COGNITIVE STUDIES | ÉTUDES COGNITIVES, 17
Warsaw 2017

Article No.: 1468
DOI: 10.11649/cs.1468

Citation: Mykowiecka, A., Marciniak, M., & Rychlik, P. (2017). Testing word embeddings for Polish. *Cognitive Studies | Études cognitives*, *2017*(17). https://doi.org/10.11649/cs.1468

AGNIESZKA MYKOWIECKA[A], **MAŁGORZATA MARCINIAK**[B], & **PIOTR RYCHLIK**[C]

Institute of Computer Science, Polish Academy of Sciences, Warsaw, Poland
[A]agn@ipipan.waw.pl ; [B]mm@ipipan.waw.pl ; [C]rychlik@ipipan.waw.pl

TESTING WORD EMBEDDINGS FOR POLISH

Abstract

Distributional Semantics postulates the representation of word meaning in the form of numeric vectors which represent words which occur in context in large text data. This paper addresses the problem of constructing such models for the Polish language. The paper compares the effectiveness of models based on lemmas and forms created with Continuous Bag of Words (CBOW) and skip-gram approaches based on different Polish corpora. For the purposes of this comparison, the results of two typical tasks solved with the help of distributional semantics, i.e. synonymy and analogy recognition, are compared. The results show that it is not possible to identify one universal approach to vector creation applicable to various tasks. The most important feature is the quality and size of the data, but different strategy choices can also lead to significantly different results.
Keywords: distributional semantics; word embeddings; model evaluation; synonymy; analogy

1 Introduction

Distributional Semantics (DS) is currently widely-used in many tasks in the domain of Natural Language Processing (NLP). Its main assumption is that the meaning of a word can be inferred (to some extent) from its usage. Therefore, in DS models words are represented as vectors whose positions directly or indirectly represent information about the frequency of the particular word occurring in its context. The underlying concept of this approach is not new. The suggestion that "the meaning of words lies in their use" was formulated by Wittgenstein in 1953 and, in 1957, Firth stated "You shall know a word by the company it keeps!". At around the same time Harris, in the paper "*Distributional structure*" (Harris, 1954), formulated an idea which can be directly implemented in computer programs: "The distribution of an element will be understood as the sum of all its environments. An environment of an element A is an existing array of its co-occurrents". However, in spite of this theoretical support, the idea of distributional semantics was to remain rather marginal for quite some time. This situation changed in the late 1990s, with enhancements in both language corpora availability and technical capabilities, and when distributional methods had proven themselves effective in both modelling cognitive phenomena and in practical applications. They have been used, for example, for word sense disambiguation problems (Schutze, 1998);

to model human similarity judgements (McDonald, 2000); to enhance n-gram language models with long range semantic information (Bellegarda, 2000; Coccaro & Jurafsky, 1998); to identify synonyms (Landauer & Dumais, 1997); and to model semantic priming (Lund & Burgess, 1996). Positive and easily achievable results, and the creation of publicly available tools for building distributional models, has increased the popularity of this approach even further.

Lying at the core of the distributional approach is the vector representation of words. A vector model is built on the basis of an appropriate corpus — a set of texts, either plain or annotated with some morphosyntactic features. As the distribution data is more reliable when a source text is large, it is common practice to use an existing large corpus of general texts or even to combine several corpora. As usual with NLP technology, texts should cover the appropriate domain and genre. Building more complex models requires the addition of different types of annotation, which can be done with the help of linguistic tools or manually, e.g. using the crowdsourcing approach. For every word from a given corpus, one can count the contexts in which it occurs. Collected contexts create a huge matrix, which is then transformed (e.g. using linear algebra) into a matrix approximating meaning. Each row in this matrix is a vector representation of one entity (usually a word). The similarity of vectors can be measured with standard mathematical functions, for example the cosine of the angle between them. Similar vectors are considered to represent related words. The relatedness of words is general and cannot be precisely defined. In this paper, as in Budanitsky and Hirst (2006), it consists of well-established relations such as: synonymy (*amazing – wonderful*) and antonymy (*good – bad*); hyperonymy and hyponymy (*bird – crow*; *cutlery – spoon*); co-hyponymy (*coffee – tea*, *dog – cat*); meronymy (*flat – room*); and other functional associations (*coffee – cup*, *state – legislation*).

Transforming corpus data into vector representations can be done in several ways. The most direct, count-based strategy consists of collecting all context data from all word occurrences and then transforming the resulting matrix using some kind of weighting function. Weighting is aimed at strengthening surprising events and weakening highly expected events, because it is more informative if something rare occurs than if something quite common takes place. In DS models, this means that having a rare context in common, e.g. 'roar', should make words more similar than having more typical common contexts, e.g. 'run'. The most commonly used method of formalizing the idea of rare and frequent words for term-document matrices is the *tf–idf* (term frequency $\times$ inverse document frequency) function (Spark Jones, 1972). In information theory, a surprising event has a higher information content than an expected event (Shannon, 1948). A frequently used alternative to *tf–idf* is PMI (Pointwise Mutual Information; Church & Hanks, 1989; Turney, 2001). The final, optional, step in building a DS model is dimensionality reduction, which aims to establish the most informative dimensions, usually from hundreds of thousands of different contexts. Dimensionality reduction can be performed by feature selection but it is typically done by SVM (Singular Value Decomposition), being the core of the Latent Semantic Analysis/Latent Semantic Indexing (LSA/LSI) method (Landauer, Foltz, & Laham, 1998). It constructs a low-rank approximation to the word-context matrix.

The second way to transform context counts into vectors representing word meanings (word embeddings) is called Global Vectors (Pennington, Socher, & Manning, 2014). The main concept behind GloVe is the observation that ratios of word-word co-occurrence probabilities can encode some sense of meaning. The training objective of GloVe is to transform original frequency-based word vectors so that their dot product equals the logarithm of the words' probability of co-occurrence.

The third method of building distributional models, and the one which has probably gained the most spectacular popularity, is to train a neural network (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013) to predict a word given a context (CBOW approach), or a context given a word (skip-gram approach), on the basis of a corpus in which every word occurrence represents one learning example. In this approach, word sense is represented as a vector of the neural network layer. This method was implemented by the author as the word2vec algorithm, which uses a very efficient learning strategy, allowing much faster neural networks model building than those

previously used.

All the methods of constructing vector representations for words have been applied to many NLP tasks, e.g. finding synonyms (Griffiths, Steyvers, & Tenenbaum, 2007), word clustering (Kovatchev, Salamo, & Marti, 2016), sentiment analysis (Duyu, Wei, Yang, Ming, Ting, & Bing, 2014), word sense disambiguation (Basile, Caputo, & Semeraro, 2014), metaphor recognition (Shutova, Sun, Gutierrez, Lichtenstein, & Narayanan, 2017), sentence paraphrasing (Dinu & Baroni, 2014), and documents classification (Kim, Kim, & Cho, 2017). Many problems are best solved with the help of DS models, e.g. identifying synonyms. A description of the main assumptions of distributional semantics, as well as the basic bibliographical references, can be found, for example, in Baroni and Lenci (2010); Turney and Pantel (2010); Clark (2015).

Although distributional semantics has become very popular, there are only a few published papers concerning the vector representation of Polish words. There is a tool — Supermatrix — which builds a distributional model, taking into account a predefined set of features and SVD dimensionality reduction, and computes word similarity (Broda & Piasecki, 2008, 2013). Kędzia, Czachor, Piasecki, and Kocoń (2016) published a skip-gram model of Polish created by word2vec with 100-dimensional feature vectors. The presentation *word2vec dla Polskiego Internetu* "word2vec for Polish Internet" (Stokowiec, 2015) is available on the Internet. The problem of synonyms and lexical variants is described in the paper (Tatjewski, Bańko, Kucińska, & Rączaszek-Leonardi, 2017). Rogalski and Szczepaniak (2016) published a paper concerning the creation of word embeddings and embeddings themselves. They re-implemented the Mikolov, Sutskever, et al. (2013) algorithm and created vectors from Polish Wikipedia.

