2013 IEEE 13th International Conference on Data Mining Workshops

Interpreting or describing? Measuring verb abstraction

Dominika Rogozińska Aleksander Wawer
Institute of Computer Science, Polish Academy of Sciences
ul. Jana Kazimierza 5, 01-248 Warszawa, Poland
dominika.rogozinska@ipipan.waw.pl, axw@ipipan.waw.pl

***Abstract*—The paper describes the results of machine learning experiments with verb classification according to the Linguistic Category Model (LCM)[1]. The LCM typology is a well-established tool to measure language abstraction, linked to sentiment and applicable in sentiment-analysis related areas. Our goal is to create automated methods of recognizing LCM verb classes. The method, demonstrated in the Polish language, turns out to be very promising, especially given the upper bounds set by inter-annotator agreement.**

I. Introduction

This paper describes experiments aimed at automating categorization of verbs in the Linguistic Category Model (LCM) [1].. The typology of verbs behind the LCM is closely related to sentiment analysis and applicable in selected opinion mining tasks. In LCM, verbs are classified according to to their level of abstraction and their link to sentiment (evaluative properties). The goal of this work is to describe automated methods of predicting LCM verb type using distributions of nominal verb arguments.

The work described in this paper is novel for three reasons. It is probably the first attempt at automated verb classification, according to the Linguistic Category Model. This typology, linked to sentiment and subjectivity, provides a means of measuring language abstraction. It is widely used in psycholinguistic studies, including those on deception detection (for instance, detecting opinion spam). However, the existing research on automated classification of verbs apply Levin's typology [2], which is (due to its syntactic backgrounds) of limited use in areas related to sentiment and emotion recognition and more widely, in psycholinguistics. Second, the design of feature space, as proposed in our work and aimed at distinguishing LCM verb classes, is also novel. Third, our work is the first effort to use machine learning and text processing techniques to internationalize LCM in a supervised, automatic fashion.

The paper is organized as follows. Section II describes the LCM typology and discusses the LCM verb classes. Section III contains an overview of the most related studies on automated verb classification. Section IV discusses the preparation of the hand-labelled verb list in Polish. In Section V we present the results and in Section VI conclude our work.

II. Linguistic Category Model

The typology of verbs of the Linguistic Category Model is directly relevant to sentiment analysis due to at least two reasons. First, sentiment is directly embedded in the LCM typology, it is one of the criteria of distinguishing verb types. Second, LCM has applications in many sentiment analysis-related fields, especially those where one needs to measure language abstraction. Such applications include opinion spam detection, as fake text writers tend to use more abstract vocabulary [3]. In the Polish language, [4] demonstrate that adding LCM features raises the precision of detecting fake product reviews.

The most general, top level distinction of the Linguistic Category Model is the one between state verbs and action verbs. As its authors put it, *state verbs (SV) refer to mental and emotional states or changes therein. SVs refer to either a cognitive (to think, to understand, etc.) or an affective state (to hate, to admire, etc.)*. This verb category is the most abstract one and also present in Levin's typology.

The other more concrete type of verbs in the LCM are action verbs. This type is always instantiated as one of its two sub-types, descriptive and interpretative action verbs (DAV and IAV) that all refer to specific actions (e.g., to hit, to help, to gossip, etc.) with a clearly defined beginning and end. SVs, in contrast, represent enduring states that don't have a clearly defined beginning and end.

The distinction between DAVs and IAVs is based on double criteria. The first states that DAVs have at least one physically invariant feature (eg. to kick - leg, to kiss - mouth), whereas IAVs do not (therefore, are more abstract than DAVs). The second criterion, sentiment, states that *IAVs have a pronounced evaluative component (e.g., positive IAVs such as to help, to encourage vs. negative IAVs such as to cheat, to bully), whereas DAVs do not (e.g., to phone, to talk). Descriptive action verbs (DAVs) are neutral in themselves (p.e. to push) but can gain an evaluative aspect dependent on the context (to push someone in front of a bus vs. to push someone away from an approaching bus).*

In practice, the criteria sometimes overlap. Some verbs have physical invariants but also have clear evaluative orientation. For instance, "to cry" always involves tears (an invariant physical feature), but carries negative sentiment.

