpora.

The third corpus consists of several hundreds 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 of text (68 144 446 words).

The fourth corpus is a collection of transcripts from the Polish Parliament, its special investigation committees and Solidarność meetings. They contain mainly transcribed speech but also some comments on the situation in rooms. The corpus includes 325 megabytes of text (31 578 382 words). It is not as big as others but the only one containing spoken language. What is more its topics are law oriented, which corresponds very well with our project, which provides ASR system for Police, administration and other governmental institutions.

We collected another two large literature corpora which we named Literature 2 and Literature 3. They contain large books collected from Internet.

In all cases perplexity is very high comparing to typical English corpora. It is because of inflective nature of Polish and significant fraction of proper names in the corpora.

III. STATISTICS WITHOUT CORRECTIONS

Table I 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 [9]. It has to be mentioned here that around 40 % of 1-grams appeared just once for most of the corpora (see Table II). This is why the ratio of a number of 1-grams and basic forms is quite low. It should change after combaining all corpora into one statistics.

The most popular 1-grams in Polish are often pronouns, what is not surprising. They are presented in Table III, where their English translations were provided. However, it is quite difficult to translate pronouns without a context. This is why, there are sometimes 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. A fullstop is considered as a word in these statistics as it brings modelling information for ASR system.

Some English, Russian, Chinese and other foreign words appeared in the statistics as well as single letters. Such words could be effects of including some foreign citations in articles. However, most of the foreign words are proper names and they appeared in Polish sentences.

Collected statistics of individual corpora 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.

IV. CORRECTION PROCESS

After an analysis of results of collecting n-gram statistics from various corpora we decided that some supervised correction is necessary. Because of the amount of data the choice of strategy in this process was crucial from financial point of view.

We designed and implemented software Fixgram to optimise n-gram corrections from time point of view. This is why the interface was prepared in a way a person unfamiliar with databases could conduct corrections. All displays and buttons were located to save user time according to standard human-computer interaction rules [10], [11]. So, first, all buttons are

TABLE III

THE MOST POPULAR 1-GRAMS IN SOME OF THE ANALYSED POLISH LANGUAGE CORPORA PRESENTED WITH THE PERCENTAGE REGARDING TO THE WHOLE TEXT. 1% STANDS FOR 2 843 204 OCCUREANCES. (R.P. - REFLEXIVE PRONOUN)

Pol.	Eng.	%	Pol.	Eng.	%
.	.	6.066	przed	before, in front of	0.099
w	in	3.143			
i	and	1.800	9	9	0.098
na	on, at	1.499	ten	this	0.094
z	with	1.390	jeszcze	still	0.093
się	r.p.	1.373	lat	years	0.092
do	to, till	1.010	tej	this	0.092
nie	no, not	1.007	by	would	0.086
to	it, this	0.622	12	12	0.085
że	that	0.583	była	was	0.084
jest	is	0.524	15	15	0.084
o	about, at	0.497	bardzo	very	0.082
a	and	0.488	gdy	when	0.081
1	1	0.379	50	50	0.080
od	from, since	0.359	został	became	0.079
po	after	0.333	mu	him	0.079
przez	through	0.311	sobie	ourself	0.075
2	2	0.310	również	also	0.074
0	0	0.304	kiedy	when	0.074
procent	percent	0.275	we	in	0.073
za	behind	0.275	nad	over	0.073
	by, for		latach	years	0.073
3	3	0.269	nawet	even	0.072
jak	how, like	0.265	można	may	0.072
roku	year	0.231	11	11	0.071
	(genit. & loc.)		30	30	0.071
co	what	0.225	2006	2006	0.071
ale	but	0.221	mnie	me	0.070
5	5	0.207	2007	2007	0.069
tym	this	0.201	niż	then	0.069
dla	for	0.201	21	21	0.068
jego	his	0.193	22	22	0.068
4	4	0.188	bez	without	0.067
tak	yes	0.170	jeśli	if	0.067
6	6	0.167	18	18	0.067
oraz	and	0.165	linki	links	0.066
są	are	0.164	25	25	0.066
był	was	0.158	polski	Polish	0.065
tego	that, hereof	0.154	tys	thousand(s)	0.065
już	already	0.149	14	14	0.064
czy	if	0.146	mi	me	0.064
ma	has	0.144	między	between	0.064
ze	of, by,	0.144	13	13	0.063
	about, with		on	he, him	0.063
tylko	only	0.144	więc	so	0.063
też	also	0.142	16	16	0.062
pod	under	0.137	osób	people	0.062
jako	as	0.133	zewnętrzne	external	0.062
może	maybe	0.132	gdzie	where	0.062
jej	her	0.132	polsce	Poland (loc.)	0.062
jednak	however	0.132	19	19	0.061
ich	their	0.127	miejscowość	town	0.060
7	7	0.126	tu	here	0.059
10	10	0.118	która	which (fem.)	0.058
go	him	0.117	u	in, at	0.058
8	8	0.116	mln	mln	0.058
który	which	0.113	tych	these	0.057
0	0	0.111	innymi	others	0.057
zł	PLN	0.110	17	17	0.057
było	was	0.108	pan	mr	0.057
20	20	0.107	były	were	0.056
także	also	0.104	23	23	0.056
lub	or	0.103	powiedział	said	0.056
które	which	0.103	miał	had	0.055
	(pl., fem.)		ją	her	0.055
przy	next to	0.101	teraz	now	0.054
być	to be	0.101	tam	there	0.054
będzie	will be	0.099	bo	because	0.053
te	these	0.053	ii	2 (latin)	0.046
nich	them	0.053	natomiast	however	0.045
według	acording	0.051	pracy	work (loc.)	0.045
podczas	while	0.051	1998	1998	0.045
nim	him	0.050	40	40	0.045
których	which	0.050	zobacz	look	0.044
urodzony	born	0.049	center	center	0.044
aby	to	0.049	je	eats, them(fem.)	0.044
miejsce	place	0.049	innych	others	0.043
danych	data	0.048	wszystko	everything	0.043
ja	I	0.048	dwa	two	0.043
24	24	0.047	jeden	one	0.043
rok	a year	0.047	potem	later	0.042

