Metaphor Detection with Cross-Lingual Model Transfer

Yulia Tsvetkov **Leonid Boytsov** **Anatole Gershman** **Eric Nyberg** **Chris Dyer**
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA 15213 USA
{ytsvetko, srchvrs, anatoleg, ehn, cdyer}@cs.cmu.edu

Abstract

We show that it is possible to reliably discriminate whether a syntactic construction is meant literally or metaphorically using lexical semantic features of the words that participate in the construction. Our model is constructed using English resources, and we obtain state-of-the-art performance relative to previous work in this language. Using a model transfer approach by pivoting through a bilingual dictionary, we show our model can identify metaphoric expressions in other languages. We provide results on three new test sets in Spanish, Farsi, and Russian. The results support the hypothesis that metaphors are conceptual, rather than lexical, in nature.

1 Introduction

Lakoff and Johnson (1980) characterize metaphor as reasoning about one thing in terms of another, i.e., a metaphor is a type of *conceptual mapping*, where words or phrases are applied to objects and actions in ways that do not permit a literal interpretation. They argue that metaphors play a fundamental communicative role in verbal and written interactions, claiming that much of our everyday language is delivered in metaphorical terms. There is empirical evidence supporting the claim: recent corpus studies have estimated that the proportion of words used metaphorically ranges from 5% to 20% (Steen et al., 2010), and Thibodeau and Boroditsky (2011) provide evidence that a choice of metaphors affects decision making.

Given the prevalence and importance of metaphoric language, effective automatic detection of metaphors would have a number of benefits, both practical and scientific. Language processing applications that need to understand language or preserve meaning (information extraction, machine translation, dialog systems, sentiment analysis, and text analytics, etc.) would have access to a potentially useful high-level bit of information about whether something is to be understood literally or not. Second, scientific hypotheses about metaphoric language could be tested more easily at a larger scale with automation.

However, metaphor detection is a hard problem. On one hand, there is a subjective component: humans may disagree whether a particular expression is used metaphorically or not, as there is no clear-cut semantic distinction between figurative and metaphorical language (Shutova, 2010). On the other, metaphors can be domain- and context-dependent.[1]

Previous work has focused on metaphor identification in English, using both extensive manually-created linguistic resources (Mason, 2004; Gedigian et al., 2006; Krishnakumaran and Zhu, 2007; Turney et al., 2011; Broadwell et al., 2013) and corpus-based approaches (Birke and Sarkar, 2007; Shutova et al., 2013; Neuman et al., 2013; Shutova and Sun, 2013; Hovy et al., 2013). We build on this foundation and also extend metaphor detection into other languages in which few resources may exist. Our work makes the following contributions: (1) we develop a new state-of-the-art English metaphor detection system that uses *conceptual* semantic features, such as a degree of abstractness and semantic supersenses;[2] (2) we create new metaphor-annotated corpora for Russian and English;[3] (3) using a paradigm of model transfer (McDonald et al., 2011; Täckström et al., 2013; Kozhenikov and Titov, 2013), we provide support for the hypothesis that metaphors are concep-

[1] For example, *drowning students* could be used metaphorically to describe the situation where students are overwhelmed with work, but in the sentence *a lifeguard saved drowning students*, this phrase is used literally.

[2] https://github.com/ytsvetko/metaphor

[3] http://www.cs.cmu.edu/~ytsvetko/metaphor/datasets.zip

tual (rather than lexical) in nature by showing that our English-trained model can detect metaphors in Spanish, Farsi, and Russian.

2 Methodology

Our task in this work is to define features that distinguish between metaphoric and literal uses of two syntactic constructions: subject-verb-object (SVO) and adjective-noun (AN) tuples.[4] We give examples of a prototypical metaphoric usage of each type:

- **SVO metaphors.** A sentence containing a metaphoric SVO relation is *my car drinks gasoline*. According to Wilks (1978), this metaphor represents a violation of selectional preferences for the verb *drink*, which is normally associated with animate subjects (the car is inanimate and, hence, cannot drink in the literal sense of the verb).

- **AN metaphors.** The phrase *broken promise* is an AN metaphor, where attributes from a concrete domain (associated with the concrete word *broken*) are transferred to a more abstract domain, which is represented by the relatively abstract word *promise*. That is, we map an abstract concept *promise* to a concrete domain of physical things, where things can be literally broken to pieces.