The distinction between DAVs and IAVs was crucial for our efforts and for applications related to sentiment analysis. For example, one immediate way of using DAV and IAV information is to analyse opinionated texts, such as product reviews, as more descriptive or more interpretative.

In fact, the distinction between IAVs and DAVs may be seen as distinguishing subjective and objective verbs. Inter-

978-0-7695-5109-8/13 $31.00 © 2013 IEEE
DOI 10.1109/ICDMW.2013.92

pretations are by their nature subjective, while descriptions – objective.

III. Related Work

Because our work appears to be unique in that it deals with the LCM verb typology, the presentation of related works has to concentrate on studies on automated verb type classification using argument distributions and corpus frequencies.

Due to the fact that the most well-known verb typology appears to be Beth Levin's [2] distinction of 49 semantic classes, the experiments on automated recognition of verb types typically follow Levin's typology.

While the LCM focuses on psycholinguistic verb properties, useful in sentiment analysis, Levin's categorization links verb semantics to its syntactic behaviour.

Below we describe several of the key and most prominent studies.

Schulte im Walde [5] describes an approach to recognize Levin's classes using frequency counts of verbs for a number of sub-categorization frames and an unsupervised classification algorithm.

Merlo and Stevenson [6] train supervised classifiers on large annotated corpora to recognize three major types of English verbs. Their verb classification does not directly follow Levin's, but instead uses thematic roles of participants.

Decadt and Daelemans [7] use rule-based machine learning to distinguish six selected Levin's verb classes. Their approach was based on inducing lexical-level rules with object and subject nouns, extracted using a shallow parser from the BNC corpus.

The papers listed above share certain similarities. The methods are based on extracting verb arguments from corpora and applying classification algorithms to distinguish verb classes. What distinguishes our work is the verb typology (LCM), not used in this context, a dedicated feature space and focus on porting to languages other than English.

IV. Translating the LCM verb list

The list of manually-labelled LCM verbs in English can be found in the General Inquirer [8]. The list of 1516 entries (word senses rather than lexemes) contains three types of verbs: SV, IAV and DAV. We translated the list into Polish using the Microsoft Translation API.

The translation engine is context-sensitive and produces poor results when used for word-to-word translations (for example, use only English verb as an input, even preceded by "to"). After a number of experiments we found that the best results could be achieved by using the following sentence template: *"I can <verb>."*. The quality of translations was verified by human annotators. The translation process, after corrections, ended up with 1170 Polish verbs.

The list was verified independently by two annotators, familiar with the LCM annotation guideline, available from http://cratylus.org. Annotators had to make corrections to both LCM tags and verb sentiment.

To evaluate difficulty of the task, we computed inter-annotator agreement between annotators (LCM labels of Polish 1170 verbs), named **LCM-PL-A** and **LCM-PL-B**, and also between LCM labels by each of the annotators and LCM labels of the English equivalent of a Polish verb (**LCM-EN**). Table I summarizes these differences computed as Cohen's Kappa.

	Kappa
LCM-PL-B vs **LCM-PL-A**	0.78
LCM-PL-A vs **LCM-EN**	0.83
LCM-PL-B vs **LCM-EN**	0.87

Table I. Inter-annotator agreement

Generally, provided correct translations, it appears that English and Polish are not far in terms of their LCM labels. The agreement is reasonably high, but it is also clear that the task is far from entirely easy and free from ambiguities.

Unfortunately, the translated LCM labels acquired from English equivalents cover only a small subset of all verbs in Polish. The number of Polish verbs in infinitive forms could be approximated at over 10k, which emphasizes the need for automated methods.