TABLE IV

THE MOST POPULAR 2-GRAMS IN THE ANALYSED POLISH LANGUAGE CORPORA (R.P. - REFLEXIVE PRONOUN). 1 % STANDS FOR 2 671 293 OCCUREANCES

Polish	English	%
się w	r.p. in	0.1237
się na	r.p. {on, at}	0.0975
w tym	in this	0.0842
się z	r.p. with	0.0717
się do	r.p. to	0.0699
linki zewnętrzne	external links	0.0653
w latach	in years	0.0633
w polsce	in Poland	0.0619
nie ma	does not have	0.0483
zobacz też	look also	0.0446
nie jest	is not	0.0429
się że	r.p. that	0.0418
roku .	year.	0.0408
0 0	0 0	0.0385
jest to	is this	0.0383
że nie	that no	0.0347
na przykład	in example	0.0338
i w	and in	0.0329
w gminie	in municipality	0.0314
w stanie	in state	0.0307
pod względem	regarding	0.0306
między innymi	including	0.0297
o tym	about this	0.0288
że w	that in	0.0287
zł .	PLN.	0.0286
w tej	in this	0.0285
się .	r.p. .	0.0281
i nie	and no	0.0280
a w	and in	0.0277
wieś w	village in	0.0266
w województwie	in a voivodship	0.0265
w ciągu	during	0.0259
tys zł	thousands of PLN	0.0258
w roku	in a year	0.0256
na to	for this	0.0253
mln zł	mln PLN	0.0252
to nie	this {not,no}	0.0249
w powiecie	in a county	0.0248
a także	and also	0.0246
w czasie	in time	0.0239
wraz z	toegether with	0.0238
znajduje się	located r.p.	0.0233
kilometrów kwadratowych	km^2	0.0232
z nich	of them	0.0231
w warszawie	in Warsaw	0.0229
się nie	r.p not	0.02267

grouped to allow different decisions with just slight moves of the mouse. The edit box is near buttons for the same reason. However, the editing decision is rarlier used than *delete* or *keep*. Secondly, the most important information is displayed in the left, upper part of the window, close to the buttons, because looking at this part of a screen is typically the first thing people instinctively do. More detailed information are displayed at the lower part. Thanks to it a user has to look there only in case of more complicated situations.

The list of words for corrections is preprepared on a server. This is why it is partly unsupervised method. Two schemes of preparing words were used so far. The first one is finding pairs of words which are different only by orthographic notation, in example *rz* and *ż*. The second is by finding words with any unusual letters (not Polish ones). Both processes are conducted before the human work. The user of Fixgram receives a database of words chosen for corrections to save time spent on automatic search for them in a database during human work. The third scheme is by comparison with *myspell* dictionary. The words which do not exist in *myspell* are also more likely to be errors then others.

All chosen words are given to the Fixgram user in order by the number of times they appeared in a corpus. In this way, a threshold can be set above which human decisions are necessary. All rarelier cases will be done automatically, typically by deleting. There is no reason in spending human time for rare cases which are likely to be incorrect and not crucial for statistics. Sometimes human decisions can be generalised and used for rarlier cases.

Another very important issue is that all human decisions are saved not only by changing the database with statistics but also as patterns of corrections in case the same problem appears in another corpus.

A few types of problems were encountered. The first one are Chinese and some English proper names. Chinese words often are written with *ch* an *h* like *Chien* and *Hien*. They appeared quite frequently in newspaper corpus. Chinese proper names tend to be also often in addition to a Polish word, so one orthographic transcription is for a Polish word and the other for Chinese proper name.

Another type of a problem are words which were split into two words with a space so they appeared as two separate words in n-grams. These are difficult to be found automatically.

There are also words which are wrongly formatted (not in UTF-8). Most of them are not in any of known to us standards for Polish letters. This is because we changed all typical standards to UTF-8 before collecting the statistics. These words can still be recognised by a human, as typically there is only one special Polish letter and other are standard Latin letters. Fixgram presents contexts of each word (2 and 3-grams) what makes much easier to correct these cases.

