Cross-Lingual Metaphor Detection Using Common Semantic Features

Yulia Tsvetkov Elena Mukomel Anatole Gershman
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA, 15213, USA
{ytsvetko,helenm,anatoleg}@cs.cmu.edu

Abstract

We present the CSF - Common Semantic Features method for metaphor detection. This method has two distinguishing characteristics: it is cross-lingual and it does not rely on the availability of extensive manually-compiled lexical resources in target languages other than English. A metaphor detecting classifier is trained on English samples and then applied to the target language. The method includes procedures for obtaining semantic features from sentences in the target language. Our experiments with Russian and English sentences show comparable results, supporting our hypothesis that a CSF-based classifier can be applied across languages. We obtain state-of-the-art performance in both languages.

1 Introduction

Metaphors are very powerful pervasive communication tools that help deliver complex concepts and ideas simply and effectively (Lakoff and Johnson, 1980). Automatic detection and interpretation of metaphors is critical for many practical language processing tasks such as information extraction, summarization, opinion mining, and translation. In this paper, we focus on the automatic metaphor detection task. This problem gained much attention in natural language processing research mostly using the detection principles articulated by the Pragglejaz Group (2007). According to these principles, a lexical unit (a word or expression) is used metaphorically if its contextual meaning is different from its “basic contemporary” meaning. To apply this method, we need to be able to determine the basic meaning of a lexical unit and then test if this interpretation makes sense in the current context.

Several approaches to automatic detection of metaphors have been proposed (Gedigian et al., 2006; Krishnakumaran and Zhu, 2007; Shutova et al., 2010), all of which rely on the availability of extensive manually crafted lexical resources such as WordNet, VerbNet, FrameNet, TreeBank, etc. Unfortunately, such resources exist only for a few resource-rich languages such as English. For most other languages, such resources either do not exist or are of a low quality.

To our knowledge this work is the first empirical study of cross-lingual metaphor detection. We present the Common Semantic Features (CSF) approach to metaphor detection in languages without extensive lexical resources. In a target language it requires only a dependency parser and a target-English dictionary. We classify sentences into literal and metaphoric using automatically extracted coarse-grained semantic properties of words such as their propensity to refer to abstract versus concrete concepts, animate entities, artifacts, body parts, etc. These properties serve as features for the key relations in a sentence, which include Subject-Verb-Object (SVO) and Adjective-Noun (AN). A classifier trained on English sentences obtains a 0.78 F-score. The same classifier, trained solely on English sentences, achieves a similar level of performance on sentences from other languages such as Russian; this is the central contribution of this work. An additional important contribution is that in Russian we obtain the necessary semantic features

Proceedings of the First Workshop on Metaphor in NLP, pages 45–51,
Atlanta, Georgia, 13 June 2013.

without recourse to sophisticated non-English lexical resources. In this paper, we focus on the sentences where verbs are used metaphorically, leaving Adjective-Noun relations for future work. Based on our examination of over 500 metaphorical sentences in English and Russian collected from general news articles, we estimate that verb-based metaphors constitute about 40-50% of all metaphors.

We present and discuss our experiments with three sets of features: (1) features corresponding to the $lexicographer\ file\ names$ defined in WordNet 3.0 (Fellbaum, 1998), (2) features based on abstractness vs. concreteness computed using Vector Space Models (VSM), and (3) features based on the types of named entities, if present. Our main target language in these experiments has been Russian, but we also present preliminary experiments with Spanish.

The paper is organized as follows: Section 2 contains an overview of the resources we use; Section 3 discusses the methodology; Section 4 presents the experiments; in Section 5, we discuss related work, and we conclude with suggestions for future research in Section 6.

2 Datasets

We use the following English lexical resources to train our model:

TroFi Example Base[1] (Birke and Sarkar, 2007) of 3,737 English sentences from the *Wall Street Journal*. Each sentence contains one of the seed verbs and is marked L by human annotators if the verb is used in a literal sense. Otherwise, the sentence is marked N (non-literal). The model was evaluated on 25 target verbs with manually annotated 1 to 115 sentences per verb. TroFi does not define the basic meanings of these verbs, but provides examples of literal and metaphoric sentences which we use to train and evaluate our metaphor identification method.