Motivated by Lakoff's (1980) argument that metaphors are systematic conceptual mappings, we will use coarse-grained *conceptual*, rather than fine-grained *lexical* features, in our classifier. Conceptual features pertain to concepts and ideas as opposed to individual words or phrases expressed in a particular language. In this sense, as long as two words in two different languages refer to the same concepts, their conceptual features should be the same. Furthermore, we hypothesize that our coarse semantic features give us a language-invariant representation suitable for metaphor detection. To test this hypothesis, we use a cross-lingual model transfer approach: we use bilingual dictionaries to project words from other syntactic constructions found in other languages into English and then apply the English model on the derived conceptual representations.

[4]Our decision to focus on SVO and AN metaphors is justified by corpus studies that estimate that verb- and adjective-based metaphors account for a substantial proportion of all metaphoric expressions, approximately 60% and 24%, respectively (Shutova and Teufel, 2010; Gandy et al., 2013).

Each SVO (or AN) instance will be represented by a triple (duple) from which a feature vector will be extracted.[5] The vector will consist of the concatenation of the conceptual features (which we discuss below) for all participating words, and conjunction features for word pairs.[6] For example, to generate the feature vector for the SVO triple $(car, drink, gasoline)$, we compute all the features for the individual words *car*, *drink*, *gasoline* and combine them with the conjunction features for the pairs *car drink* and *drink gasoline*.

We define three main feature categories (1) abstractness and imageability, (2) supersenses, (3) unsupervised vector-space word representations; each category corresponds to a group of features with a common theme and representation.

- **Abstractness and imageability.** Abstractness and imageability were shown to be useful in detection of metaphors (it is easier to invoke mental pictures of concrete and imageable words) (Turney et al., 2011; Broadwell et al., 2013). We expect that abstractness, used in conjunction features (e.g., a feature denoting that the subject is abstract and the verb is concrete), is especially useful: semantically, an abstract agent performing a concrete action is a strong signal of metaphorical usage.

  Although often correlated with abstractness, imageability is not a redundant property. While most abstract things are hard to visualize, some call up images, e.g., *vengeance* calls up an emotional image, *torture* calls up emotions and even visual images. There are concrete things that are hard to visualize too, for example, *abbey* is harder to visualize than *banana* (B. MacWhinney, personal communication).

- **Supersenses.** Supersenses[7] are coarse semantic categories originating in WordNet. For nouns and verbs there are 45 classes: 26 for nouns and 15 for verbs, for example,

[5]Looking at components of the syntactic constructions independent of their context has its limitations, as discussed above with the *drowning students* example; however, it simplifies the representation challenges considerably.

[6]If word one is represented by features $\mathbf{u} \in \mathbb{R}^n$ and word two by features $\mathbf{v} \in \mathbb{R}^m$ then the conjunction feature vector is the vectorization of the outer product $\mathbf{u}\mathbf{v}^\top$.

[7]Supersenses are called "lexicographer classes" in WordNet documentation (Fellbaum, 1998), `http://wordnet.princeton.edu/man/lexnames.5WN.html`

noun.body, *noun.animal*, *verb.consumption*, or *verb.motion* (Ciaramita and Altun, 2006). English adjectives do not, as yet, have a similar high-level semantic partitioning in WordNet, thus we use a 13-class taxonomy of adjective supersenses constructed by Tsvetkov et al. (2014) (discussed in §3.2).

Supersenses are particularly attractive features for metaphor detection: coarse sense taxonomies can be viewed as semantic concepts, and since concept mapping is a process in which metaphors are born, we expect different supersense co-occurrences in metaphoric and literal combinations. In "drinks gasoline", for example, mapping to supersenses would yield a pair <*verb.consumption, noun.substance*>, contrasted with <*verb.consumption, noun.food*> for "drinks juice". In addition, this coarse semantic categorization is preserved in translation (Schneider et al., 2013), which makes supersense features suitable for cross-lingual approaches such as ours.

- **Vector space word representations.** Vector space word representations learned using unsupervised algorithms are often effective features in supervised learning methods (Turian et al., 2010). In particular, many such representations are designed to capture lexical semantic properties and are quite effective features in semantic processing, including named entity recognition (Turian et al., 2009), word sense disambiguation (Huang et al., 2012), and lexical entailment (Baroni et al., 2012). In a recent study, Mikolov et al. (2013) reveal an interesting cross-lingual property of distributed word representations: there is a strong similarity between the vector spaces across languages that can be easily captured by linear mapping. Thus, vector space models can also be seen as vectors of (latent) semantic concepts, that preserve their "meaning" across languages.

3 Model and Feature Extraction

In this section we describe a classification model, and provide details on mono- and cross-lingual implementation of features.