This paper verifies the distributional semantic models (DSM) for Polish created by word2vec from the genism package (Řehůřek & Sojka, 2010), `https://radimrehurek.com/gensim/`, and compares them with previously published resources. The functionalities available in the word2vec tool are tested to discover which parameter values are the best for processing Polish — a highly inflectional language. Models based on lemmas and forms for corpora, consisting of Polish Wikipedia (WikiPL) and the National Corpus of Polish (NKJP; Przepiórkowski, Bańko, Górski, & Lewandowska-Tomaszczyk, 2012), are created. The results obtained by CBOW and the skip-gram architecture using 100- and 300-dimensional vectors are compared. Moreover, the paper examines how the removal of infrequent forms from the data influences the results.

The evaluation of DSM has been the subject of many studies. Two ways of performing this evaluation are possible: intrinsic evaluation (testing a system in itself), e.g. Tsvetkov, Faruqui, Ling, Lample, and Dyer (2015), and extrinsic evaluation, measuring its performance in a task or application, e.g. Cheung and Penn (2012), which reports on testing syntactically invariant inference. The problem with performing an extrinsic evaluation is that task-oriented benchmarks adopted in distributional semantics tasks, such as the TOEFL synonim detection task, have not been specifically designed to evaluate DSMs. Thus, the results obtained reveal more about the particular solution of the task than about a specific element of the processing flow, i.e. in this case a DS model. To gain a real insight into the abilities of DSM, Baroni and Lenci (2011) postulate that existing benchmarks must be complemented with a more intrinsically oriented approach. Although aware of the many problems also identified for intrinsic DSM evaluation, described for example in Faruqui, Tsvetkov, and Rastogi (2016) and Jastrzebski, Leśniak, and Czarnecki (2017), it was decided to perfom such an evaluation using already available data, in order to gain some knowlegde about the differences in the quality of various word models for Polish. As there are still no sets designed to test the specific aspects of lexical knowledge for Polish, it was decided to use two existing lexicons of synomyns. In order to make the comparison more robust, a set of analogy pairs covering many types of relations apart from synonymy were defined.

2 Corpora description

The experiments with DS models employ the NKJP and WikiPL corpora, as well as the combined set of these two corpora. A small, openly-accessible subset of NKJP is downloadable from `http://clip.ipipan.waw.pl/NationalCorpusOfPolish`, but for this paper the full data of the NKJP project consortium was used, by permission of the project leader. NKJP and Wikipedia, dump of late 2016 (`https://dumps.wikimedia.org/plwiki`), were annotated using Concraft-pl (Waszczuk, 2012), a morphosyntactic tagger for Polish based on constrained conditional random fields. Several input sets for building DS models were prepared. They can be divided into two main groups: one containing orthographic forms and one containing lemmas generated by a tagger. All sentences were scanned to remove tokens that are punctuation marks, or which contain characters other than a letter or digit. All words were converted to lower case, unless capital letters had been found in their lemmas.

Unfortunately, using data annotated by Concraft-pl has a potential drawback that may influence the results of experiments. Some words, mainly verb forms, are divided into several tokens. For example, the word *chciałbym* 'I would like' will be split into three tokens: *chciał* (past tense), *by* (qublik), and *m* (agglutinate). Similarly, the word *biało-czerwony* 'white and red' will be split into *biało*, punctuation mark '-', and *czerwony*. To test the influence of some potentially not very informative tokens like '-', '*by*' or '*m*', restricted data sets were prepared, which only included tokens classified as nouns, adjectives, adverbs, verb forms, and abbreviations, which constitute 19 parts of speech (POS) out of the 34 foreseen in NKJP. All other words are treated as if they do not exist in the data. The sizes of the corpora used are shown in Table 1 below.

Table 1: Corpora sizes (in millions)

	sentences	tokens	unique forms	unique lemmas
WikiPL	12	184	3	2.6
WikiPL-restricted	12	137	2.7	2.2
NKJP	107	1,482	9.2	8.4
NKJP-restricted	107	1,044	8.2	7.4

3 Models

There are many assumptions that may influence the performance of a particular vector model in a particular task. These assumptions may be divided into three different categories and concerns:

- model elements, i.e. if a model is built for word forms or lemmas or for more complicated structures (such as a word being a noun, or a particular word being the object of a particular verb);
- context definition, i.e. whether all or only selected words will be taken into account as context values, and which features to include, e.g. only word forms, their POS, grammatical relations, etc.;
- the method used for transforming raw data into a final model;
- the values of parameters specific for a chosen method.

All the models used in this paper were built using genism word2vec. In the description below, a naming convention for the models is given in brackets.

Both the CBOW (c) and skip-gram (s) approaches were tested. The models were built on NKJP data (N), Wikipedia (W), and the two corpora joined together (NW) consisting of either:

- word forms (fa);
- lemmas (la);
- word forms restricted to 19 out of 34 POS (fr);
- lemmas restricted to 19 out of 34 POS (lr);
- lemmas combined with a part of speech name (-pos).

As learning strategies, the experiment used either hierarchical softmax (h) or negative sampling (n) in the standard configuration of 5 positive examples and 1 negative. The number of features, i.e. different types of contexts represented for one word, was either 100 (1) or 300 (3). The context size is equal to 5, the minimal number of occurrences is 5 and there are 10 learning steps. To test the influence of rare words (and some no-words, spelling errors, etc.) selected models were built, limited to words occurring no fewer than 50 times for NKJP data or no fewer than 30 times for Wikipedia data. These models are: *NWfa-3-s-h50* (NKJP plus WikiPL, all word forms, CBOW, 300 features, hierarchical softmax, word form occurrences no fewer than 50), *NWfa-3-c-n50* (the same as before, but with negative sampling), *Wfa-3-c-h30* and *Wfa-3-c-n30*. For selected models, it was also tested whether increasing the number of training steps to 100 influences the results (-it).

As well as these models, publicly available models (named *pl-emb-c* and *pl-emb-s*) from the paper (Rogalski & Szczepaniak, 2016) were also used. These are CBOW and skip-gram models with negative sampling trained on pre-processed data from Wikipedia. All Wikipedia text was changed to lower case, numbers were divided into separate digits and converted to words, and some non-text elements were deleted. The skip-gram model published by Kędzia et al. (2016) was not used due to technical problems with processing it. Moreover, there are no details about the corpus or the exact technique that was used to obtain the data.

4 Tasks description

The main problem when comparing many alternative models is to establish a relatively large scale through a repeated experiment which uses open and high-quality data. Satisfying all these requirements is very difficult, and in many cases even unfeasible, as preparing test data is highly labour-intensive. It is well-documented that the similarity of word embeddings reflects many different relations between the given words, e.g. synonymy, antonymy, hypernymy or hyponymy (Scheible, Schulte im Walde, & Springorum, 2013; Weeds, Clark, Reffin, Weir, & Bill, 2014). For example, in one of the models used in this paper the most similar words to *przyjazd* 'arrival, using some ground vehicle' are: *przylot* 'arival by plane', *wyjazd* 'departure' and *przybycie* 'arrival, no means specified'; to *cichy* 'quiet, noiseless' the most similar words are: *wesoły* 'cheerful', *delikatny* 'delicate', *cichutki* 'quiet, barely audible', *miły* 'nice'. To allow for massive and maximally objective tests two specific problems were chosen to be solved using all of the models: synonym identification and analogy testing

4.1 Synonymy

From the many possible relatedness relations only one, synonymy, was selected, as it facilitated the preparation of the test data and the interpretation of the results. It was assumed that if word embeddings correctly represent word senses, then they should also correctly indicate synonyms as words whose embeddings are very close. The larger the number of synonyms at the top of the ranked list of similar words, the better the model represents word senses. It should be stressed that the goal was not to elaborate the most efficient way of finding synonyms, but rather to ascertain which model represents word synonymy in the best way. For this reason, we did not add any additional methods for filtering non-synonyms from the ranked similar word lists, but instead evaluated the original lists obtained using different model settings.

For the evaluation of the different embedding models, two publicly available collections of synonyms were selected. The first is a free online resource, created and edited by volunteers. The lexicon contains more than 600,000 synonyms of almost 150,000 words and it is available at `www.synonim.net`. The second was created and is maintained by Wojciech Broniarek and its original version was published as a synonym lexicon entitled *Gdy Ci słowa zabraknie* "When You Are Lost For Words". It is still edited and extended by the author, and is also available on-line at `www.synonimy.pl`. This set is smaller, as the adapted rules for assigning synonyms to lexical entries are more restrictive. For example, for the word *słowo* 'word', the first set contains 74 synonyms, while the second one only 15. Most of the synonyms are single words, but some multi-word phrases are also included. Both lexicons differentiate word meanings, but as we do not build representation for senses but for words, we merge synonyms for all word senses together. Tests were limited to words from the three main syntactic categories: nouns, verbs and adjectives. Each category is represented by 50 frequent words selected from the list of NKJP forms. Moreover, an attempt was made to select forms/words which are not ambiguous in terms of parts of speech. When selecting nouns, those which can also be gerunds were avoided. In Table 2 the least and the most frequent words, together with the number of their occurrences in the combined NKJP and WikiPL, corpus are given.