WordNet (Fellbaum, 1998) is an English lexical database where each entry contains a set of synonyms (a synset) all representing the same concept. This database is compiled from a set of 45 lexicographer files[2] such as "noun.body" or "verb.cognition" identified by a number from 0 to 44, called $lexicographer\ file\ number$ (henceforth $lexFN$). The $lexFN$ of each synset is contained in the database. We use $lexFNs$ as coarse-grain semantic features of nouns and verbs.

MRC Psycholinguistic Database[3] (Wilson, 1988) is a dictionary containing 150,837 words with up to 26 linguistic and psycholinguistic attributes rated by human subjects in psycholinguistic experiments. It includes 4,295 words rated with degrees of abstractness; the ratings range from 158 (highly abstract) to 670 (highly concrete). We use these words as a seed when we calculate the values of abstractness and concreteness features for nouns and verbs in our training and test sets.

Word Representations via Global Context is a collection of 100,232 words and their vector representations.[4] These representations were extracted from a statistical model embedding both local and global contexts of words (Huang et al., 2012), intended to capture better the semantics of words. We use these vectors to calculate the values of abstractness and concreteness features of a word.

3 Methodology

We treat the metaphor detection problem as a task of binary classification of sentences. A sentence is represented by one or more key relations such as Subject-Verb-Object triples and Adjective-Noun pairs. In this paper, we focus only on the SVO relations and we allow either the S part or the O part to be empty. If all relations representing a sentence are classified literal by our model then the whole sentence is tagged literal. Otherwise, the sentence is tagged metaphoric.

[1] http://www.cs.sfu.ca/ anoop/students/jbirke/

[2] See http://wordnet.princeton.edu/man/lexnames.5WN.html for a full list of lexicographer file names.

[3] http://ota.oucs.ox.ac.uk/headers/1054.xml

[4] http://www.socher.org/index.php/Main/Improving-WordRepresentationsViaGlobalContextAndMultipleWordPrototypes

3.1 Model

We classify an SVO relation $\mathbf{x}$ as literal vs. metaphorical using a logistic regression classifier:

$$p(y \mid \mathbf{x}) \propto \exp \sum_j \lambda_j h_j(y, \mathbf{x}),$$

where $h_j(\cdot)$ are feature values computed for each word in $\mathbf{x}$, λ_j are the corresponding weights, and $y \in \{\mathtt{L}, \mathtt{M}\}$ refer to our classes: L for literal and M for metaphoric. The parameters λ_j are learned during training.

3.2 Features

An SVO relation is a concatenation of features for the S, V, and O parts. The S and O parts contain three types of features: (1) semantic categories of a word, (2) degree of abstractness of a word, and (3) types of named entities. The V part contains only the first two types of features.

Semantic categories are features corresponding to the WordNet $lexFNs$, introduced in Section 2. Since S and O are assumed to be nouns,[5] each has 26 semantic category features corresponding to the $lexFNs$ for nouns (3 through 28). These categories include *noun.animal, noun.artefact, noun.body, noun.cognition, noun.food, noun.location*, etc. The V part has 15 semantic category features corresponding to lexical ids for verbs (29 through 43), for example, *verb.motion* and *verb.cognition*. A lexical item can belong to several synsets with different $lexFNs$. For example, the word "head" when used as a noun participates in 33 synsets, 3 of which have $lexFN$ 08 (*noun.body*). The value of the feature corresponding to this $lexFN$ is $3/33 = 0.09$.

For a non-English word, we first obtain its most common translations to English and then select all corresponding English WordNet synsets. For example, when Russian word 'голова' is translated as *'head'* and *'brain'*, we select all the synsets for the nouns *head* and *brain*. There are 38 such synsets (33 for *head* and 5 for *brain*). Four of these synsets have $lexFN$ 08 (*noun.body*). Therefore, the value of the feature corresponding to this $lexFN$ is $4/38 = 0.10$. This dictionary-based mapping of non-English words into WN synsets is rather coarse. A more discriminating approach may improve the overall performance. In addition, WN synsets may not always capture all the meanings of non-English words. For example, Russian word 'нога' refers to both the *'foot'* and the *'leg'*. WN has synsets for *foot*, *leg* and *extremity*, but not for *lower extremity*.