3.1 Classification using Random Forests

To make classification decisions, we use a random forest classifier (Breiman, 2001), an ensemble of decision tree classifiers learned from many independent subsamples of the training data. Given an input, each tree classifier assigns a probability to each label; those probabilities are averaged to compute the probability distribution across the ensemble. Random forest ensembles are particularly suitable for our resource-scarce scenario: rather than overfitting, they produce a limiting value of the generalization error as the number of trees increases,[8] and no hyperparameter tuning is required. In addition, decision-tree classifiers learn non-linear responses to inputs and often outperform logistic regression (Perlich et al., 2003).[9] Our random forest classifier models the probability that the input syntactic relation is metaphorical. If this probability is above a threshold, the relation is classified as metaphoric, otherwise it is literal. We used the `scikit-learn` toolkit to train our classifiers (Pedregosa et al., 2011).

3.2 Feature extraction

Abstractness and imageability. The MRC psycholinguistic database is a large dictionary listing linguistic and psycholinguistic attributes obtained experimentally (Wilson, 1988).[10] It includes, among other data, 4,295 words rated by the degrees of abstractness and 1,156 words rated by the imageability. Similarly to Tsvetkov et al. (2013), we use a logistic regression classifier to propagate abstractness and imageability scores from MRC ratings to all words for which we have vector space representations. More specifically, we calculate the degree of abstractness and imageability of all English items that have a vector space representation, using vector elements as features. We train two separate classifiers for abstractness and imageability on a seed set of words from the MRC database. Degrees of abstractness and imageability are posterior probabilities of classifier predictions. We binarize these posteriors into abstract-concrete (or imageable-unimageable) boolean indicators using pre-defined thresholds.[11] Perfor-

[8] See Theorem 1.2 in (Breiman, 2001) for details.

[9] In our experiments, random forests model slightly outperformed logistic regression and SVM classifiers.

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

[11] Thresholds are equal to 0.8 for abstractness and to 0.9 for imageability. They were chosen empirically based on ac-

mance of these classifiers, tested on a sampled held-out data, is 0.94 and 0.85 for the abstractness and imageability classifiers, respectively.

Supersenses. In the case of SVO relations, we incorporate supersense features for nouns and verbs; noun and adjective supersenses are used in the case of AN relations.

Supersenses of nouns and verbs. A lexical item can belong to several synsets, which are associated with different supersenses. Degrees of membership in different supersenses are represented by feature vectors, where each element corresponds to one supersense. For example, the word *head* (when used as a noun) participates in 33 synsets, three of which are related to the supersense *noun.body*. The value of the feature corresponding to this supersense is $3/33 \approx 0.09$.

Supersenses of adjectives. WordNet lacks coarse-grained semantic categories for adjectives. To divide adjectives into groups, Tsvetkov et al. (2014) use 13 top-level classes from the adapted taxonomy of Hundsnurscher and Splett (1982), which is incorporated in GermaNet (Hamp and Feldweg, 1997). For example, the top-level classes in GermaNet include: *adj.feeling* (e.g., willing, pleasant, cheerful); *adj.substance* (e.g., dry, ripe, creamy); *adj.spatial* (e.g., adjacent, gigantic).[12] For each adjective type in WordNet, they produce a vector with a classifier posterior probabilities corresponding to degrees of membership of this word in one of the 13 semantic classes,[13] similar to the feature vectors we build for nouns and verbs. For example, for a word *calm* the top-2 categories (with the first and second highest degrees of membership) are *adj.behavior* and *adj.feeling*.

Vector space word representations. We employ 64-dimensional vector-space word representations constructed by Faruqui and Dyer (2014).[14] Vector construction algorithm is a variation on traditional latent semantic analysis (Deerwester et al., 1990) that uses multilingual information to produce representations in which synonymous words have similar vectors. The vectors were trained on the news commentary corpus released by WMT-2011,[15] comprising 180,834 types.

3.3 Cross-lingual feature projection

For languages other than English, feature vectors are projected to English features using translation dictionaries. We used the Babylon dictionary,[16] which is a proprietary resource, but any bilingual dictionary can in principle be used. For a non-English word in a source language, we first obtain all translations into English. Then, we average all feature vectors related to these translations. Consider an example related to projection of WordNet supersenses. A Russian word голова is translated as *head* and *brain*. Hence, we select all the synsets of the nouns *head* and *brain*. There are 38 such synsets (33 for *head* and 5 for *brain*). Four of these synsets are associated with the supersense *noun.body*. Therefore, the value of the feature *noun.body* is $4/38 \approx 0.11$.

