Discriminating Homonymy from Polysemy in Wordnets: English, Spanish and Polish Nouns

Arkadiusz Janz, Marek Maziarz
Wrocław University of Science and Technology, Poland
{arkadiusz.janz|marek.maziarz}@pwr.edu.pl

Abstract

We propose a novel method of homonymy-polysemy discrimination for three Indo-European Languages (English, Spanish and Polish). Support vector machines and LASSO logistic regression were successfully used in this task, outperforming baselines. The feature set utilised lemma properties, gloss similarities, graph distances and polysemy patterns. The proposed ML models performed equally well for English and the other two languages (constituting testing data sets). The algorithms not only ruled out most cases of homonymy but also were efficacious in distinguishing between closer and indirect semantic relatedness.

1 Introduction

Lexical polysemy is the word property of being a signifier for different but semantically related senses. Homonymy, on the other hand, is the accidental identity of word-forms, with no traces of real semantic relatedness. Homonyms have different etymologies, while *polysemes* are the product of the sense extending diachronic processes (Lyons, 1995, pp. 54-60). In fact, homonymous words – as semantically unrelated – should be treated as separate words. For Natural Language Processing this task is completely valid since homonyms frequently appear in textual corpora. Wordnets suffer from the absence of explicit links between related meanings and do not distinguish between the two types of ambiguity, making the task harder (Freihat et al., 2013b; Mihalcea, 2003).

In this paper we present a machine learning approach to automatic discrimination of homonymy from polysemy in three languages: English, Spanish and Polish. We randomly drew samples of noun polysemous lemmas from each wordnet and generated all possible sense couplings. Then we cross-checked them in traditional dictionaries in search of their homonymy/polysemy status (Section 3.1). Each pair was annotated with four different groups of features, representing: lemma properties (Sec. 3.2.1), semantic similarities between glosses (3.2.2), graph properties of nodes (3.2.3) and polysemy patterns (3.2.4). Having trained ML models on English data, we checked their efficacy on Spanish and Polish sense pairs (Sec. 4.1). Then, we passed to the analysis of each model behavior on the subset of English words with known sense distances (we transformed macro- and microstructures of Oxford Lexico and Merriam-Webster Dictionary into graphs, Sec. 4.2). We also introduced a definitional guidelines for distinction between *close* and *indirect* polysemy relationship. At the end, we manually inspected 300 sense pairs to assess how well homonymy-polysemy discrimination served close polysemy recognition (Sec. 6).

We define *homonyms* or *homographs* as etymologically unrelated sets of senses, having the same part of speech (POS) category and signified by the same lemma. We abstract from other grammatical properties of nouns, such as the mass/countable noun distinction in English or gender differences in Spanish. We say that two nominal senses represent homonymy, if they share the same lemma and whose dictionary equivalents are noted under distinct entries (i.e., in disjoint entry microstructures).

2 Related Work

Polysemy and homonymy attracted huge researchers' attention. Approaches to dissolve the problem could be divided roughly into three main camps: (i) regular polysemy pattern recognition, (ii) polysemy instance recognition and (iii) ontology-based discrimination. Our method belongs to the second group.

(i) Numerous papers were devoted to recognising regular polysemy types (patterns), i.e. classes of polysemy instances (actual sense pairs), sharing the same two superordinate semantic categories (Apresjan, 1974). Those pairs of categories include *animal – food*, *container – content*, *institution – building* etc. To computational linguistics this approach was introduced first by Buitelaar (Buitelaar, 2000, 1998). Wordnets were searched for polysemy patterns since then by many scientific teams. Peters and Peters (2000) and Peters et al. (1998) tested WordNet unique beginners, as well as, different combinations of indirect hypernyms as representatives of semantic domains ("conceptual signposts") for English, Spanish, and Dutch, see also (Peters, 2003). Freihat et al. (2016), Freihat et al. (2013b) and Freihat et al. (2013a) identified polysemy patterns and homonymy through the automatic analysis of WordNet taxonomy and logical-like inferences. They searched for parental semantic categories below the upper levels of WordNet hierarchy. Precisions as high as 90% were reported by different research groups. Since the automatic assessment of recall was impossible in this methodology (cf. Barque and Chaumartin (2009)), instead, the coverage ratio for wordnets was often given.

(ii) Regular polysemy does not exhaust the possible polysemous link types, since polysemous senses might be related irregularly, according to metonymy or metaphor paths specific to one or very few pairs. This topic is of high interest for Word Sense Disambiguation, because finding precise semantic links between senses may lead directly to sense merging and – the so called – polysemy reduction (Palmer et al., 2007; Navigli, 2009; Mihalcea, 2003). Some general kinds of polysemy are being distinguished, like metaphor, metonymy or specialisation/generalisation (Peters et al., 1998). Barque and Chaumartin (2009) and Peters (2006) constructed rules and imposed keyword constrains on glosses. Veale (2004) investigated the broader range of possible rules relying not only on glosses but also on local graph topological properties. New models/algorithms may be used to adding new instances of polysemy to WordNet, cf. for instance a metonymy enrichment in (Hayes et al., 2004).