Degree of abstractness According to Turney et al. (2011), "Abstract words refer to ideas and concepts that are distant from immediate perception, such as economics, calculating and disputable." Concrete words refer to physical objects and actions. Words with multiple senses can refer to both concrete and abstract concepts. Evidence from several languages suggests that concrete verbs tend to have concrete subjects and objects. If either the subject or an object of a concrete verb is abstract, then the verb is typically used in a figurative sense, indicating the presence of a metaphor. For example, when we hear that "an idea was born", we know that the word "born" is used figuratively. This observation motivates our decision to include the degree of abstractness in our feature set.

To calculate the degree of abstractness of English lexical items we use the vector space representations of words computed by Huang et al. (2012) and a separate supervised logistic regression classifier trained on a set of abstract and concrete words from the MRC dataset. Each value in a word's vector is a feature, thus, semantically similar words have similar feature values. Degrees of abstractness are posterior probabilities of the classifier predictions.

For non-English words, we use the following procedure. Suppose word w has n English translations whose degrees of abstractness are $a_1, a_2, \ldots a_n$ in decreasing order. If the majority is deemed abstract then $ABSTRACT(w) = a_1$, otherwise $ABSTRACT(w) = a_n$. This heuristic prefers the extreme interpretations, and is based on an observation that translations tend to be skewed to one side or the other of "abstractness". Our results may improve if we map non-English words more precisely into the most contextually-appropriate English senses.

Named entities (NE) is an additional category of features instrumental in metaphor identification. Specifically, we would like to distinguish whether an action (a verb in SVO) is performed by a human,

[5] We currently exclude pronouns from the relations that we learn.

an organization or a geographical entity. These distinctions are often needed to detect metonymy, as in "the White House said". Often, these entities are mentioned by their names which are not found in common dictionaries. Fortunately, there are many named entity recognizers (NER) for all major languages. In addition, Shah et al. (2010) showed that named entities tend to survive popular machine translation engines and can be relatively reliably detected even without a native NER. Based on these observations, we decided to include three boolean features corresponding to these NE categories: person, organization, and location.

4 Experiments

We train two classifiers: the first to calculate the degree of abstractness of a given word and the second to classify an SVO relation as metaphoric or literal. Both are logistic regression classifiers trained with the `creg` regression modeling framework.[6] To minimize the number of free parameters in our model we use ℓ_1 regularization.

4.1 Measuring abstractness

To train the abstractness classifier, we normalize abstractness scores of nouns from the MRC dataset to probabilities, and select 1,225 most abstract and 1,225 most concrete words. From these words, we set aside 25 randomly selected samples from each category for testing. We obtain the vector space representations of the remaining 1,400 samples and use the dimensions of these representations as features. We train the abstractness classifier on the 1,400 labeled samples and test it on the 50 samples that were set aside, obtaining 76% accuracy. The degree of abstractness of a word is the posterior probability produced by the abstractness classifier.

4.2 Metaphor detection

We train the metaphor classifier using labeled English SVO relations. To obtain these relations, we use the Turbo parser (Martins et al., 2010) to parse 1,592 literal and 1,609 metaphorical manually annotated sentences from the TroFi Example Base and extract 1,660 sentences that have SVO relations that contain annotated verbs: 696 literal and 964 metaphorical training instances. For example, the verb *flourish* is used literally in *"Methane-making bacteria flourish in the stomach"* and metaphorically in *"Economies flourish in free markets"*. From the first sentence we extract SVO relation `<bacteria, flourish, NIL>`, and `<economies, flourish, NIL>` from the second. We then build feature vectors, using feature categories described in Section 3.

We train several versions of the metaphor classifier for each feature category and for their combinations. The feature categories are designated as follows:

- WN - Semantic categories based on WordNet $lexFNs$
- VSM - Degree of abstractness based on word vectors
- NE - Named Entity categories

We evaluate the metaphor classifiers using 10-fold cross validation. The results are listed in Table 1.