4 Datasets

In this section we describe a training and testing dataset as well a data collection procedure.

4.1 English training sets

To train an SVO metaphor classifier, we employ the TroFi (Trope Finder) dataset.[17] TroFi includes 3,737 manually annotated English sentences from the *Wall Street Journal* (Birke and Sarkar, 2007). Each sentence contains either literal or metaphorical use for one of 50 English verbs. First, we use a dependency parser (Martins et al., 2010) to extract subject-verb-object (SVO) relations. Then, we filter extracted relations to eliminate parsing-related errors, and relations with verbs which are not in the TroFi verb list. After filtering, there are 953 metaphorical and 656 literal SVO relations which we use as a training set.

In the case of AN relations, we construct and make publicly available a training set containing 884 metaphorical AN pairs and 884 pairs with literal meaning. It was collected by two annotators using public resources (collections of metaphors on the web). At least one additional person carefully examined and culled the collected metaphors, by removing duplicates, weak metaphors, and metaphorical phrases (such as

curacy during cross-validation.

[12] For the full taxonomy see http://www.sfs.uni-tuebingen.de/lsd/adjectives.shtml

[13] http://www.cs.cmu.edu/~ytsvetko/adj-supersenses.tar.gz

[14] http://www.cs.cmu.edu/~mfaruqui/soft.html

[15] http://www.statmt.org/wmt11/

[16] http://www.babylon.com

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

drowning students) whose interpretation depends on the context.

4.2 Multilingual test sets

We collect and annotate metaphoric and literal test sentences in four languages. Thus, we compile eight test datasets, four for SVO relations, and four for AN relations. Each dataset has an equal number of metaphors and non-metaphors, i.e., the datasets are balanced. English (EN) and Russian (RU) datasets have been compiled by our team and are publicly available. Spanish (ES) and Farsi (FA) datasets are published elsewhere (Levin et al., 2014). Table 1 lists test set sizes.

	SVO	AN
EN	222	200
RU	240	200
ES	220	120
FA	44	320

Table 1: Sizes of the eight test sets. Each dataset is balanced, i.e., it has an equal number of metaphors and non-metaphors. For example, English SVO dataset has 222 relations: 111 metaphoric and 111 literal.

We used the following procedure to compile the EN and RU test sets. A moderator started with seed lists of 1000 most common verbs and adjectives.[18]

Then she used the SketchEngine, which provides searching capability for the TenTen Web corpus,[19] to extract sentences with words that frequently co-occurred with words from the seed lists. From these sentences, she removed sentences that contained more than one metaphor, and sentences with non-SVO and non-AN metaphors. Remaining sentences were annotated by several native speakers (five for English and six for Russian), who judged AN and SVO phrases in context. The annotation instructions were general: *"Please, mark in bold all words that, in your opinion, are used non-literally in the following sentences. In many sentences, all the words may be used literally."* The Fleiss' Kappas for 5 English and 6 Russian annotators are: EN-AN = .76, RU-AN = .85, EN-SVO = .75, RU-SVO = .78. For the final selection, we filtered out low-agreement ($<.8$) sentences.

The test candidate sentences were selected by a person who did not participate in the selection of the training samples. No English annotators of the test set, and only one Russian annotator out of 6 participated in the selection of the training samples. Thus, we trust that annotator judgments were not biased towards the cases that the system is trained to process.

5 Experiments

5.1 English experiments

Our task, as defined in Section 2, is to classify SVO and AN relations as either metaphoric or literal. We first conduct a 10-fold cross-validation experiment on the training set defined in Section 4.1. We represent each candidate relation using the features described in Section 3.2, and evaluate performance of the three feature categories and their combinations. This is done by computing an accuracy in the 10-fold cross validation. Experimental results are given in Table 2, where we also provide the number of features in each feature set.

	SVO		AN	
	# FEAT	ACC	# FEAT	ACC
AbsImg	20	0.73*	16	0.76*
Supersense	67	0.77*	116	0.79*
AbsImg+Sup.	87	0.78*	132	0.80*
VSM	192	0.81	228	0.84*
All	279	**0.82**	360	**0.86**

Table 2: 10-fold cross validation results for three feature categories and their combination, for classifiers trained on English SVO and AN training sets. # FEAT column shows a number of features. ACC column reports an accuracy score in the 10-fold cross validation. Statistically significant differences ($p < 0.01$) from the all-feature combination are marked with a star.