(iii) Instead of investigating sense pair status, Utt and Padó (2011) carried out the division at the *lemma* level. They made use of polysemy patterns, based on basic types of Buitelaar's CoreLex, and looked for *n* most frequent polysemy types. They found the fact that polysemous words tended to have more frequent patterns than homonyms useful in homonymy-polysemy discrimination.

3 Method

3.1 Resources and Samples

In discriminating homonymy from polysemy we relied on traditional dictionaries. For English they were Oxford *Lexico*[1] and American English *Merriam-Webster Dictionary*[2]. For Spanish we utilised *Diccionario de la lengua española* of Real Academia Española[3] and Lexico *Spanish Dictionary* of Oxford University Press[4]. Polish dictionaries included *Uniwersalny słownik języka polskiego*[5], Doroszewski's *Słownik języka polskiego*[6] and *Słownik języka polskiego*[7], all published by Wydawnictwo Naukowe PWN.

We used Open Multilingual WordNet (version 1, Bond and Paik (2012); Bond and Foster (2013)) as the source of polysemous lemmas. We focused on English WordNet (Fellbaum, 1998), Spanish part of the Multilingual Central Repository (Atserias et al., 2004; Gonzalez-Agirre et al., 2012) and Polish WordNet (Maziarz et al., 2016). For each wordnet we randomly sampled a set of polysemous noun lemmas. Then each lemma was checked in the dictionaries in order to find whether it was homonymous or not. If so, we carefully checked all couplings of lemma senses and decided their homonymy/polysemy status. For lemmas that were considered polysemous we automatically assumed polysemy of their sense pairs. Then we added a couple dozen potentially homonymous nouns to increase the number of homonymy cases.[8] We searched the potential homonyms in the literature on homonymy. These new nouns were then cross-checked with dictionaries, pair by pair. In

[1] https://www.lexico.com/

[2] https://www.merriam-webster.com/

[3] https://dle.rae.es/

[4] https://www.lexico.com/es/

[5] https://usjp.pwn.pl/

[6] http://doroszewski.pwn.pl/

[7] https://sjp.pwn.pl/

[8] Having added new homonymy cases to our data sets, we must have distorted the real proportion between homonymy and polysemy. This choice affected the subsequent measurements of precision and recall. Consequently, the calculated homonymy recognition precision will be treated as the upper bound for the real homonymy precision, while the obtained polysemy precision will be regarded as the lower bound for the real polysemy precision. Random sampling enables us to directly assess recall of ML models, which is an obvious advantage.

the case of English we simply borrowed 25 English homonymous lemmas from a previously made resource,[9] see Sec. 4.2 for details. Table 1 presents statistics of final data sets. As could be seen the data set was unbalanced.

wordnet		sample		# sense pairs		
lang	#nS	#L	#S/L	H	P	$\sum$
eng	82k	159	4.1	325	1,241	1,566
spa	26k	135	4.0	87	1,060	1,147
pol	29k	111	2.5	39	232	271

Table 1: English, Spanish and Polish polysemous wordnet nouns with annotated homonymy cases. Symbol: #nS – number of noun synsets in a wordnet, #L – number of lemmas in a sample, #S/L – an average number of senses per lemma.

3.2 Features

We used a set of 19 features, representing different properties of sense pairs (Fig. 1). We started from Open Multilingual Wordnet and its language-dependent lemma-synset pairings. Having obtained the set of – let's say – n noun senses, we generated all possible $\frac{n\times(n-1)}{2}$ combinatorial pairs of senses. We treated PWN network structure, synset glosses, synset semantic domains etc. as means of meaning description. English served as the metalanguage for language specific lemma-sense pairs, not only in the case of Spanish and Polish, but also in the case of English itself. Thus English language was used in a two-fold way: as a semantic metalanguage (via PWN), and also as the object of semantic description (via OMW). Thanks to such an approach, our analysis and developed statistical models hopefully could be applicable to virtually any OMW language.

The features that we used could be roughly divided into four main groups: (a) lemma properties (standardised to obtain language independent measures), (b) gloss similarities, (c) graph measures and (d) polysemy patterns. We give them a sharp description below.

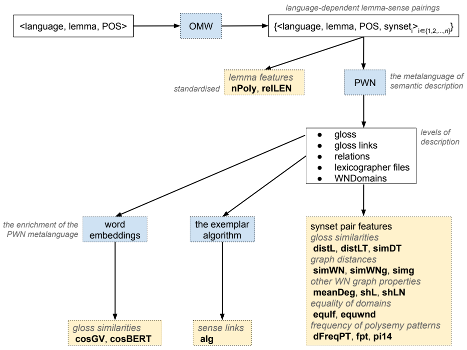

Figure 1: Feature calculation stages.

3.2.1 Lemma properties

nPoly – is a standardised number of senses of a given lemma l:

$$nPoly(l) := \frac{nsen(l) - m}{s}, \tag{1}$$

where $nsen(l)$ is the number of lemma l senses, m – a mean lemma sense number in a language and s – a standard deviation of lemma sense number in the language.

relLEN – is a standardised length of a given lemma l in characters, given by the formula:

$$relLEN(l) := \frac{nchar(l) - m}{s}, \tag{2}$$

where $nchar(l)$ is the lemma l length in characters, m – a mean lemma length in a language and s – a standard deviation of lemma length in the language.