Feature categories	Accuracy
WN	63.7%
VSM	64.1%
WN+VSM	67.7%
WN+NE	64.5%
WN+VSM+NE	**69.0%**

Table 1: 10-fold cross validation results of the metaphor classifier.

Our results are comparable to the accuracy of 64.9% reported by Birke and Sarkar (2007) on the TroFi dataset. The combination of all feature categories significantly improves over this baseline.

4.2.1 English metaphor detection

We compute precision, recall and F-score on a test set of 98 English sentences. This test set consists of 50 literal and 48 metaphorical sentences, where each metaphoric sentence contains a verb used in a figurative sense. The test sentences were selected from general news articles by independent collectors. Table 2 shows the results.

In this experiment, the WN group of features contributes the most. The addition of NE, while not improving the overall F-score, helps to reduce false positives and better balance precision and recall. The VSM features are considerably weaker perhaps

[6] https://github.com/redpony/creg

Feature categories	Precision	Recall	F-score
WN	0.75	0.81	0.78
VSM	0.57	0.71	0.63
WN+VSM	0.66	0.90	0.76
WN+NE	0.78	0.79	**0.78**
WN+VSM+NE	0.68	0.71	0.69

Table 2: Evaluation of the metaphor classifier on the test set of 50 literal and 48 metaphoric English sentences from news articles.

because we used single model vector space representations where each word uses only one vector that combines all its senses.

4.2.2 Russian metaphor detection

In a cross-lingual experiment, we evaluate our algorithm on a set of 140 Russian sentences: 62 literal and 78 metaphoric, selected from general news articles by two independent collectors. As in English, each metaphoric sentence contains a verb used in a figurative sense. We used the AOT parser[7] to obtain the SVO relations and the Babylon dictionary[8] to obtain English translations of individual words. The example sentence in Figure 1 contains one SVO relation with missing O part. We show the set of features and their values that were extracted from words in this relation.

The results of the Russian test set, listed in Table 3, are similar to the English results, supporting our hypothesis that a semantic classifier can work across languages. As in the previous experiment, the WN features are the most effective and the NE features contribute to improved precision.

Feature categories	Precision	Recall	F-score
WN	0.74	0.76	0.75
VSM	0.66	0.73	0.69
WN+VSM	0.70	0.73	0.71
WN+NE	0.82	0.71	**0.76**
WN+VSM+NE	0.74	0.72	0.73

Table 3: Evaluation of the metaphor classifier on the test set of 62 literal and 78 metaphoric Russian sentences from news articles.

While we did not conduct a full-scale experiment with Spanish, we ran a pilot using 51 sentences: 24 literal and 27 metaphoric. We obtained the F-score of 0.66 for the WN+VSM combination. We take it as a positive sign and will conduct more experiments.

5 Related work

Our work builds on the research of Birke and Sarkar (2007) who used an active learning approach to create an annotated corpus of sentences with literal and figurative senses of 50 common English verbs. The result was the TroFi Example Base set of 3,737 labeled sentences, which was used by the authors to train several classifiers. These algorithms were tested on sentences containing 25 English verbs not included in the original set. The authors report F-scores around 64.9%. We used this dataset for training and evaluation, and Birke and Sarkar's (2007) results as a baseline.

In a more recent work, Turney et al. (2011) suggested that the degree of abstractness of a word's context is correlated with the likelihood that the word is used metaphorically. To compute the abstractness of a word, the authors use a variation of Turney and Littman's (2003) algorithm comparing the word to twenty typically abstract words and twenty typically concrete words. Latent Semantic Analysis (Deerwester et al., 1990) is used to measure semantic similarity between each pair of words. A feature vector is generated for each word and a logistic regression classifier is used. The result is an average F-score of 63.9% on the TroFi dataset,[9] compared to Birke and Sarkar's (2007) 64.9%. In another experiment on 100 adjective-noun phrases labeled as literal or non-literal, according to the sense of the adjective, this algorithm obtains an average accuracy of 79%. While we obtain comparable results, our work extends this method in several important directions. First, we show how to apply a metaphor classifier across languages. Second, we extend our feature set beyond abstractness criteria. Finally, we propose an alternative technique to measure degrees of abstractness.