These results show superior performance over previous state-of-the-art results, confirming our hypothesis that conceptual features are effective in metaphor classification. For the SVO task, the cross-validation accuracy is about 10% better than that of Tsvetkov et al. (2013). For the AN task, the cross validation accuracy is better by 8% than the result of Turney et al. (2011) (two baseline

[18] Selection of 1000 most common verbs and adjectives achieves much broader lexical and domain coverage than what can be realistically obtained from continuous text. Our test sentence domains are, therefore, diverse: economic, political, sports, etc.

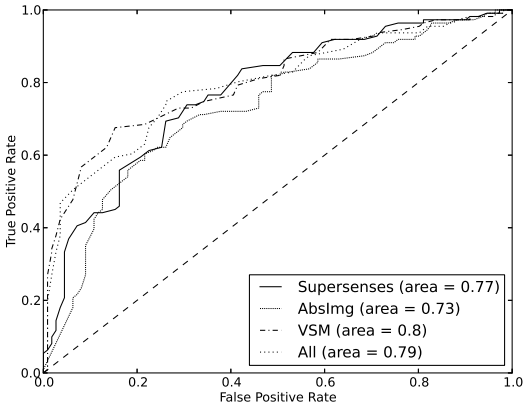
[19] http://trac.sketchengine.co.uk/wiki/Corpora/enTenTen

methods are described in Section 5.2). We can see that all types of features have good performance on their own (VSM is the strongest feature type). Noun supersense features alone allows us to achieve an accuracy of 75%, i.e., adjective supersense features contribute 4% to adjective-noun supersense feature combination. Experiments with the pairs of features yield better results than individual features, implying that the feature categories are not redundant. Yet, combining all features leads to even higher accuracy during cross-validation. In the case of the AN task, a difference between the All feature combination and any other combination of features listed in Table 2 is statistically significant ($p < 0.01$ for both the sign and the permutation test).

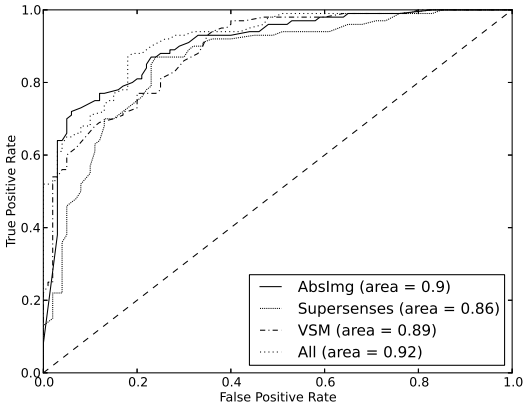
Although the first experiment shows very high scores, the 10-fold cross-validation cannot fully reflect the generality of the model, because all folds are parts of the same corpus. They are collected by the same human judges and belong to the same domain. Therefore, experiments on out-of-domain data are crucial. We carry out such experiments using held-out SVO and AN EN test sets, described in Section 4.2 and Table 1. In this experiment, we measure the f-score. We classify SVO and AN relations using a classifier trained on the All feature combination and balanced thresholds. The values of the f-score are 0.76, both for SVO and AN tasks. This out-of-domain experiment suggests that our classifier is portable across domains and genres.

However, (1) different application may have different requirements for recall/precision, and (2) classification results may be skewed towards having high precision and low recall (or vice versa). It is possible to trade precision for recall by choosing a different threshold. Thus, in addition to giving a single f-score value for balanced thresholds, we present a Receiver Operator Characteristic (ROC) curve, where we plot a fraction of true positives against the fraction of false positives for 100 threshold values in the range from zero to one. The area under the ROC curve (AUC) can be interpreted as the probability that a classifier will assign a higher score to a randomly chosen positive example than to a randomly chosen negative example.[20] For a randomly guessing classifier, the ROC curve is a dashed diagonal line. A bad classifier has an ROC curve that goes close to the dashed diagonal or even below it.

[20] Assuming that positive examples are labeled by ones, and negative examples are labeled by zeros.



(a) **SVO**



(b) **AN**

Figure 1: ROC curves for classifiers trained using different feature sets (English SVO and AN test sets).

According to ROC plots in Figure 1, all three feature sets are effective, both for SVO and for AN tasks. Abstractness and Imageability features work better for adjectives and nouns, which is in line with previous findings (Turney et al., 2011; Broadwell et al., 2013). It can be also seen that VSM features are very effective. This is in line with results of Hovy et al. (2013), who found that it is hard to improve over the classifier that uses only VSM features.

5.2 Comparison to baselines

In this section, we compare our method to state-of-the-art methods of Tsvetkov et al. (2013) and of Turney et al. (2011), who focused on classifying SVO and AN relations, respectively.