3.2.2 Gloss similarities

distL – is a synset gloss dissimilarity measured on strings of letters through Levensthein edition distance:

$$distL(g(s_1), g(s_2)) := \frac{L(g(s_1), g(s_2))}{max(nchar(g(s_1)), nchar(g(s_2)))}, \tag{3}$$

where s_1, s_2 are synsets, $g(s)$ - is a gloss of synset s, $nchar(\cdot)$ - is a string length in characters, $L(\cdot,\cdot)$ - is a Levensthein edition distance between two strings.

distLT – is a synset gloss dissimilarity measured on sequences of tagged glosses through Levens-

[9] https://github.com/MarekMaziarz/PolysemyTheories/blob/master/LEX-MW-merged-graph-distances.txt

thein edition distance:

$$distLT(g(s_1), g(s_2)) := \frac{L(T(g(s_1)), T(g(s_2)))}{max(nchar(T(g(s_1))), nchar(T(g(s_2))))}, \tag{4}$$

where $T(\cdot)$ denotes a sequence of glosses lemmatised by the Stanford Tagger, and all other symbols defined exactly as in the definition of $distL$.

simOV – is a synset gloss similarity measured as the overlap between two sets of gloss lemmas. Let $V = (v_1, ..., v_n)$ and $W = (w_1, ..., v_m)$ be the sequences of words constituting glosses $g(s_1) = V$ and $g(s_2) = W$, respectively. Let then $ST(X)$ denotes the set of tagged words constituting the gloss sequence $X = (x_1, ..., x_n)$, i.e., $ST(X) = ST(x_1, ..., x_n) = \{T(x_1), ..., T(x_n)\}$. Thus we define $simOV$ similarity as follows:

$$simOV(g(s_1), g(s_2)) := \frac{|ST(V) \cap ST(W)|}{min(|ST(V)|, |ST(W)|)}, \tag{5}$$

where $|A|$ denotes a cardinality of the set A.

cosGV – is a cosine of two 50D vectors representing mean of GloVe vectors for all words constituting a gloss of a synset. Let $GV(w)$ be a 50D GloVe vector of the word w. We define $\overline{GV(g(s))} = \overline{GV(W)}$ as a mean vector of all words constituting the sequence W of length n, i.e.

$$\overline{GV(W)} = \frac{GV(w_1) + ... + GV(w_n)}{n}. \tag{6}$$

Then, if $V = g(s_1)$ and $W = g(s_2)$,

$$cosGV(V, W) := cos(\overline{GV(V)}, \overline{GV(W)}). \tag{7}$$

cosBERT – cosine of BERT vectors representing two glosses of paired synsets.

3.2.3 Graph properties

Six measures were based on graph properties:

simWN – was measured on the bidirectional graph of sole WordNet relations:

$$simWN(s_1, s_2) := \frac{1}{dist_{WN}(s_1, s_2)^2 + 1}. \tag{8}$$

Here $dist_{WN}(s_1, s_2)$ describes Dijkstra's distance on WordNet graph.

simWNg – was defined on WordNet graph expanded with bidirectional gloss relations:

$$simWNg(s_1, s_2) := \frac{1}{dist_{WNg}(s_1, s_2)^2 + 1}. \tag{9}$$

simg – was defined accordingly on the graph of gloss relations:

$$simg(s_1, s_2) := \frac{1}{dist_g(s_1, s_2)^2 + 1}. \tag{10}$$

meanDeg – is a mean degree of two synsets measured in bidirectional WordNet graph as follows:

$$meanDeg(s_1, s_2) := \frac{F(s_1) + F(s_2)}{2}, \tag{11}$$

where $F(s)$ is a geometric mean of a square root of the instance degree $\sqrt{deg_i(s)}$ of the synset s (total number of instance relations coming to and from the node s) and the type degree $deg_t(s)$ (total number of relation types the node s is involved within), i.e.

$$F(s) := \frac{2 \cdot deg_t(s) \cdot \sqrt{deg_i(s)}}{deg_t(s) + \sqrt{deg_i(s)}}, \tag{12}$$

shL – is a shared lemma index, i.e. the intersection of sets of lemma synsets divided by the cardinality of the smallest lemma set:

$$shL(s_1, s_2) := \frac{|lem(s_1) \cap lem(s_2)|}{min(|lem(s_1)|, |lem(s_1)|)}, \tag{13}$$

$lem(s)$ being the set of all lemmas of the synset s.

shLN – is a shared lemma neighborhood index. Let $Nb(s) = \{s_1, ..., s_m\}$ be the set of all m synsets that are one step apart from the synset s in bidirectional WordNet graph. Let Lem be a function such that

$$Lem(Nb(s)) = lem(s_1) \cup ... \cup lem(s_m). \tag{14}$$

The shLN measure is given by the formula:

$$shLN(s_1, s_2) := \frac{|Lem(Nb(s_1)) \cap Lem(Nb(s_2))|}{min(|Lem(Nb(s_1))|, |Lem(Nb(s_1))|)}. \tag{15}$$

3.2.4 Polysemy patterns

The last group of features relies on our ability to capture polysemy patterns and relations.