Apart from that, we discovered quite a lot of Russian words and single letters (in Cyrillic). All of those are removed.

Several automatically detected words were correct. In example there are plenty of similar surnames with an only difference in one Polish special character. There were some other words which are correct with both orthographic transcriptions but different senses, like morze (Eng. sea) and może (Eng. maybe). These cases were kept in the n-gram database by a human decision.

V. Conclusions

N-gram statistics for Polish were presented. The strategy to automatically detect errors was described, as well as the Fixgram software tool to allow a human to take final decisions for possible errors. All decisions are kept to allow fully automatic decisions in the future.

VI. Acknowledgements

This work was supported by MNISW grant number OR00001905.

References

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

[2] 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.

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

[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] 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.

[7] 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.

[8] 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.

[9] 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.

[10] J. Preece, Y. Rogers, and H. Sharp, *Interaction design:Beyond human-computer interaction. 2nd edn.* Wiley, 2007.

[11] A. Dix, J. Finlay, G. Abowd, and R. Beale, *Human-Computer Interaction, 3rd edition.* Pearson Education, 2004.

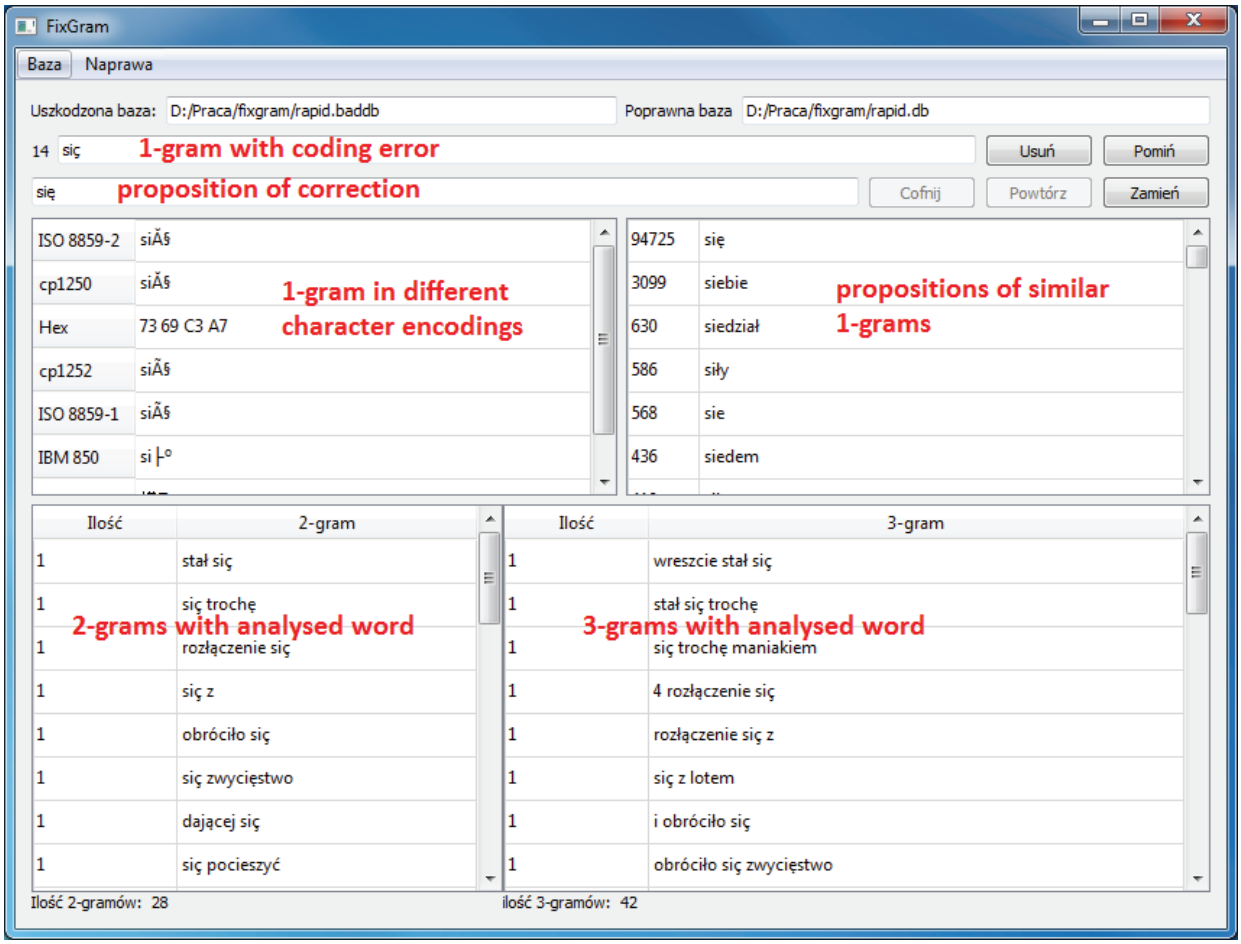

Fig. 1. Screenshot of our Fixgram software to correct n-gram statistics