Table 2: The least and the most frequent words together with the number of their occurrences in the combined NKJP and WikiPL corpus (first number) and WikiPL only (second number)

	The least frequent word				The most frequent word			
	form	count	lemma	count	form	count	lemma	count
adjective	pikantny	375	pikantny	2,675	cały	340,566	inny	3,886,370
	'spicy'	55	'spicy'	159	'whole'	20,079	'other'	412,575
noun	pieniądz	8,295	mowa	201,400	czas	637,066	czas	2,775,604
	'money'	341	'talk/speech'	6,550	'time'	36,358	'time'	285,295
verb	mącić	492	mącić	3,536	powiedzieć	341,157	zostać	3,757,034
		3		31	'say'	2,522	'stay'	912,736

It was decided to perform the tests separately for words of different categories (the selection concerns only the sets of tested words, lists of the most similar words were not filtered from words of different categories) to check if the difference in syntactic structures in which these words occur might influence the results. We also wanted to test how inflection influences the quality of embeddings. In Polish, all nouns have gender and have to agree in gender with the modifying adjectives. Gender agreement is also visible between a noun which is the subject of a verb in the third person in the past tense and the verb itself, e.g. "$Szkoła_{fem}$ $była_{fem}$ $zamknięta_{fem}$." 'The school was closed.', "$Dworzec_{masc}$ $był_{masc}$ $zamknięty_{masc}$." 'The station was closed.'. Third person constructions occur very frequently in texts — in Wikipedia, there are 60 times more 3rd person than 1st person constructions and in NKJP the proportion is roughly 3.5 to 1. This may lead to different lists of the most similar suggestions for synonyms of different genders. For example, the list of related words for $osoba_{fem}$ 'person' contains mainly feminine nouns like *kobieta* 'woman', *dziewczyna* 'girl', *osóbka* 'wench', *duszyczka* 'soul', *prostytutka* 'prostitute' while the list of related words for $człowiek_{masc}$ 'man' contains masculine nouns like *mężczyzna* 'man', *facet* 'guy', *chłopak* 'boy', *osobnik* 'individual', *chłop* 'peasant', and *on* 'he'. To see whether gender inflection influences the similarity results, models were built on word lemmas and directly on word forms. To evaluate this second set of models, we generated all the forms of synonyms taken from both lexicons and all the forms of the selected lexicon entries. While models based on lemmas could possibly overcome

inflection features disagreement, we wanted to ascertain whether or not they are influenced by the relatively poor quality of Polish lemmatizers (in the case when there is more than one word of the same syntactic category, Polish taggers do not assign lemmas with great precision, e.g. they quite frequently assign to the word form *mają* 'they have' lemma 'to adorn with verdure' (*maić*) not 'to have'.). Table 3 shows the names and the number of elements of the two synonym sets defined for the three syntactic categories.[1]

Table 3: Cardinality of test sets for synonyms

	adjectives	nouns	verbs
synonim.net	A1 – 7186	N1 – 6709	V1 – 4107
synonimy.pl	A2 – 1326	N2 – 984	–

4.2 Analogy

The second task concerns the identification of analogy (Baroni, Dinu, & Kruszewski, 2014; Mikolov, Yih, & Zweig, 2013). This type of relation is open and is defined by a pair of words that are in this relation, e.g. *jesień-deszcz* 'autumn-rain' or *Polska-Warszawa* 'Poland-Warsaw'. The algorithm has to identify the word that is in the same relation with a new word given as an input. Thus, in the first case, for the word *zima* 'winter', we would expect *śnieg* 'snow' (for data in Polish at least), and in the second case, for *Francja* 'France', we would expect *Paryż* 'Paris'. In this task, the selection of both the initial pairs and the test words is crucial. The relations between two words can be hard to recognize, as in *filiżanka-kot* 'cup-cat', if we have in mind 'something that can be broken by a cat'. This task has its source in the college admission test in the United States — SAT (Shaw, 2015), which includes this type of questions. There are a lot of examples of such tests, in particular on the word2vec page `https://code.google.com/archive/p/word2vec/source/default/source` there is a file, question-words.txt, for model checking in English. A list of Polish pairs was prepared, taking into account the relations represented in this file and adding several other types of relations. The list consists of 200 elements, each consisting of two pairs of words that are in the same relation.

To test if a pair of words $(a - b; c - d)$ represents analogy, we performed the standard test on the vectors representing words: $(\overrightarrow{b} + \overrightarrow{c}) - \overrightarrow{d}$ and it was expected that, as a result, the nearest vector to the resulting one $(\overrightarrow{d})$, would represent the word d. For each pair, we tested the 10 first nearest vectors and checked if they were consistent with the word given in the pair. The tested relations, divided into groups of similar ones (together with the number of examples), are given in Table 4. Moreover, we prepared 20 additional analogies representing grammatical relations, e.g. *kot–kotom* 'cat-cat$_{\text{pl,dat}}$' *pisał–pisała* 'wrote–wrote$_{\text{fem}}$' and *mały–mniejszy* 'small–smaller'. These were only tested on form-based models. The test contained: noun-noun in plural (3); noun in the nominative case-noun in a different case (3); noun–noun in various cases and numbers (7); adjective-adjective in the higher degree (2); adjective-adjective in different number and gender (1), verbs in the present-past tenses (3) and verb in the singular-plural (1).

[1]The reason why we did not use Polish WordNet as an alternative source of synonyms was the fact that it did not provide any synonyms for more than 15% of the words in our test sets.

Table 4: Analogies list

	Relation description	#	Example	Translation
1	family	9	matka-córka; ojciec-syn	mother-daughter; father-son
2	profession	3	kobieta-kucharka; mężczyzna-kucharz	woman-cook; man-cook
3	diminutives	10	kot-kotek; pies-piesek	cat-kitten; dog-little dog
4	adjective/adverb	2	wesoły-wesoło; szybki-szybko	merry-merrily; fast-fast
5	Performer/action	2	biegacz-biegać; pływak-pływać	runner-run; swimmer-swim
6	animal offspring	9	koń-źrebak; pies-szczeniak	horse-colt; dog-puppy
7	animal/sex	20	byk-krowa; baran-owca	bull-cow; ram-sheep
8	fruits and vegetables	6	gruszka-owoc; pomidor-warzywo	pear-fruit; tomato-vegetable
9	drink/vessel	6	kawa-filiżanka; wino-kieliszek	coffee-cup; wine-glass
10	meal/food	4	śniadanie-kanapka; obiad-kotlet	breakfast-sandwich; dinner-cutlet
11	trees/leaves	1	klon-liść; świerk-igła	maple-leaf; spruce-needle
12	senses	1	oko-obraz; ucho-dźwięk	eye-picture; ear-sound
13	animal type/supertype	6	małpa-ssak; wąż-gad	monkey-mammal; serpent-reptile
14	animal supertype/type	1	pies-buldog; kot-ragdoll	dog-bulldog; cat-ragdoll
15	mean of transport/drive	3	żaglówka-żagiel; motorówka-silnik	sailboat-sail; motorboat-engine
16	season/phenomenon	3	lato-ciepło; zima-zimno	summer-warm; winter-cold
17	drink/made from	3	wino-winogrono; cydr-jabłko	wine-grape; cider-apple
18	profession/place of work	6	kucharka-kuchnia; nauczyciel-szkoła	cook-kitchen; teacher-school
19	profession/product	7	piekarz-chleb; cukiernik-tort	baker-bread; confectioner-cake
20	place of growth/ plant	3	ogródek-warzywo; łąka-trawa	garden-vegetable; meadow-grass
21	food/made from	3	chleb-zboże; ser-mleko	bread-grain; cheese-milk
22	cultural place/event	3	teatr-sztuka; filharmonia-koncert	theatre-play; philharmonic-concert
23	work/author	6	powieść-pisarz; symfonia-kompozytor	novel-writer; symphony-composer
24	person/vehicle	3	kapitan-statek; kierowca-samochód	captain-ship; driver-car
25	vehicle/route	3	samochód-droga; statek-rzeka	car-road; ship-river
26	parts of plants	1	pień-drzewo; łodyga-kwiat	trunk-tree; stalk-flower
27	product/made of	6	drewno-mebel; skóra-but	wood-furniture; leather-shoe
28	clothes/body parts	3	szal-szyja; rękawiczka-ręka	shawl-neck; glove-hand
29	doctor/patient	3	lekarz-pacjent; weterynarz-pies	physician-patient; vet-dog
30	physician/body part	4	dentysta-ząb; okulista-oko	dentist-tooth; ophthalmologist-eye
31	athlete/equipment	3	kolarz-rower; narciarz-narty	cyclist-bike; skier-skis
32	clergyman/place	6	ksiądz-kościół; zakonnik-klasztor	priest-church; monk-convent
33	expression/feeling	1	płacz-smutek; śmiech-radość	cry-sadness; laughter-joy