alg – is a binary function which checks whether a given sense pair is predicted by a sense linking algorithm, called *exemplar* algorithm (cf. Ramiro et al. (2018)). The exemplar algorithm links word senses into a polysemy net according to their proximity in WordNet+glosses graph. At each step we join a new sense that is the closest to all already linked senses. The algorithm starts from the synset with the highest vertex degree (given by the formula (12)).

equlf – is a binary function that checks equality of semantic domains as defines by lexicographer files. Let $LF(s)$ be a semantic domain of the synset s given in lexicographer files.

$$equlf(s_1, s_2) = \begin{cases} 1 & \text{if } LF(s_1) = LF(s_2) \\ 0 & \text{otherwise.} \end{cases} \tag{16}$$

equwnd – is a binary function that checks equality of semantic domains as defined by WordNet Domains.[10] Let $WND(s)$ be a semantic domain of the synset s given by WND, then

$$equlf(s_1, s_2) = \begin{cases} 1 & \text{if } WND(s_1) = WND(s_2) \\ 0 & \text{otherwise.} \end{cases} \tag{17}$$

fpt – is a binary function that checks whether a given sense pair belongs to the 5% of the most frequent polysemy patterns. Let's define a polysemy type PT as a pair of semantic domains, i.e. $PT(s_1, s_2) = (LF(s_{\tilde{1}}), LF(s_{\tilde{2}}))$ (ordered alphabetically from left to right, which we mark symbolically with a tilde mark). If we arrange PTs into a ranking list according to their frequency in WordNet (that is in the set of all possible pairs of polysemous senses) and establish the set $FreqPT$ of most frequent PTs which accounts altogether for at most 5% of PT occurrences in WordNet, then

$$fpt(s_1, s_2) = \begin{cases} 1 & \text{if } PT(s_1, s_2) \in FreqPT \\ 0 & \text{otherwise.} \end{cases} \tag{18}$$

dFreqPT – is a cumulative distribution of a given polysemy pattern. Let N be the total number of all polysemy pairs in a wordnet, m be the total number of polysemy patterns, i be the rank of the polysemy type PT_i and $Freq(PT_i)$ be the count of all occurrences of PT in the wordnet.

$$N = \sum_{i=1}^{m} Freq(PT_i). \tag{19}$$

Then

$$dFreqPT(s_1, s_2) = \frac{\sum_{i=1}^{p} Freq(PT_i)}{N}, \tag{20}$$

where $PT_p = PT(s_1, s_2), p \leq m$.

pi14 – is the measure inspired by the π_{81}-score from (Utt and Padó, 2011). Let $FreqPT$ be the set of most frequent polysemy patterns, as defined above, P_w be the set of polysemy patterns PT of a given lemma l, then

$$pi14(l) := \frac{|FreqPT \cap P_w|}{|P_w|}, \tag{21}$$

i.e., it is the ratio of language most frequent PTs in the set of PT characteristic for a given l.

4 Results

4.1 Polysemy vs. Homonymy

To cope with the unbalanced data problem a resampling methodology was applied for the smaller 'H' class.[11] Data were divided into 10 folds according to their *lemmas*.[12] Models were presented iteratively 9 training folds and then evaluation was performed on the 10^{th} fold. The final tests were performed on Spanish and Polish (broken down into 5 folds for the comparison with baselines). LR was optimised through LASSO methodology (with 1 SE optimisation).[13] SVM was run with default parameters. Tables 2, 3 and 4 present the results for English, Spanish and Polish, respectively.

The obtained results prove that our two classes are not easily separable. This is caused not only by the choice of particular features, but also by the actual nature of polysemy. As soon as we matched various sense pairs of polysemous words, we had to deal with different parts of polysemy nets. Some

[10] If there were more than one category ascribed to a synset, we manually picked the most representative.

[11] We presented a ML model each homonymy pair four times.

[12] Sense pairs representing the same lemma landed in the same fold

[13] We optimalised λ parameter of the LASSO logistic regression on each training data set, then for the whole training data set the optimalised λ (corresponding to the most parsimonious model within 1 SE from the minimal error model) was established. The number of features was reduced to 13, with the most prominent being relLEN, simWN, simWNG, cosBERT, equlf, meanDeg, nPoly, pi14 and shL.

English		prediction		efficiency	
	class	P	H	Prec.	Recall
SVM	P	716	525	$\mathbf{.94}^{*}_{*}$	$\mathbf{.58}_{*}$
	H	43	282	$\mathbf{.35}_{*}$	$\mathbf{.87}^{*}_{*}$
LR	P	661	580	$\mathbf{.97}^{*}_{*}$	.53
	H	19	306	$\mathbf{.34}_{*}$	$\mathbf{.94}^{*}_{*}$
mBL	P	1241	0	.79	$\mathbf{1}^{*}_{*}$
	H	325	0	–	0
rBL	P	620.5	620.5	.79	.50
	H	162.5	162.5	.21	.50

Table 2: Confusion matrices for English test set, 10-fold cross-validation. Baselines: 'mBL' – the majority class, 'rBL' – random (uniform distribution). Results significant at 5% significance level are marked with an asterisk (Holm's correction for multiple comparisons was applied, see Holm (1979)). In superscripts we give the comparison with mBL, while in subscripts – to rBL. In the case of mBL baseline, in superscript we present comparison to SVM and in subscript – to LR.