In the case of SVO relations, we use software

and datasets from Tsvetkov et al. (2013). These datasets, denoted as an SVO-baseline, consist of 98 English and 149 Russian sentences. We train SVO metaphor detection tools on SVO relations extracted from TroFi sentences and evaluate them on the SVO-baseline dataset. We also use the same thresholds for classifier posterior probabilities as Tsvetkov et al. (2013). Our approach is different from that of Tsvetkov et al. (2013) in that it uses additional features (vector space word representations) and a different classification method (we use random forests while Tsvetkov et al. (2013) use logistic regression). According to Table 3, we obtain higher performance scores for both Russian and English.

	EN	**RU**
SVO-baseline	0.78	0.76
This work	0.86	0.85

Table 3: Comparing f-scores of our SVO metaphor detection method to the baselines.

In the case of AN relations, we use the dataset (denoted as an AN-baseline) created by Turney et al. (2011) (see Section 4.1 in the referred paper for details). Turney et al. (2011) manually annotated 100 pairs where an adjective was one of the following: *dark*, *deep*, *hard*, *sweet*, and *worm*. The pairs were presented to five human judges who rated each pair on a scale from 1 (very literal/denotative) to 4 (very non-literal/connotative). Turney et al. (2011) train logistic-regression employing only abstractness ratings as features. Performance of the method was evaluated using the 10-fold cross-validation separately for each judge.

We replicate the above described evaluation procedure of Turney et al. (2011) using their model and features. In our classifier, we use the All feature combination and the balanced threshold as described in Section 5.1.

According to results in Table 4, almost all of the judge-specific f-scores are slightly higher for our system, as well as the overall average f-score.

In both baseline comparisons, we obtain performance at least as good as in previously published studies.

5.3 Cross-lingual experiments

In the next experiment we corroborate the main hypothesis of this paper: a model trained on English data can be successfully applied to other languages. Namely, we use a trained English model discussed in Section 5.1 to classify literal and metaphoric SVO and AN relations in English, Spanish, Farsi and Russian test sets, listed in Section 4.2. This time we used all available features.

	AN-baseline	**This work**
Judge 1	0.73	0.75
Judge 2	0.81	0.84
Judge 3	0.84	0.88
Judge 4	0.79	0.81
Judge 5	0.78	0.77
average	0.79	0.81

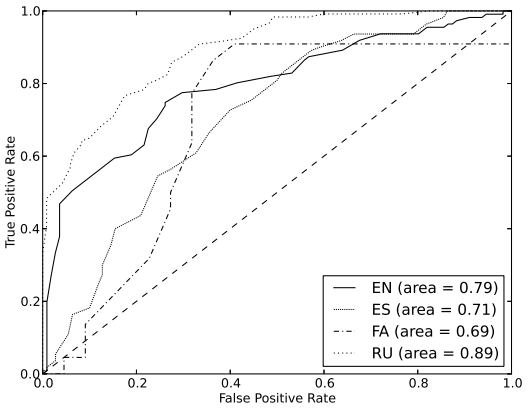
Table 4: Comparing AN metaphor detection method to the baselines: accuracy of the 10-fold cross validation on annotations of five human judges.

Experimental results for all four languages, are given in Figure 2. The ROC curves for SVO and AN tasks are plotted in Figure 2a and Figure 2b, respectively. Each curve corresponds to a test set described in Table 1. In addition, we perform an oracle experiment, to obtain actual f-score values for best thresholds. Detailed results are shown in Table 5.

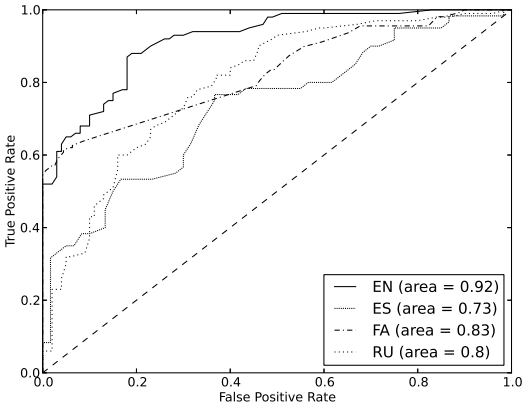
Consistent results with high f-scores are obtained across all four languages. Note that higher scores are obtained for the Russian test set. We hypothesize that this happens due to a higher-quality translation dictionary (which allows a more accurate model transfer). Relatively lower (yet reasonable) results for Farsi can be explained by a smaller size of the bilingual dictionary (thus, fewer feature projections can be obtained). Also note that, in our experience, most of Farsi metaphors are adjective-noun constructions. This is why the AN FA dataset in Table 1 is significantly larger than SVO FA. In that, for the AN Farsi task we observe high performance scores.