34	thing/element	4	drzwi-dom; brama-ogród	door-house; gate-garden
35	tool/applied to	1	młotek-gwóźdź; śrubokręt-śruba	hammer-nail; screwdriver-screw
36	tool/action	1	młotek-wbijać; wiertarka-wiercić	hammer-hammer; drill-drill
37	creator/action	1	pisarz-pisać; malarz-malować	writer-write; painter-paint
38	teacher/learner	1	nauczyciel-uczeń; wykładowca-student	teacher-pupil; lecturer-student
39	element/natural disaster	1	woda-powódź; ogień-pożar	water-flood; fire-fire
40	sex/cloth	1	kobieta-garsonka; mężczyzna-garnitur	woman-woman's suit; man-suit
41	parts of clothes	1	garsonka-spódnica; garnitur-spodnie	woman's suit-skirt; suit-trousers
42	syntactic derivatives	3	dom-domowy; biuro-biurowy	home-home; office-of office
43	antonyms	4	czysty-brudny; stary-nowy	clean-dirty; old-new
44	state/capital	11	Francja-Paryż; Japonia-Tokio	France-Paris; Japan-Tokyo
45	town/river	6	Paryż-Sekwana; Warszawa-Wisła	Paris-Seine; Warsaw-Vistula
46	state/continent	3	Francja-Europa; Chiny-Azja	France-Europe; China-Asia
47	state/inhabitant	4	Polska-Polak; Hiszpania-Hiszpan	Poland-Pole; Spain-Spaniard
48	geo. place/adjective	5	Szwecja-szwedzki; Polska-polski	Sweden-Swedish; Poland-Polish
49	capital/region	1	Warszawa-Mazowsze; Kraków-Małopolska	Warsaw-Mazovia; Cracow-Malopolska

5 Results

5.1 Synonymy

To assess the performance of the models in the task of synonymy identification, we produced lists of the most similar words (using the cosine similarity measure) to all the members of the fifty-element test sets of nouns, adjectives and verbs. We then searched these lists for elements from the synonym sets from Table 3. For lemma-based models, the comparison is straightforward. In order to test the results for the form-based models, we computed all forms of the nouns and adjectives given on the synonyms lists using the morphosyntactic analyser and generator Morfeusz 2 (Woliński, 2014). All verbs were left in their infinitive forms. We counted the precision of the results at every ten elements of these lists up to the 50^{th} position. That is to say, we counted the percentage of the words from the synonyms list among the first ten, twenty, thirty results and so on. For models based on word forms, we counted as synonyms the inflected forms of the tested word and its synonyms. Included below is the top of one of these lists for the word *trudny* 'difficult', together with similarity values (1 meaning the word itself) and an indication as to whether the word is found on the synonyms list. Here we have $prec_{10} = 0.6$ and $prec_{20} = 0.45$. As we did not observe any major differences in the distribution of good answers between models, we have only included the results counted for the entire fifty-element lists.

trudny (difficult):	niełatwy (rather difficult)	0.8114 - FOUND
	ważny (important)	0.7519
	łatwy (easy)	0.7364
	ciężki (hard)	0.7175 - FOUND
	najtrudniejszy (the hardest)	0.7081
	bolesny (grievous,painful)	0.6787 - FOUND
	kosztowny (costly)	0.6719
	niebezpieczny (dangerous)	0.6719 - FOUND
	niewygodny (awkward)	0.6671 - FOUND
	kłopotliwy (inconvenient)	0.6630 - FOUND
	ryzykowny (risky)	0.6600 - FOUND
	ciekawy (interesting)	0.6508
	udany (successful)	0.6500
	stresujący (stressful)	0.6499
	interesujący (interesting)	0.6362
	poważny (serious)	0.6324 - FOUND
	dobry (good)	0.6293
	pożyteczny (useful)	0.6227
	męczący (trying)	0.6224 - FOUND
	skomplikowany (complicated)	0.6179 - FOUND

Figures 1 and 2 show the performance of all the models trained on the three corpora for form- and lemma-based models respectively. (Models trained on restricted POS data are not shown here in the interests of greater readability.) The data shows how many words from the synonym sets from Table 3 are found within the first 50 most similar words of 50 elements checked (so, the maximum value could be 2500 if every word had 50 synonyms and all of them were placed at the top 50 position of the similarity lists). For N2 and A2 sets, this could possibly be 100% of the appropriate set (as, in the second set, the number of synonyms are usually much lower than 50). The differences between results for various parts of speech are not clear, but different models are the most efficient for a particular word category. Due to the different sizes of test sets, the figures illustrate the changes in a models efficiency for every test set separately but it cannot be used directly to compare the performance for different test sets.

Table 5: Precision of selected models for all test sets counted either for 2500 elements or for the test set size (for N2 and A2)

	N1	N2	A1	A2	V1		N1	N2	A1	A2	V1
NWla-3-c-n50	**0.23**	0.11	**0.23**	0.11	**0.21**	NWfa-3-s-n50	**0.25**	**0.32**	**0.30**	**0.33**	**0,19**
NWla-3-c-n	0.22	0.20	0.22	**0.19**	**0.21**	NWfa-3-s-n	0.16	0.18	0.18	0.14	0.18
NWla-1-c-n	0.21	**0.23**	0.19	**0.18**	0.17	NWfa-1-c-n	0.16	0.13	0.18	0.14	0.18
NWla-3-s-n	0.07	0.08	0.09	0.11	0.14	pl-emb-s	0.23	0.27	0.17	0.18	0.12
Wla-3-c-h	**0.25**	0.20	0.16	0.12	0.15	Wfa-3-c-n30	0.12	0.31	0.17	0.14	0.13

Table 5 contains the overall precision for the selected models. We have not reported recall, as the results are highly influenced by the fact that the N1, A1 and V1 test sets are much larger than the given threshold (50 top most similar words). The best achieved precision was 0,33 for

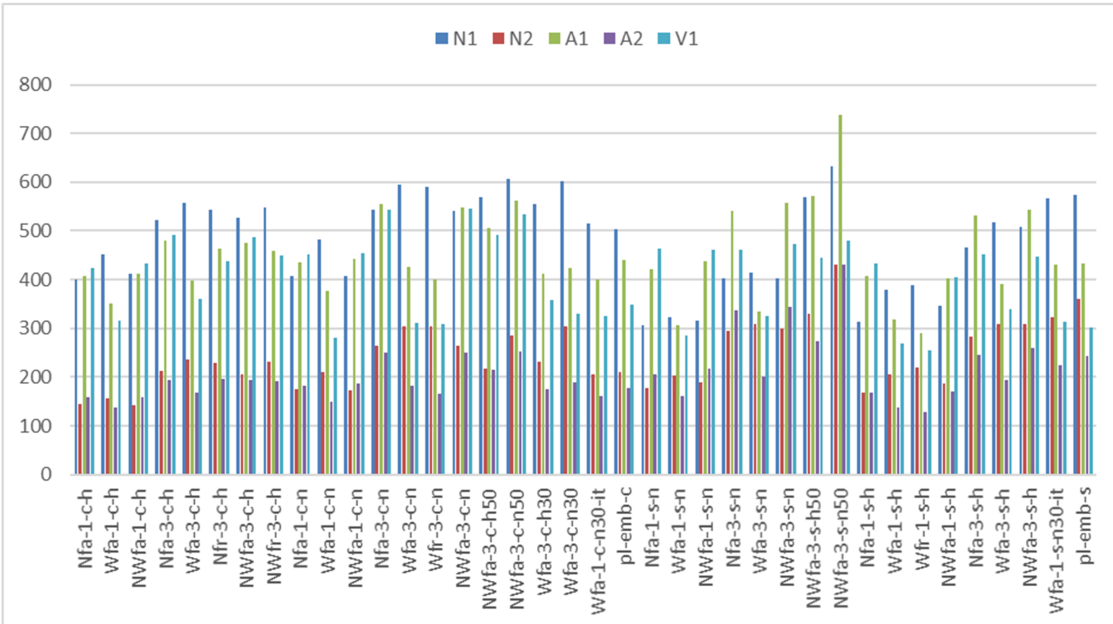

Figure 1: The number of retrieved synonyms in 2500 list consisting of the 50 first most similar words for all 50 test words; form-based models