Spanish		prediction		efficiency	
	class	P	H	Prec.	Recall
SVM	P	600	460	$\mathbf{.98}^{*}_{*}$	.57
	H	10	77	$\mathbf{.14}_{*}$	$\mathbf{.88}^{*}_{*}$
LR	P	525	535	$\mathbf{1}^{*}_{*}$	.50
	H	0	87	$\mathbf{.14}_{*}$	$\mathbf{1}^{*}_{*}$
mBL	P	1060	0	.92	$\mathbf{1}^{*}_{*}$
	H	87	0	–	0
rBL	P	530	530	.92	.50
	H	47.5	47.5	.08	.50

Table 3: Confusion matrices for Spanish test set. 5-fold cross-validation (with Benjamini-Hochberg correction, see Benjamini and Hochberg (1995)).

pairs were semantically as close as an extended sense and its base sense. Another represented distant relationships, i.e. indirect links. Although the polysemy class contained only related senses, the real nature of their semantic proximity was not determined. As a result the polysemy relationship class might have been torn apart between closer relationships and the heavy body of (resampled) homonymy class – representing the opposite poles of the polysemy-homonymy axis.[14]

Polish		prediction		efficiency	
	class	P	H	Prec.	Recall
SVM	P	149	83	$\mathbf{.96}^{*}_{*}$	$\mathbf{.64}_{*}$
	H	6	33	$\mathbf{.28}_{*}$	$\mathbf{.85}^{*}_{*}$
LR	P	116	116	$\mathbf{.96}^{*}_{*}$	.50
	H	5	34	$\mathbf{.23}_{*}$	$\mathbf{.87}^{*}_{*}$
mBL	P	232	0	.86	$\mathbf{1}^{*}_{*}$
	H	39	0	–	0
rBL	P	116	116	.86	.50
	H	19.5	19.5	.14	.50

Table 4: Confusion matrices for Polish test set. 5-fold cross-validation (Benjamini-Hochberg correction).

It seems that both SVM and logistic regression aimed at capturing as many cases of homonymy as possible, with slight predominance of the logistic regression in this task. Both models led to ruling out many semantically related pairs, thus as a result we obtained a heterogeneous class of homonymy predictions and homogeneous polysemy class. Despite these weaknesses both models easily outperformed majority baselines, as well as random ones.

4.2 Close vs. Distant Polysemy

To check how well the two models cope with close (direct) and distant (indirect) polysemy, we contrasted their outputs with external data from Lexico and Merriam-Webster Dictionary. We analysed 57 nominal lemmas.[15] Onto each noun and its senses we mapped corresponding Princeton WordNet synsets. Then we transformed dictionary microstructures into graphs – according to sense ordering and polysemy hierarchy. It enabled us to measure distances between PWN senses in both dictionary-based graphs.[16]

Figure 2 presents both prediction classes "P" and "H" projected onto the plane of semantic distances measured either in Lexico graph ("distLEX"), or in Merriam-Webster ("distMW"). Homonymy resides in the top right most corner

[14] Lyons (1977, p. 550) perceived polysemy as a non-binary relation ranging from vagueness of meaning shades to total unrelatedness of homonymy.

[15] This comprised following nouns: *angle*, *band*, *bank*, *bark*, *bat*, *board*, *can*, *chapter*, *chop*, *clip*, *concealment*, *crest*, *cylinder*, *date*, *degree*, *duck*, *fall*, *fame*, *file*, *fly*, *gloss*, *intellect*, *lump*, *master*, *match*, *palm*, *pasturage*, *plant*, *ring*, *rock*, *rose*, *saw*, *scale*, *score*, *sentence*, *shilling*, *sink*, *skimmer*, *spring*, *stage*, *stalk*, *table*, *term*, *tie*, *tongue*, *trepan*, *trip*, *tune*, *veneer*, *vermin*, *victim*, *voucher*, *well*, *whirl*, *wrapping* and *wreck*.

[16] The transformation followed two main rules: (1) link main senses into a chain according to their ordering, (2) link a subsense to its superordinate.

of the plane, while direct polysemy occupies the area close to the origin of the coordinate system, i.e. the point $(0, 0)$. Graph distances themselves are highly correlated if we include homonymy cases.[17]. Spearman's rank correlation $\rho = 0.771$. If we exclude homonymy the correlation drops to the moderate values, $\rho = 0.453$ (sole polysemy cases). Intuitively, we could define *close polysemy* as a pair of senses which are (at least in one dictionary graph):

- either adjacent nodes in the chain of ordered senses,
- or a main sense and its subsense.

More formally we would say that two senses s_i and s_j of the same word represents the relation of close polysemy (cP) if the following condition holds:

$$cP := \{(s_i, s_j) \in S \times S : dist_{LEX}(s_i, s_j) \leq 1 \lor dist_{MW}(s_i, s_j) \leq 2\}, \tag{22}$$

where $S = s_1, ..., s_n$ is the set of n senses of the same word, while $dist_{LEX}$ and $dist_{MW}$ are measured on Lexico and Merriam-Webster graphs, respectively.[18] The 'dP' class was a set-theoretic complement of the 'cP' set to the 'P' class, i.e.

$$dP := \{(s_i, s_j) \in S \times S : (s_i, s_j) \notin H \land dist_{LEX}(s_i, s_j) > 1 \land dist_{MW}(s_i, s_j) > 2\}, \tag{23}$$

where H is the set of homonymy cases.