Figure 2 and Table 5 confirm, that *we obtain similar, robust results on four very different languages, using the same English classifiers*. We view this result as a strong evidence of language-independent nature of our metaphor detection method. In particular, this shows that proposed conceptual features can be used to detect selectional preferences violation across languages.

To summarize the experimental section, our metaphor detection approach obtains state-of-the-



(a) **SVO**



(b) **AN**

Figure 2: Cross-lingual experiment: ROC curves for classifiers trained on the English data using a combination of all features, and applied to SVO and AN metaphoric and literal relations in four test languages: English, Russian, Spanish, and Farsi.

art performance in English, is effective when applied to out-of-domain English data, and works cross-lingually.

5.4 Examples

Manual data analysis on adjective-noun pairs supports an abstractness-concreteness hypothesis formulated by several independent research studies. For example, in English we classify as metaphoric *dirty word* and *cloudy future*. Word pairs *dirty diaper* and *cloudy weather* have same adjectives. Yet they are classified as literal. Indeed, *diaper* is a more concrete term than *word* and *weather* is more concrete than *future*. Same pattern is observed in non-English datasets. In Russian, больное общество "sick society" and пустой звук "empty sound" are classified as metaphoric, while

	SVO	AN
EN	0.79	0.85
RU	0.84	0.77
ES	0.76	0.72
FA	0.75	0.74

Table 5: Cross-lingual experiment: f-scores for classifiers trained on the English data using a combination of all features, and applied, with optimal thresholds, to SVO and AN metaphoric and literal relations in four test languages: English, Russian, Spanish, and Farsi.

больная бабушка "sick grandmother" and пустая чашка "empty cup" are classified as literal. Spanish example of an adjective-noun metaphor is a well-known *músculo económico* "economic muscle". We also observe that non-metaphoric adjective noun pairs tend to have more imageable adjectives, such as literal *derecho humano* "human right". In Spanish, *human* is more imageable than *economic*.

Verb-based examples that are correctly classified by our model are: *blunder escaped notice* (metaphoric) and *prisoner escaped jail* (literal). We hypothesize that supersense features are instrumental in the correct classification of these examples: <*noun.person,verb.motion*> is usually used literally, while <*noun.act,verb.motion*> is used metaphorically.

6 Related Work

For a historic overview and a survey of common approaches to metaphor detection, we refer the reader to recent reviews by Shutova et al. (Shutova, 2010; Shutova et al., 2013). Here we focus only on recent approaches.

Shutova et al. (2010) proposed a bottom-up method: one starts from a set of seed metaphors and seeks phrases where verbs and/or nouns belong to the same cluster as verbs or nouns in seed examples.

Turney et al. (2011) show how abstractness scores could be used to detect metaphorical AN phrases. Neuman et al. (2013) describe a Concrete Category Overlap algorithm, where co-occurrence statistics and Turney's abstractness scores are used to determine WordNet supersenses that correspond to literal usage of a given adjective or verb. For example, given an adjective, we can learn that it modifies concrete nouns that usually have the

supersense *noun.body*. If this adjective modifies a noun with the supersense *noun.feeling*, we conclude that a metaphor is found.

Broadwell et al. (2013) argue that metaphors are highly imageable words that do not belong to a discussion topic. To implement this idea, they extend MRC imageability scores to all dictionary words using links among WordNet supersenses (mostly hypernym and hyponym relations). Strzalkowski et al. (2013) carry out experiments in a specific (government-related) domain for four languages: English, Spanish, Farsi, and Russian. Strzalkowski et al. (2013) explain the algorithm only for English and say that is the same for Spanish, Farsi, and Russian. Because they heavily rely on WordNet and availability of imageability scores, their approach may not be applicable to low-resource languages.

Hovy et al. (2013) applied tree kernels to metaphor detection. Their method also employs WordNet supersenses, but it is not clear from the description whether WordNet is essential or can be replaced with some other lexical resource. We cannot compare directly our model with this work because our classifier is restricted to detection of only SVO and AN metaphors.

Tsvetkov et al. (2013) propose a cross-lingual detection method that uses only English lexical resources and a dependency parser. Their study focuses only on the verb-based metaphors. Tsvetkov et al. (2013) employ only English and Russian data. Current work builds on this study, and incorporates new syntactic relations as metaphor candidates, adds several new feature sets and different, more reliable datasets for evaluating results. We demonstrate results on two new languages, Spanish and Farsi, to emphasize the generality of the method.