In the following experiments we selected only those verbs, where both annotators agreed (used the same LCM labels). Frequencies of non-ambiguous verbs are as follows: 164 DAV verbs, 604 IAV verbs and 27 SV verbs.

V. Experiments and Results

In this section we present the experiments on automated classification of verbs according to the LCM.

The idea behind our experiments is that the distinction between IAV and DAV verbs is reflected in argument structure, specifically in the distribution of nouns in hyperonymy taxonomies. The intuition is that physically invariant features of DAVs might be reflected in a verb's tendency to occur with lower parts of selected trees only. IAV argument distributions might have more variation both within and between hyperonymy trees. Generally, the hypothesis is that distributional properties of verb arguments, measured on hyperonymy taxonomies, are relevant for the distinction between DAV and IAV. Therefore, the experiments are based on measuring abstraction of verb arguments (nouns in specific grammatical forms) using the Polish WordNet (Słowosieć) [9].

We begin with extracting a random sample of up to 1000 verb occurrences from the the National Corpus of Polish [11]. For each verb occurrence, we seek nouns that immediately follow it, assuming that these are its arguments. In order to filter out nouns which are not verb arguments, we pick only nouns in genitive or accusative (only in those two cases) appearing on the first or second position after the verb. We assume this to be a reasonable approximation to capture verb's arguments. For each extracted noun, we seek its synsets and compute its distance to the top node in the hyperonymy taxonomy. The distance is measured as the number of synsets (nodes) above the noun to reach the top hyperonymy node.

Then, for each verb we compute average distance of nouns towards the top of each hyperonymy taxonomy. Verbs with lower values of average distances should be more typically abstract ones (assumingly, DAV) and vice versa. For each verb, we form its feature space of average distances, computed for its nominal arguments within each hyperonymy taxonomy.

Since there are no reasonably performing and universal word sense disambiguation methods for the Polish language, we take all possible synsets of each noun (this often results in multiple hyperonymy taxonomies for a single noun), or alternatively the most frequent one.

On that feature space we perform the experiments described in the next parts of this section. The machine learning algorithm is Support Vector Machines, implementation in Sklearn machine learning environment[12]. The best performing parameters are polynomial kernel of degree two. Due to class imbalance, we enabled automatic class weighting. All the results presented are means of 3-fold cross validation, avg represents frequency-weighted average.

As the baseline for our experiment we assumed the most frequent class labelling (IAV). Due to class imbalance it obviously yields better overall results than the uniform random distribution. We present the baseline in Table II.

	precision	recall	f1-score
DAV	0.00	0.00	0.00
IAV	0.76	1.00	0.86
SV	0.00	0.00	0.00
avg	0.58	0.76	0.65

Table II. BASELINE RESULTS: THE MOST FREQUENT CLASS

Marking all verbs as the most frequent class (IAV) results in average score of 0.58 precision and 0.76 recall.

In the first experiment, we classify the verbs using the feature space with only 56 features obtained from 28 top frequent hyperonymy taxonomies, obtained by matching of the Polish Wordnet with Princeton WordNet. In other words, we selected only those top hyperonymy taxonomies that have their equivalents in the English language WordNet. We created features from all possible synsets of each noun (verb argument), which results in increasing the number of hyperonymy taxonomies for a noun. Table III presents the results of classification using this feature space.

	precision	recall	f1-score
DAV	0.38	0.61	0.46
IAV	0.88	0.64	0.74
SV	0.12	0.40	0.20
avg	0.75	0.63	0.66

Table III. CLASSIFICATION RESULTS: MANUALLY SELECTED 28 HYPERONYMY TAXONOMIES (56 FEATURES)

In the next experiment we focus on evaluating the performance of hyperonymy selection by matching against Princeton Wordnet. Therefore, we select exactly the same number of features (56 hyperonymy taxonomies), but this time we apply automated feature selection according to the χ^2 measure. As before, we take all possible synsets of each noun. Table IV presents the results of this feature space.