Figure 3 and Table 5 illustrate how well the logistic classifier and the SVM model deal with the two different types of polysemy: close, 'cP', and distant, 'dP', as well as with homonymy pairs, 'H'. As could be seen, the prediction class 'H' comprises almost all homonymy cases and most cases of distant polysemy. Almost half cases of close polysemy belongs there also. When one looks at the prediction 'P' class, the reversed picture is revealed. It contains nearly no cases of homonymy, and 2 times more close polysemy pairs than distant polysemy. It seems that the prediction class 'P' approximates close polysemy (with 67% precision and 50% recall), although we did not teach models the direct recognition of this class.

English	class	prediction P	prediction H	efficiency P	efficiency R
SVM	cP	170	125	**.59**$^*_*$	.58$_*$
	dP	105	172	**.67**$^*_*$	**.68**$^*_*$
	H	11	80		
LR	cP	157	138	**.66**$^*_*$	.53
	dP	80	197	**.67**$^*_*$	**.78**$^*_*$
	H	2	89		
mBL	cP	295	0	.44	**1**$^*_*$
	dPH	368	0	–	0
rBL	cP	147.5	147.5	.44	.50
	dPH	184	184	.56	.50

Table 5: The subset of LR and SVM confusion matrices presented in Table 2 limited to Lexico and Merriam-Webster data. Three grades of semantic similarity/dissimilarity represent: close polysemy ('cP') – distant polysemy ('dP') – homonymy ('H'), as cross-tabulated with binary logistic predictions ('P', 'H'). Efficiency measures were calculated for the 'cP' class and for the joint 'dP' + 'H' class. Two baselines were calculated: 'mBL', i.e., the majority class and 'rBL' – random baseline. Benjamini-Hochberg correction was applied in the comparison with baselines on 5 random folds (5% significance was marked with asterisks).

4.3 Manual evaluation

Table 6 presents results of the independent manual evaluation by the first (#1) and the second (#2) author of this paper. #2 annotated 300 sense pairs (100 for each language), randomly selected from the outcome 'P' class of the English logistic regression model. #1 evaluated a subset of 100 of those pairs. Sense pairs were judged against their PWN definitions. Two senses were considered a close polysemy pair ('cP') if only they could be classified as one of the following polysemy subtypes: (i) metaphor, (ii) metonymy (including situation-argument relationships), (iii) sense broadening/narrowing, (iv) co-hyponymy, (v) antonymy and (vi) near-synonymy (cf. (Cruse, 2006, pp. 133-4), (Taylor, 2000, pp. 128-9)). Otherwise they were considered 'dP' (if they were semantically related) or 'H' case (if there was no relationship at all).

The resulting agreement was moderate, with Cohen's $\kappa = 0.4$ ('dP' and 'H' class were identified). 24 remaining disagreement cases were then again independently rejudged, resulting a higher

[17] Transforming infinities to maximum values for homonymy, i.e. $Inf \longrightarrow max(dist) + 1$.

[18] Since Merriam-Webster has more fine-grained sense distinctions, we used different thresholds for both dictionaries.

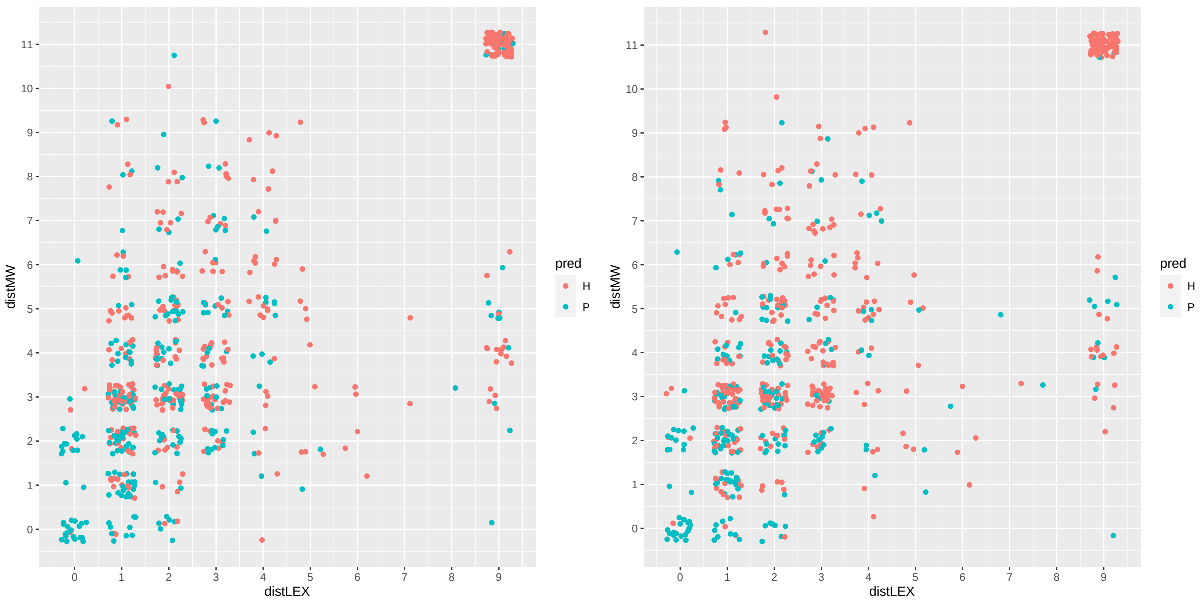

Figure 2: On the left – SVM, on the right: logistic regression. Prediction classes “P” and “H” compared with Lexico and Merriam-Webster distances (distLEX and distMW, respectively). Real homonymy cases occupy the top right corner, while direct polysemy cases take up the bottom left area. Please note, this is the subset of 10-fold cross-validation data (Table 2) limited to Lexico and Merriam-Webster lemmas.