the NWfa-3-s-n50 model on the A2 test set. The results are low, partially due to the fact that the lists contain words of different syntactic categories. Moreover, in cases when one word meaning is predominant, the top of the list of similar words reflects this one sense only, while the test sets contain synonyms for all, even rare, senses.

The results obtained for different models and each test set were checked using the t-student test for averages. For the lemma-based models, eliminating low frequency words did not improve the results. Model NWla-3-c-n50 was equally as good as NWla-3-c-n for N1, A1 and V1 but worse for N2 and A2. The skip-gram approach produces significantly worse results than CBOW — the NWla-3-s-n model is significantly worse than NWla-3-c-n. The smaller number of features (NWla-3-c-n) worsens the results for A1 and V1 only. The Wla-3-c-h model is equally good for N1 as the best models which are based on much more data.

For form-based models, the NWfa-3-s-n50 model was statistically significantly better than other form-based models for the N2, A1 and A2 sets. For N1, the results of the pl-emb-s model are statistically equally as good as for NWfa-3-s-n50. For V1, both NWfa-3-s-n and NWfa-1-c-n models give similar results. For verbs, the pl-emb-s and Wfa-3-c-n30 models are worse than others (a statistically significant difference). The Wfa-3-c-n30 model is equally as good for N2 as the best models, but for the other sets it produces significantly worse results.

The results obtained from the word form-based models are more uniform across various model parameters than those for the lemma-based models. In the latter case, CBOW has a clear advantage over the skip-gram approach. The results for CBOW models are twice as good.

Figure 3 shows the results for nouns, adjectives and verbs for all models separately. The left column of the diagram shows word form models, the right column — lemma-based ones. For lemma-based models, the advantage of the CBOW approach is visible for all test sets. Using skip-gram with negative sampling is the worst strategy here, and there is only a slight improvement for models based on 300 features for nouns. For adjectives and verbs, there is slightly more improvement. For form-based models, the choice of learning strategy is not very important. For all but verbs, the skip-gram model based on a large corpus, with low frequency words eliminated, is

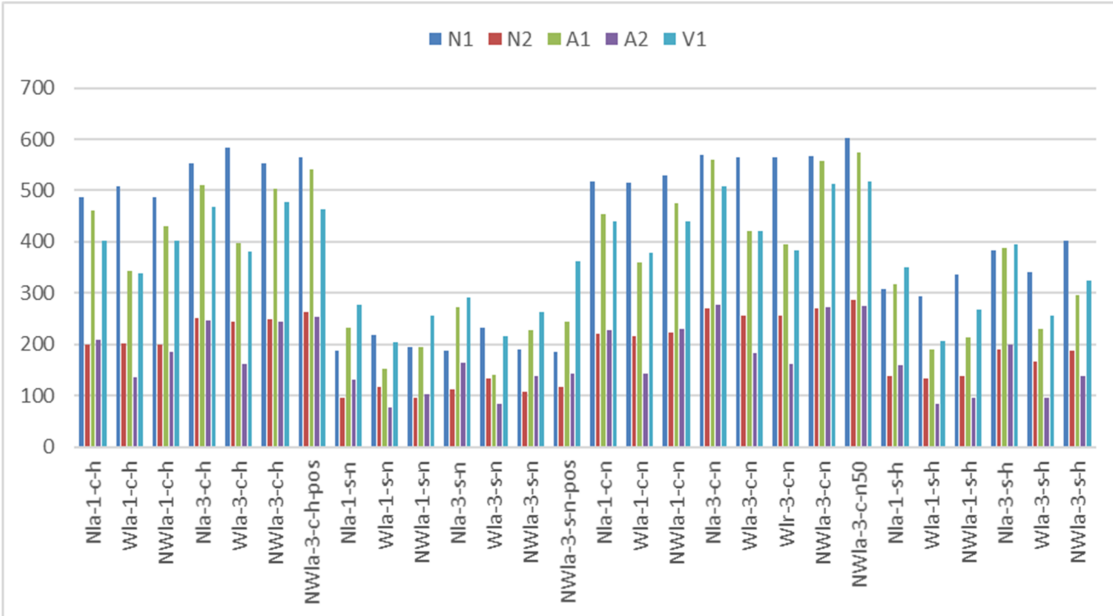

Figure 2: The number of retrieved synonyms in 2500 list consisting of the 50 first most similar words for all 50 test words; lemma-based models

the best choice and sometimes its results are far superior to the others. *Pl-emb* models are good for smaller sets of noun and adjective synonyms (N2 and A2), while they are only slightly better than word2vec models for larger sets and perform at the same level for verbs. Eliminating words with a count lower than 50 did not change the results significantly for lemma-based CBOW with negative sampling models, but was very important for most form-based models of this type (green lines on Figure 3). Using only selected POS categories to train a model (models with 'r') worsens the results for lemmas but improves them (very slightly) for forms.

5.2 Analogy

The same models were used to search for analogies. Figure 4 illustrates the overall performance of all the models in this task. It shows the number of correctly recognised analogies (the whole set described in Table 4) and compares the models based on lemmas (blue) and forms (red). It shows the higher efficiency of lemma-based models which are systematically better than the respective form-based models. The best results were obtained from the models based on NKJP and Wiki, but the models based solely on NKJP are similarly effective for some parameters. The *Nlr-3-c-n* model gives the best results on the tested set of analogies, but differences in the results of several other good models are within the margin of error. Generally, models based solely on Wikipedia are less effective than others, but both *pl-emb* models are even better than some models based on the large amount of NKJP data.

Figure 5 illustrates a more detailed comparison of models based on word forms from Wikipedia, highlighting the differences between *pl-emb* vectors and those created by us with the word2vec tool. *Pl-emb* vectors are better than most of our models with the same (100) feature vector length, and even those with 300 features. It shows that either there were some unreported interesting changes in the learning algorithm, or that data pre-processing can significantly influence the results. This would mean that it is essential to clean data before training a model. However, restricting data to certain part of speech categories (see Sec. 2) did not prove itself valuable as it did not improve the results. The effects are usually comparable to those obtained from all data. Interestingly, for