[7] www.aot.ru

[8] www.babylon.com

[9] Turney et al. (2011) report on two experimental setups with TroFi, our setup is closer to their first experiment.

Общество зреет десятилетиями .
'Society ripens over decades'

SVO = <Общество, зреет, NIL>

	Subject		Verb	
WN	noun.group	0.54	verb.change	0.75
	noun.state	0.23	verb.body	0.125
	noun.possession	0.15	verb.communication	0.125
	noun.location	0.08		
VSM	Abstractness	0.87	Abstractness	0.93

Figure 1: Features extracted for a Russian test sentence classified as metaphoric by our model.

6 Conclusions and future work

We presented CSF – an approach to metaphor detection based on semantic rather than lexical features. We described our experiments with an initial set of fairly coarse-grained features and showed how these features can be obtained in languages that lack extensive lexical resources. Semantic, as opposed to lexical features, are common to all languages which allows a classifier trained to detect metaphors in one language to be successfully applied to sentences in another language. Our results suggest that metaphors can be detected on a conceptual level, independently of whether they are expressed in Russian or English, supporting Lakoff and Johnson's (1980) claim that metaphors are parts of a pervasive conceptual system.

Our current work has been limited to the detection of figurative SVO relations, which account for about half of all metaphors in English and Russian. Other languages such as Farsi have a greater proportion of metaphors based on figurative use of adjectives and nouns. We plan to include more relations and expand our set of semantic features as part of the future research.

Acknowledgments

We are grateful to Chris Dyer for his invaluable advice. We are also grateful to the three anonymous reviewers for their constructive suggestions. Supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Defense US Army Research Laboratory contract number W911NF-12-C-0020. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoD/ARL, or the U.S. Government.

References

Julia Birke and Anoop Sarkar. 2007. Active learning for the identification of nonliteral language. In *Proceedings of the Workshop on Computational Approaches to Figurative Language*, FigLanguages '07, pages 21–28.

Scott C. Deerwester, Susan T. Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, 41(6):391–407.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. Language, Speech and Communication. MIT Press.

Matt Gedigian, John Bryant, Srini Narayanan, and Branimir Ciric. 2006. Catching metaphors. In *Proceedings of the 3rd Workshop on Scalable Natural Language Understanding*, pages 41–48.

Pragglejaz Group. 2007. MIP: A method for identifying metaphorically used words in discourse. *Metaphor and Symbol*, 22(1):1–39.

Eric H. Huang, Richard Socher, Christopher D. Manning, and Andrew Y. Ng. 2012. Improving word representations via global context and multiple word prototypes. In *Annual Meeting of the Association for Computational Linguistics*, ACL 2012.

Saisuresh Krishnakumaran and Xiaojin Zhu. 2007. Hunting elusive metaphors using lexical resources. In *Proceedings of the Workshop on Computational approaches to Figurative Language*, pages 13–20.

George Lakoff and Mark Johnson. 1980. Conceptual metaphor in everyday language. *The Journal of Philosophy*, pages 453–486.

André F. T. Martins, Noah A. Smith, Eric P. Xing, Pedro M. Q. Aguiar, and Mário A. T. Figueiredo. 2010. Turbo parsers: dependency parsing by approximate variational inference. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pages 34–44.

Rushin Shah, Bo Lin, Anatole Gershman, and Robert Frederking. 2010. SYNERGY: a named entity recognition system for resource-scarce languages such as Swahili using online machine translation. In *Proceedings of the Second Workshop on African Language Technology*, AfLaT 2010, pages 21–26.

Ekaterina Shutova, Lin Sun, and Anna Korhonen. 2010. Metaphor identification using verb and noun clustering. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 1002–1010.

Peter D. Turney and Michael L. Littman. 2003. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information and System Security*, 21(4):315–346.

Peter D. Turney, Yair Neuman, Dan Assaf, and Yohai Cohen. 2011. Literal and metaphorical sense identification through concrete and abstract context. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 680–690.

Michael Wilson. 1988. MRC Psycholinguistic Database: Machine-usable dictionary, version 2.00. *Behavior Research Methods, Instruments, & Computers*, 20(1):6–10.