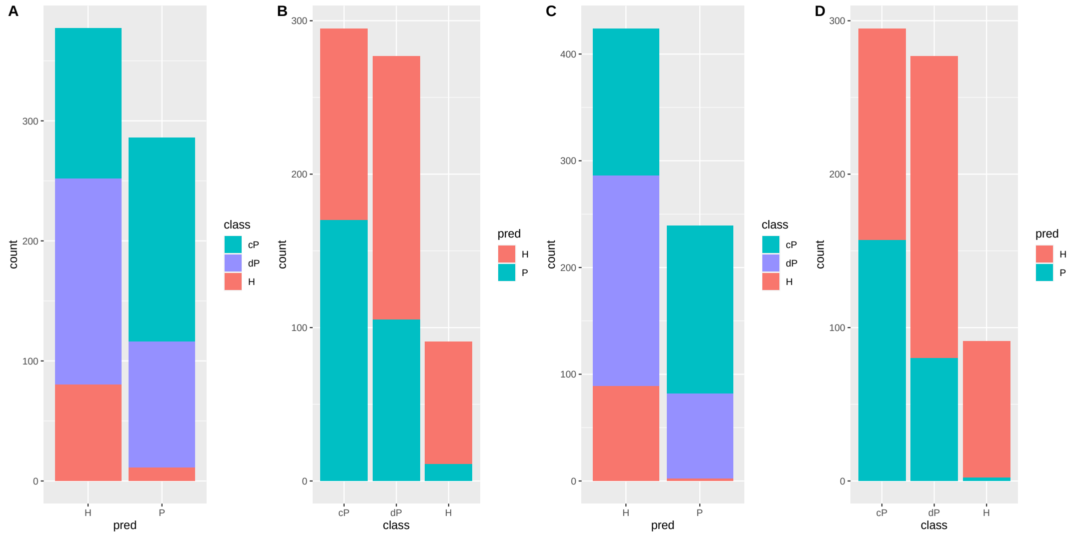

Figure 3: On the left (A, B) – SVM outcome, on the right (C and D) – LR model. Prediction classes ‘P’ and ‘H’ as compared to close and distant polysemy, ‘cP’, ‘dP’ and real homonymy ‘H’. A, C – facets by prediction classes, B, D – facets by cP, dP and real homonymy.

pred. 'P'		eng*	spa	pol	in total	
	class	n	n	n	$\sum$	CI [%]
#1	cP	24	26	28	78	69-86
	dPH	10	7	5	22	14-31
	$\sum$	34	33	33	100	100%
#2	cP	58	70	71	199	61-72
	dPH	42	30	29	101	28-39
	$\sum$	100	100	100	300	100%

Table 6: Manual evaluation of the prediction class 'P' given by the LR classifier with regard to 'cP', 'dP/H' classes. Symbols: * – cross-validation results, CI – a 95% confidence interval. The annotator #2 validated 300 cases, out of which the annotator #1 annotated 100. Cohen's $\kappa = 0.4$.

kappa, $\kappa = 0.6$, with the percentage IAA = 86%. Taking into account only the agreed 86 cases, we got CI for 'cP' equal to 68%-86%. Though the agreement was not perfect, the experiment proved that the majority of 'P' class instances was indeed close polysemy. The obtained confidence intervals are almost in perfect concordance with the automatic evaluation performed on the Lexico and Merriam-Webster graphs.

5 Conclusions

In a small-scale study of 400 nouns from three languages representing different branches of the Indo-European family we checked usefulness of two ML models (logistic regression and SVM) in discriminating homonymy from polysemy. We proposed a new set of 19 language-independent features, which comprised: lemma properties (like length), gloss similarities (including embeddings), graph properties (like graph distances) and frequent polysemy patterns. LR and SVM were trained on English data and tested on Spanish and Polish. The results were comparable, suggesting that our method could be transferred to non-congenial languages. Machine learning models performed above baselines for all languages.

Comparison with traditional dictionaries showed that trained classifiers preserved not only the polysemy-homonymy distinction, but also favoured direct polysemy over indirect relationships (in the prediction class 'P', with the reversed situation for 'H' predictions). Manual inspection of the LR 'P'-class outcome confirmed this finding: majority of sense pairs were classified as close rather than indirect semantic links.

Acknowledgments

This research was financed by the National Science Centre, Poland, grant number 2018/29/B/HS2/02919, and supported by the CLARIN-PL[19] research infrastructure.