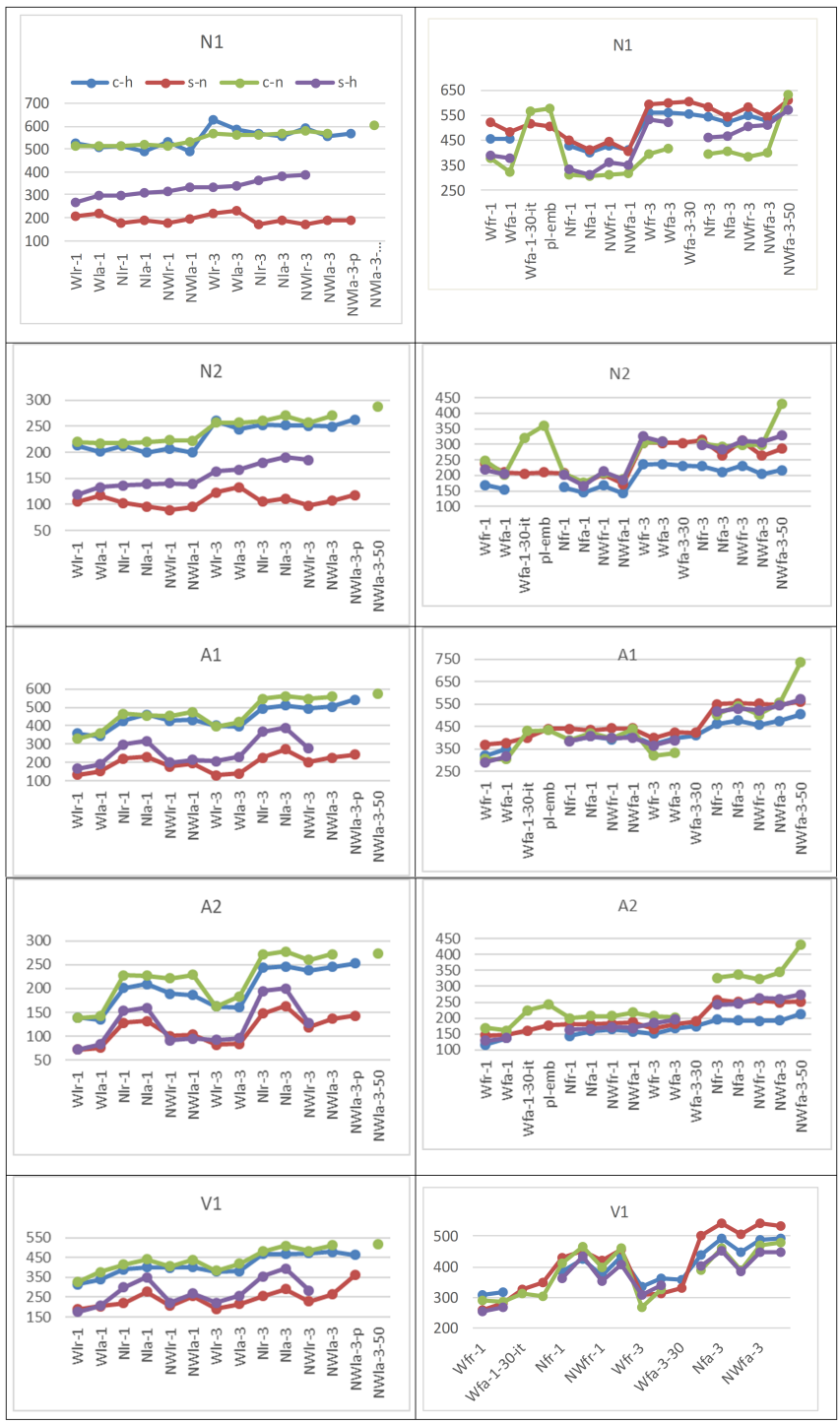

Figure 3: Synonymy results for lemma (left) and form (right) based models. The order of models reflects the increasing size of a corpus. First data are given for 100 features, then for 300 features. Four combinations of cbow and skip-gram approaches with hierarchical softmax and negative sampling were tested. The last models use either data annotation or elimination

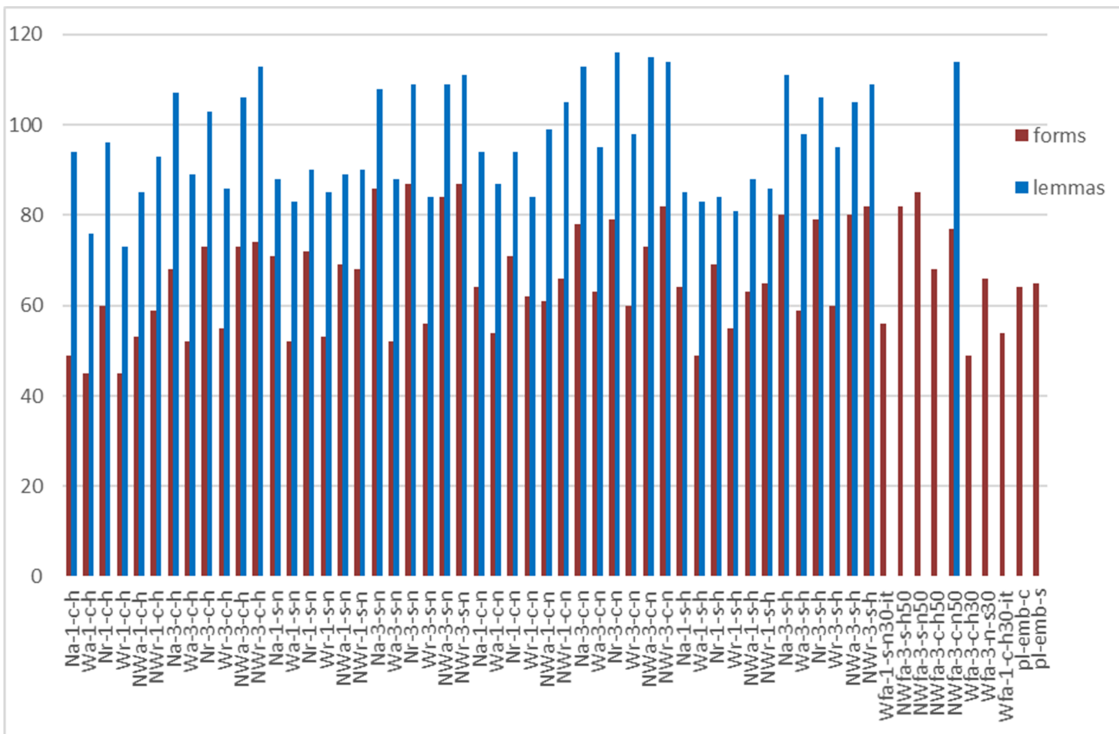

Figure 4: Overall performance for all models, both built on lemmas and word forms, for the analogy recognition task. Number of correct answers for 200 examples

form-based models, limiting word types gave slightly better results, while for lemma-based models they produced slightly worse results. The difference is most significant for verbs, which is probably because some verb forms were treated by the tagger as consisting of more than one token. Our best model was trained on the subset of Wikipedia tokens which occur more than 30 times using the skip-gram approach with negative sampling and 300 features and obtained results similar to the *pl-emb* models. This shows that the filtering of low occurrences can play a similar role to careful pre-processing. The number of learning iterations does not have a uniform impact on the results. In the case of CBOW with hierarchical softmax, increasing the amount of iteration to 100 for a 100-feature model resulted in a model which is better than an analogical model with 300 features and 10 iterations. This improvement did not occur in the skip-gram approach with negative sampling, so there is no clear answer as to whether increasing the number of iterations is justified.

Figure 6 shows a comparison of the lemma-based models grouped according to the learning technique used. Within each group, the models differ in the training data set. There are three corpora: Wikipedia; NKJP; and Wikipedia+NKJP, all in two variants: as a full set (a) or restricted (r) to selected POS. Models with a greater number of features (300) perform better than those with 100 features and the best learning combination for lemmas is CBOW with negative sampling. This model for Wikipedia is equally as good as a model for the much larger NKJP corpus. However, in this particular case data restriction worsens the results substantially. Another interesting observation is that adding Wikipedia data to the NKJP corpus does not significantly change the results. This would suggest that Wikipedia data does not contribute much to NKJP in the case of this task, i.e. all the information we look for is encoded in NKJP already. However, it is not clear why the results for the larger set are sometimes worse.

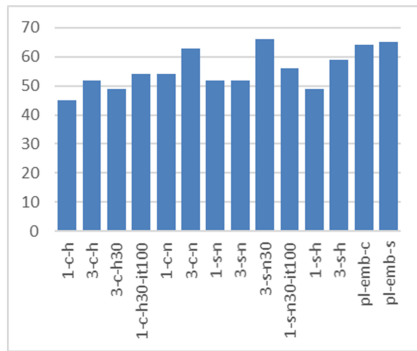

Figure 5: Overall performance of models based on Wikipedia word forms. Number of correct answers for 200 pairs

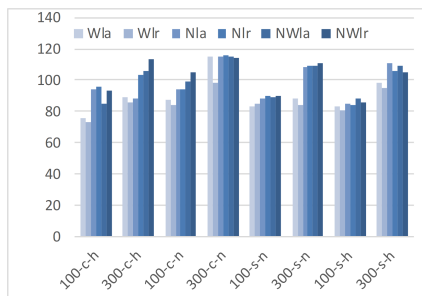

Figure 6: Comparison of lemma-based models trained on different corpora. Number of correct answers for 200 pairs

The conclusion formulated in the paper Mikolov, Sutskever, et al. (2013) that negative sampling outperforms hierarchical softmax for this task was confirmed by our results. The difference between CBOW and skip-gram is much less clear. For lemmas, models using CBOW typically give better results, while for forms, better results are obtained with the skip-gram approach.