Comparing tables III and IV reveals that the manual taxonomy selection results in higher precision (0.75 avg) but lower recall (0.63 avg), while the automated selection of the same number of features improves recall (0.71 avg) at the loss of precision (0.67 avg). In the case of manual taxonomy selection, precision is improved for each verb class, DAV and SV have notably higher recall (even up to 0.4 recall in the case of IAV), but the major reason of problems is low recall of IAV verbs (only 0.64).

	precision	recall	f1-score
DAV	0.28	0.29	0.28
IAV	0.81	0.86	0.82
SV	0.07	0.15	0.09
avg	0.67	0.71	0.69

Table IV. CLASSIFICATION RESULTS: ALL HYPERONYMY TAXONOMIES, AUTOMATIC FEATURE SELECTION OF 56 FEATURES

The third experiment is an attempt at solving the problem of low recall of IAV verbs when using the high-precision manual feature set of 28 taxonomies. We weight the nouns (verb arguments) using a frequency list computed from the National Corpus of Polish [11], in a manner similar to the TF-IDF procedure. The results, summarized in Table V, are not discouraging as the precision compared to Table III dropped for every verb category (avg precision of 0.72) while the recall for IAV verbs increased only to 0.66.

	precision	recall	f1-score
DAV	0.34	0.59	0.43
IAV	0.86	0.66	0.75
SV	0.09	0.15	0.11
avg	0.72	0.63	0.66

Table V. CLASSIFICATION RESULTS: MANUALLY SELECTED 28 HYPERONYMY TAXONOMIES (56 FEATURES), FREQUENCY-WEIGHTED

The fourth experiment introduces two more alterations of the feature space. First, we change hyperonymy-based space: instead of taking all possible synsets of each noun, as before, this time we select only its most frequent synset. Second, we use concordances (context information) to create additional features from word co-occurrences: we count words (base forms) appearing 3 tokens left and right from a verb, remembering the position. Then, we perform feature selection, selecting top 100 features according to the χ^2 measure. Table VI presents the results of this feature space.

	precision	recall	f1-score
DAV	0.38	0.62	0.47
IAV	0.89	0.65	0.75
SV	0.11	0.37	0.17
avg	0.76	0.63	0.67

Table VI. CLASSIFICATION RESULTS: THE MOST FREQUENT SYNSET, CONTEXT INFORMATION

We tried to improve over the previous method by raising the number of features to 200 and using different context representation (in a bag-of-words manner, disregarding word order), however the results did not outperform those presented in Table VI.

Finally, we extended the feature space of the fourth experiment (Table VI) by morphosyntactic information. This type of information, specific for automated processing Slavic languages, contains part-of-speech as well as morphological and selected syntactic information about word forms. We counted occurrences of morphosyntactic information at specific positions within the same context width (3 tokens left and right from verb). Table VII presents the results for this feature space.

	precision	recall	f1-score
DAV	0.38	0.61	0.46
IAV	0.89	0.67	0.76
SV	0.13	0.41	0.20
avg	0.76	0.65	0.68

Table VII. CLASSIFICATION RESULTS: WITH MORPHOSYNTACTIC TAGS

The results presented in Table VII reveal a positive influence of morphosyntactic features and are the best overall in terms of precision (avg 0.76) and recall (avg 0.65). To summarize, the feature set consists of 28 hand-selected hyperonymy taxonomies that translate into 56 features for each of the two forms of nominal verb arguments (accusative and genitive). In addition to this, we added context-based features: corpus frequencies of base word forms and morphosyntactic information.

As the SV class contains only few dozen verbs, it is possible to label it by hand. Additionally, 22 out of 27 (81%) verbs in the dataset obtained by the General Inquirer translation are in the Polish WordNet (Słowosieć) [9] and already marked as state verbs[1]. Interestingly, 16% DAVs and 11% IAVs are also marked as state verbs according to the Polish WordNet.