References

Ju D Apresjan. 1974. Regular polysemy. *Linguistics*, 12(142):5–32.

Jordi Atserias, Luıs Villarejo, German Rigau, Eneko Agirre, John Carroll, Bernardo Magnini, and Piek Vossen. 2004. The meaning multilingual central repository. In *2nd International Global Wordnet Conference, January 20-23, 2004: proceedings*, pages 23–30. Masaryk University.

Lucie Barque and François Régis Chaumartin. 2009. Regular polysemy in wordnet. *Journal for Language Technology and Computational Linguistics*, 24(2):5–18.

Yoav Benjamini and Yosef Hochberg. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300.

Francis Bond and Ryan Foster. 2013. Linking and extending an open multilingual wordnet. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1352–1362.

Francis Bond and Kyonghee Paik. 2012. A survey of wordnets and their licenses. In *Proceedings of the 6th International Global Wordnet Conference*, volume 8.

Paul Buitelaar. 1998. *CoreLex: systematic polysemy and underspecification*. Ph.D. thesis.

Paul Buitelaar. 2000. Reducing lexical semantic complexity with systematic polysemous classes and underspecification. In *NAACL-ANLP 2000 Workshop: Syntactic and Semantic Complexity in Natural Language Processing Systems*.

A. Cruse. 2006. *A Glossary of Semantics and Pragmatics*. Edinburgh University Press.

Christiane Fellbaum. 1998. Wordnet: An electronic lexical database.

[19] `http://clarin-pl.eu`

Abed Alhakim Freihat, Fausto Giunchiglia, and Biswanath Dutta. 2013a. Approaching regular polysemy in wordnet. In *proceedings of 5th International Conference on Information, Process, and Knowledge Management (eKNOW), Nice, France*.

Abed Alhakim Freihat, Fausto Giunchiglia, and Biswanath Dutta. 2013b. Regular polysemy in wordnet and pattern based approach. *International Journal On Advances in Intelligent Systems*, 6.

Abed Alhakim Freihat, Fausto Giunchiglia, and Biswanath Dutta. 2016. A taxonomic classification of wordnet polysemy types. In *Proceedings of the 8th GWC Global WordNet Conference*.

Aitor Gonzalez-Agirre, Egoitz Laparra, and German Rigau. 2012. Multilingual central repository version 3.0. In *LREC*, volume 2525, page 2529.

Jer Hayes, Tony Veale, and Nuno Seco. 2004. Enriching wordnet via generative metonymy and creative polysemy. In *LREC*. Citeseer.

Sture Holm. 1979. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pages 65–70.

John Lyons. 1977. *Semantics*, volume 1. Cambridge University Press: Cambridge.

John Lyons. 1995. *Linguistic semantics: An introduction*. Cambridge University Press.

Marek Maziarz, Maciej Piasecki, Ewa Rudnicka, Stan Szpakowicz, and Paweł Kędzia. 2016. PlWordNet 3.0 – a Comprehensive Lexical-Semantic Resource. In Nicoletta Calzolari, Yuji Matsumoto, and Rashmi Prasad, editors, *COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, December 11-16, 2016, Osaka, Japan*, pages 2259–2268. ACL, ACL.

Rada Mihalcea. 2003. Turning wordnet into an information retrieval resource: Systematic polysemy and conversion to hierarchical codes. *International journal of pattern recognition and artificial intelligence*, 17(05):689–704.

Roberto Navigli. 2009. Word sense disambiguation: A survey. *ACM computing surveys (CSUR)*, 41(2):1–69.

Martha Palmer, Hoa Trang Dang, and Christiane Fellbaum. 2007. Making fine-grained and coarse-grained sense distinctions, both manually and automatically. *Nat. Lang. Eng.*, 13(2):137–163.

Wim Peters. 2003. Metonymy as a cross-lingual phenomenon. In *Proceedings of the ACL 2003 Workshop on the Lexicon and Figurative Language*, pages 1–9.

Wim Peters. 2006. In search for more knowledge: Regular polysemy and knowledge acquisition. *Proceedings of GWC2006*.

Wim Peters and Ivonne Peters. 2000. Lexicalised systematic polysemy in wordnet. In *Proceedings of LREC-2000*.

Wim Peters, Ivonne Peters, and Piek Vossen. 1998. Automatic sense clustering in euroword-net. In *Proceedings of first international conference on language resource and evaluation: Granada, Spain, 28-30 May, 1998*, pages 409–416. ELRA.

Christian Ramiro, Mahesh Srinivasan, Barbara C. Malt, and Yang Xu. 2018. Algorithms in the historical emergence of word senses. *Proceedings of the National Academy of Sciences*, 115(10):2323–2328. URL `https://www.pnas.org/content/115/10/2323`.

John R. Taylor. 2000. *The Lexicon-Encyclopedia Interface*, chapter Approaches to word meaning: The network model (Langacker) and the two-level model (Bierwisch) in comparison, pages 115–142. Elsevier.

Jason Utt and Sebastian Padó. 2011. Ontology-based distinction between polysemy and homonymy. In *Proceedings of the Ninth International Conference on Computational Semantics (IWCS 2011)*.

Tony Veale. 2004. Polysemy and category structure in wordnet: An evidential approach. In *LREC*. Citeseer.