Figure 7 shows the results for 20 pairs representing grammatical relations. The set is small as we decided that recognition of grammatical relations is not essential to the DS models. The overall results are good, with the best results (19 good answers for 20 questions) obtained for 300 feature vectors trained on NKJP only, or NKJP together with the Wikipedia data. The conclusions are similar to those formulated above for the general relations.

Finally, Figure 8 shows how well analogies are recognised for various groups of relations (given in Table 4). We have shown results obtained from the best model, i.e.: Nlr-3-c-n. It is difficult to draw reliable conclusions from this data, as the groups are rather small, but it is clear that some types of relations are easier to recognize than others. Good results are obtained for geographic relations (apart from river-town pairs) and gender (family, animal and profession) relations. It is interesting that there are no results for analogies representing the substance from which an item is made, or a cultural event and its type. This problem needs further investigation before we can formulate the conditions under which analogies are correctly recognized.

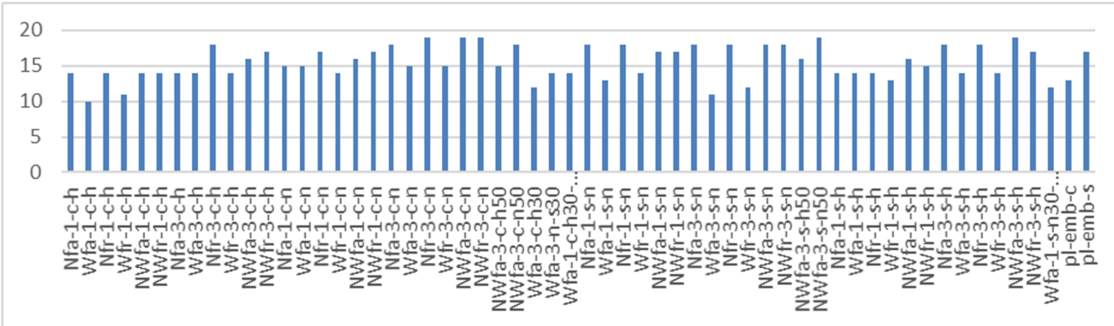

Figure 7: Performance of all word-form models on the set of grammatically related inflected forms

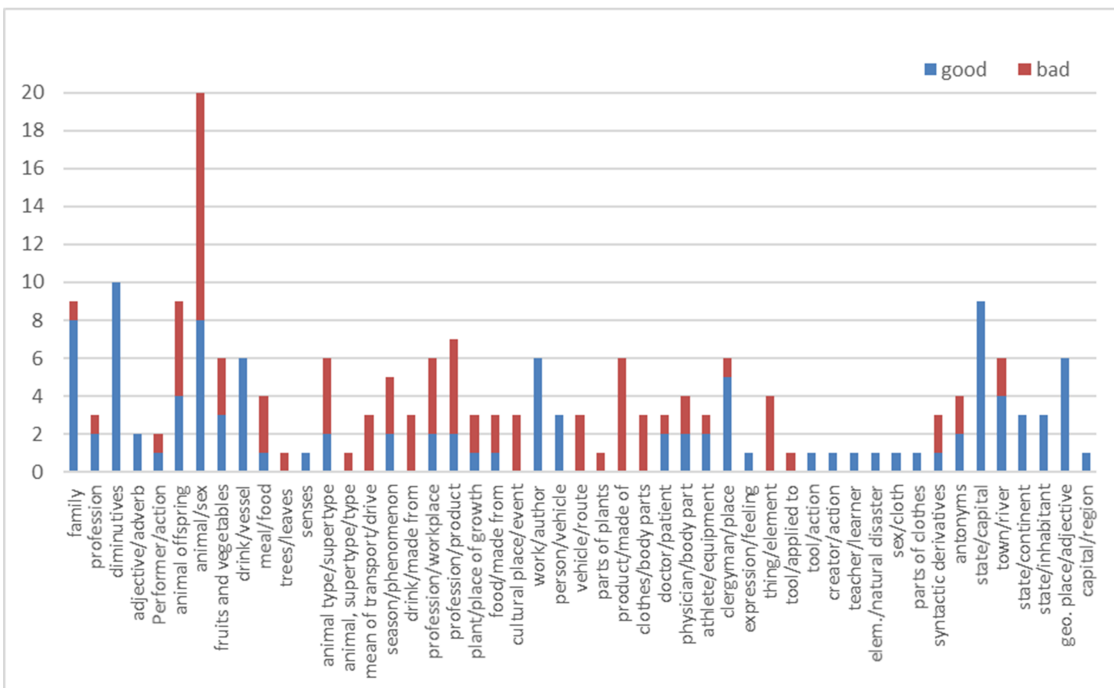

Figure 8: Recognition of different types of analogies, for each group from Table 4 the number of correct and incorrect answers are given

6 Conclusions

The aim of this paper was to test models of Polish words created with the word2vec tool with various parameters for two specific tasks: synonymy and analogy identification. As Polish is inflectional, we tested models of both lemmas and forms. The results show that word embeddings can be used to identify similarity and certain kinds of analogies for Polish words, and that the efficiency of the method is highly dependent on the chosen corpus and the model parameter values. The distributional models based on lemmas are better for analogy, while for synonymy, word forms produce better results. Moreover, it is not possible to identify one reliable, universal approach to vector creation. The CBOW approach gives better results for analogy, while skip-grams are better for synonymy. An increasing number of features, or even corpora size, do not always yield better

results. It is more important to clear the data, either by careful editing or by imposing occurrence limits. This is confirmed by the comparison of the Wikipedia models (Fig. 5) and the better results of our models for corpora with deleted infrequent words/forms. Identifying which types of analogy relation are better recognised by the method than others needs further investigation. The results obtained for synonymy are only an indication of which model gives the greater number of similar words among the closest vectors. The complete solution to the problem of synonymy identification would require both additional filtering means and a precise evaluation of the results.

References

Baroni, M., Dinu, G., & Kruszewski, G. (2014). Don't count, predict!: A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of ACL 2014 (52nd Annual Meeting of the Association for Computational Linguistics)* (pp. 238–247). East Stroudsburg, PA: Association for Computational Linguistics.

Baroni, M., & Lenci, A. (2010). A general framework for corpus-based semantics. *Computational Linguistics*, *36*(4), 673–721. https://doi.org/10.1162/coli_a_00016

Baroni, M., & Lenci, A. (2011). How we BLESSed Distributional Semantic Evaluation. In *Proceedings of the GEMS 2011 Workshop on Geometrical Models of Natural Language Semantics* (pp. 1–10). Edinburgh: Association for Computational Linguistics.

Basile, P., Caputo, A., & Semeraro, G. (2014). An enhanced Lesk word sense disambiguation algorithm through a distributional semantic model. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics*. Dublin, Irleand: Association for Computational Linguistics.

Bellegarda, J. R. (2000). Large vocabulary speech recognition with multispan statistical language models. *IEEE Transactions on Speech and Audio Processing*, *8*(1), 76–84.

Broda, B., & Piasecki, M. (2008). SuperMatrix: A general tool for lexical semantic knowledge acquisition. In *Proceedings of the International Multiconference on Computer Science and Information Technology — 3rd International Symposium Advances in Artificial Intelligence and Applications (AAIA'08)* (pp. 345–352). https://doi.org/10.1109/IMCSIT.2008.4747263

Broda, B., & Piasecki, M. (2013). Parallel, massive processing in SuperMatrix — a General tool for distrubutional semantic analysis. *International Journal of Data Mining, Modelling and Management*, *5*(1), 1–19. https://doi.org/10.1504/IJDMMM.2013.051924

Broniarek, W. (2010). *Gdy Ci słowa zabraknie*. Brwinów: Haroldson.

Budanitsky, A., & Hirst, G. (2006). Evaluating WordNet-based Measures of Lexical Semantic Relatedness. *Computational Linguistics*, *32*(1), 13–47. https://doi.org/10.1162/coli.2006.32.1.13

Cheung, J. C., & Penn, G. (2012). Evaluating distributional models of semantics for syntactically invariant inference. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 33–45). Avignon: Association for Computational Linguistic.