For these reasons, below we consider a different and simplified scenario of distinguishing only between IAV and DAV verbs. In Table VIII we present the results for two classes using the best performing feature space of these described above.

	precision	recall	f1-score
DAV	0.50	0.43	0.46
IAV	0.85	0.88	0.87
avg	0.78	0.78	0.78

Table VIII. CLASSIFICATION OF IAV AND DAV ONLY

The results demonstrate again that DAV is by far more difficult to recognize. The precision of 0.5 is the highest value obtained for this class, but the recall of 0.43 is substantially lower than in the case of top 3-class results in Table VII.

VI. CONCLUSIONS AND FUTURE WORK

In this paper, we have demonstrated a method of automated classification of verbs in the Linguistic Category Model (LCM). We develop dedicated hyperonymy-based feature space from distributions of nominal arguments to distinguish interpretative and descriptive properties of verbs. We extract the features from large corpora and apply supervised machine learning to classify the verbs. The method has been implemented and evaluated for the Polish language. The results are a step up over the baseline and very promising, especially given the upper bounds set by inter-annotator agreement.

Our future work includes exploring other extensions of feature space to raise the performance of classifiers even further. We also consider using syntactic parsers to improve the precision of nominal argument extraction on sentence-level. We intend to apply the best performing models to label all Polish verbs and finally, repeat the experiments in other languages. Since DAV and IAV verbs can express concepts, an interesting extension would be to embed the distinction in concept-level sentiment analysis [13].

[1] CZASOWNIK STANOWY NDK

REFERENCES

[1] G. R. Semin and K. Fiedler, "The cognitive functions of linguistic categories in describing persons: Social cognition and language," *Journal of Personality and Social Psychology*, vol. 54, pp. 558–568, 1988.

[2] B. Levin, *English Verb Classes and Alternations: A Preliminary Investigations*. The University of Chicago Press, Chicago IL, USA, 1993.

[3] M. L. Newman, J. W. Pennebaker, D. S. Berry, and J. M. Richards, "Lying words: Predicting deception from linguistic styles," *Personality and Social Psychology Bulletin*, vol. 29, no. 5, pp. 665–675, 2003.

[4] M. Rubikowski and A. Wawer, "The scent of deception: Recognizing fake perfume reviews in polish," in *Language Processing and Intelligent Information Systems*. Springer Berlin Heidelberg, 2013, vol. 7912, pp. 45–49.

[5] S. im Walde, "Automatic semantic classification of verbs according to their alternation behaviour," *Arbeitspapiere des Instituts fur Maschinelle Sprachverarbeitung*, vol. 3, pp. 55–96, 1998.

[6] P. Merlo and S. Stevenson, "Automatic verb classification based on statistical distribution of argument structure," *Compoutational Linguistics*, vol. 23, pp. 373–408, 2001.

[7] B. Decadt and W. Daelemans, "Verb classification - machine learning experiments in classifying verbs into semantic classes," in *Proceedings of the LREC 2004 Workshop 'Beyond Named Entity Recognition - Semantic Labelling for NLP Tasks'*, 2004, pp. 25–30.

[8] P. J. Stone, D. C. Dunphy, D. M. Ogilvie, and M. S. Smith, *The General Inquirer: A Computer Approach to Content Analysis*. MIT Press, 1966.

[9] M. Piasecki, S. Szpakowicz, and B. Broda, *A Wordnet from the Ground Up*. Wrocław: Oficyna Wydawnicza Politechniki Wroclawskiej, 2009.

[10] C. Fellbaum, *WordNet: An Electronic Lexical Database*. MIT Press, 1998.

[11] A. Przepiórkowski, M. Bańko, R. L. Górski, and B. Lewandowska-Tomaszczyk, Eds., *Narodowy Korpus Języka Polskiego*. Wydawnictwo Naukowe PWN, 2012.

[12] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[13] E. Cambria, B. Schuller, Y. Xia, and C. Havasi, "New avenues in opinion mining and sentiment analysis," *Intelligent Systems, IEEE*, vol. 28, no. 2, pp. 15–21, 2013.