Church, K. W., & Hanks, P. (1989). Word association norms, mutual information, and lexicography. In *ACL'89 Proceedings of the 27th annual meeting on Association for Computational Linguistics* (pp. 76–83). Stroudsburg, PA: Association for Computational Linguistics. https://doi.org/10.3115/981623.981633

Clark, S. (2015). Vector Space Models of Lexical Meaning. In S. Lappin & C. Fox, *Handbook of contemporary semantics* (2nd ed.). Willey-Blackwell. https://doi.org/10.1002/9781118882139.ch16

Coccaro, N., & Jurafsky, D. (1998). Towards better integration of semantic predictors in statistical language modeling. In *Proceedings of ICSLP-98* (Vol. 6, pp. 2403–2406).

Dinu, G., & Baroni, M. (2014). How to make words with vectors: Phrase generation in distributional semantics. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: Vol. 1. Long papers* (pp. 624–633). Association for Computational Linguistics. https://doi.org/10.3115/v1/P14-1059

Duyu, T., Wei, F., Yang, N., Ming, Z., Ting, L., & Bing, Q. (2014). Learning sentiment-specific word embedding for twitter sentiment classification. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: Vol. 1. Long papers* (pp. 1555–1565). Association for Computational Linguistics.

Faruqui, M., Tsvetkov, Y., & Rastogi, P. (2016). Problems with evaluation of word embeddings using word similarity tasks. In *Proceedings of the 1st Workshop on Evaluating Vector Space Representati-*

ons for NLP (pp. 30–35). Associacion of Computational Linguistics. https://doi.org/10.18653/v1/W16-2506

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological Review*, *114*(2), 211–244. https://doi.org/10.1037/0033-295X.114.2.211

Harris, Z. S. (1954). Distributional structure. *Word*, *10*(23), 146–162. https://doi.org/10.1080/00437956.1954.11659520

Jastrzebski, S., Leśniak, D., & Czarnecki, W. M. (2017). *How to evaluate word embeddings?: On importance of data efficiency and simple supervised tasks*. Retrieved 23 July 2017, from https://arxiv.org/pdf/1702.02170

Kędzia, P., Czachor, G., Piasecki, M., & Kocoń, J. (2016). *Vector representations of Polish words (Word2Vec method)*. CLARIN-PL digital repository. http://hdl.handle.net/11321/327

Kim, H. K., Kim, H., & Cho, S. (2017). Bag-of-concepts: Comprehending document representation through clustering words in distributed representation. *Neurocomputing*, *266*, 336–352. https://doi.org/10.1016/j.neucom.2017.05.046

Kovatchev, V., Salamo, M., & Marti, M. (2016). Comparing Distributional Semantics Models for identifying groups of semantically related words. *Procesamiento del Lenguaje Natural*, *2016*(57), 109–116.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*(2), 211–240. https://doi.org/10.1037/0033-295X.104.2.211

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, *25*(2–3), 259–284. https://doi.org/10.1080/01638539809545028

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers*, *28*(2), 203–208. https://doi.org/10.3758/BF03204766

McDonald, S. (2000). *Environmental determinants of lexical processing effort* (Unpublished doctoral dissertation). University of Edinburgh.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, *2013*, 3111–3119. https://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf

Mikolov, T., Yih, W., & Zweig, G. (2013). Linguistic regularities in continuous space word representations. In *Proceedings of NAACL* (pp. 746–751). Atlanta, GA.

Palmer, F. R. (Ed.). (1968). *Selected papers of J. R. Firth 1952–1959*. London: Longman. (Reprinted from *A synopsis of linguistic theory 1930–1955: Studies in linguistic analysis*, pp. 1–32, by J. R. Firth, 1957, Oxford: Philological Society).

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing* (pp. 1532–1543). Association for Computational Linguistics. https://doi.org/10.3115/v1/D14-1162

Przepiórkowski, A., Bańko, M., Górski, R. L., & Lewandowska-Tomaszczyk, B. (Eds.). (2012). *Narodowy Korpus Języka Polskiego*. Warszawa: Wydawnictwo Naukowe PWN.

Řehůřek, R., & Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks* (pp. 45–50). Valetta, Malta: ELRA.

Rogalski, M., & Szczepaniak, P. S. (2016). Word embeddings for the Polish language. In L. Rutkowski, M. Korytkowski, R. Scherer, R. Tadeusiewicz, L. Zadeh, & J. Zurada (Eds.), *Artificial Intelligence and Soft Computing, ICAISC 2016: Part I. LNAI 9692* (pp. 126–135). https://doi.org/10.1007/978-3-319-39378-0_12

Sager, J. C. (1990). *A practical course in terminology processing*. Amsterdam: John Benjamins. https://doi.org/10.1075/z.44

Scheible, S., Schulte im Walde, S., & Springorum, S. (2013). Uncovering distributional differences between synonyms and antonyms in a word space model. In *International Joint Conference on Natural Language Processing* (pp. 489–497). Ngoya, Japan.

Schutze, H. (1998). Automatic word sense discrimination. *Computational Linguistics*, *24*(1), 97–124.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*(3), 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

Shaw, E. (2015). *An SAT® Validity Primer*. College Board Research. http://research.collegeboard.org/sites/default/files/publications/2015/2/research-report-sat-validity-primer.pdf

Shutova, E., Sun, L., Gutierrez, D., Lichtenstein, P., & Narayanan, S. (2017). Multilingual metaphor processing: Experiments with semi-supervised and unsupervised learning. *Computational Linguistics*, *43*(1), 71–123. https://doi.org/10.1162/COLI_a_00275

Spark Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, *28*(1), 11–21. https://doi.org/10.1108/eb026526

Stokowiec, W. (2015). *word2vec dla Polskiego Internetu*. Retrieved 19 August 2017, from http://doczz.pl/doc/562319/word2vec-dla-polskiego-internetu

Tatjewski, M., Bańko, M., Kucińska, A., & Rączaszek-Leonardi, J. (2017). Computational distributional semantics and free associations: A comparison of two word-similarity models in a study of synonyms and lexical variants. In P. P. Waliński, *Language, corpora and cognition*. Frankfurt am Main: Peter Lang.

Tsvetkov, Y., Faruqui, M., Ling, W., Lample, G., & Dyer, C. (2015). Evaluation of Word Vector Representations by Subspace Alignment. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* (pp. 2049–2054). https://doi.org/10.18653/v1/D15-1243

Turney, P. D. (2001). Mining the Web for synonyms: PMI-IR versus LSA on TOEFL. In *Proceedings of the Twelfth European Conference on Machine Learning* (pp. 491-502). Berlin: Springer-Verlag.

Turney, P., & Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, *2010*(37), 141–188.

Waszczuk, J. (2012). Harnessing the CRF complexity with domain-specific constraints: The case of morphosyntactic tagging of a highly inflected language. In *Proceedings of COLLING 2012* (pp. 2789–2804). Mumbai, India.

Weeds, J., Clark, D., Reffin, J., Weir, D., & Bill, K. (2014). Learning to distinguish hypernyms and co-hyponyms. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers* (pp. 2249–2259). Dublin: Dublin City University and Association for Computational Linguistics.

Wittgenstein, L. (1953). *Philosophical investigations*. Oxford: Basil Blackwell.

Woliński, M. (2014). Morfeusz reloaded. In N. Calzorali, K. Chourkri, T. Declerk, H. Loftsson, B. M. Maegaard, J. Mariani, A. Moreno, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the Ninth International Conference on Language Resources and Evaluation*. Reykjavík: ELRA.

Acknowledgment

The paper is supported by the Polish National Science Centre project 2014/15/B/ST6/05186, Compositional distributional semnatics methods for discrimanation and disambiguation word and simple phrase senses in Polish texts. The test sets, selected models and results will be available at zil.ipipan.waw.pl/CoDeS.

The authors declare that they have no competing interests.

The authors' contribution was as follows: concept of the study: Agnieszka Mykowiecka, Małgorzata Marciniak; data analyses: Agnieszka Mykowiecka, Małgorzata Marciniak, Piotr Rychlik; the writing: Agnieszka Mykowiecka, Małgorzata Marciniak and Piotr Rychlik.

Publisher: Institute of Slavic Studies, Polish Academy of Sciences, University of Silesia & The Slavic Foundation