A words sense disambiguation (WSD) is a related problem, where one identifies meanings of polysemous words. The difference is that in the WSD task, we need to select an already existing sense, while for the metaphor detection, the goal is to identify cases of sense borrowing. Studies showed that cross-lingual evidence allows one to achieve a state-of-the-art performance in the WSD task, yet, most cross-lingual WSD methods employ parallel corpora (Navigli, 2009).

7 Conclusion

The key contribution of our work is that we show how to identify metaphors across languages by building a model in English and applying it—without adaptation—to other languages: Spanish, Farsi, and Russian. This model uses language-independent (rather than lexical or language specific) conceptual features. Not only do we establish benchmarks for Spanish, Farsi, and Russian, but we also achieve state-of-the-art performance in English. In addition, we present a comparison of relative contributions of several types of features. We concentrate on metaphors in the context of two kinds of syntactic relations: subject-verb-object (SVO) relations and adjective-noun (AN) relations, which account for a majority of all metaphorical phrases.

Future work will expand the scope of metaphor identification by including nominal metaphoric relations as well as explore techniques for incorporating contextual features, which can play a key role in identifying certain kinds of metaphors. Second, cross-lingual model transfer can be improved with more careful cross-lingual feature projection.

Acknowledgments

We are extremely grateful to Shuly Wintner for a thorough review that helped us improve this draft; we also thank people who helped in creating the datasets and/or provided valuable feedback on this work: Ed Hovy, Vlad Niculae, Davida Fromm, Brian MacWhinney, Carlos Ramírez, and other members of the CMU METAL team. This work was supported by the U.S. Army Research Laboratory and the U.S. Army Research Office under contract/grant number W911NF-10-1-0533.

References

Marco Baroni, Raffaella Bernardi, Ngoc-Quynh Do, and Chung-chieh Shan. 2012. Entailment above the word level in distributional semantics. In *Proc. of EACL*, pages 23–32.

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

Leo Breiman. 2001. Random forests. *Machine Learning*, 45(1):5–32.

George Aaron Broadwell, Umit Boz, Ignacio Cases, Tomek Strzalkowski, Laurie Feldman, Sarah Taylor, Samira Shaikh, Ting Liu, Kit Cho, and Nick Webb. 2013. Using imageability and topic chaining to locate metaphors in linguistic corpora. In *Social Computing, Behavioral-Cultural Modeling and Prediction*, pages 102–110. Springer.

Massimiliano Ciaramita and Yasemin Altun. 2006. Broad-coverage sense disambiguation and information extraction with a supersense sequence tagger. In *Proc. of EMNLP*, pages 594–602.

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

Manaal Faruqui and Chris Dyer. 2014. Improving vector space word representations using multilingual correlation. In *Proc. of EACL*. Association for Computational Linguistics.

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

Lisa Gandy, Nadji Allan, Mark Atallah, Ophir Frieder, Newton Howard, Sergey Kanareykin, Moshe Koppel, Mark Last, Yair Neuman, and Shlomo Argamon. 2013. Automatic identification of conceptual metaphors with limited knowledge. In *Proc. of the Twenty-Seventh AAAI Conference on Artificial Intelligence*, pages 328–334.

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.

Birgit Hamp and Helmut Feldweg. 1997. Germanet-a lexical-semantic net for German. In *Proc. of ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*, pages 9–15.

Dirk Hovy, Shashank Srivastava, Sujay Kumar Jauhar, Mrinmaya Sachan, Kartik Goyal, Huiying Li, Whitney Sanders, and Eduard Hovy. 2013. Identifying metaphorical word use with tree kernels. In *Proc. of the First Workshop on Metaphor in NLP*, page 52.

Eric H. Huang, Richard Socher, Christopher D. Manning, and Andrew Y. Ng. 2012. Improving word representations via global context and multiple word prototypes. In *Proc. of ACL*, pages 873–882.

Franz Hundsnurscher and Jochen Splett. 1982. *Semantik der Adjektive des Deutschen*. Number 3137. Westdeutscher Verlag.

Mikhail Kozhenikov and Ivan Titov. 2013. Cross-lingual transfer of semantic role labeling models. In *Proc. of ACL*, pages 1190–1200.

Saisuresh Krishnakumaran and Xiaojin Zhu. 2007. Hunting elusive metaphors using lexical resources. In *Proc. 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.

Lori Levin, Teruko Mitamura, Davida Fromm, Brian MacWhinney, Jaime Carbonell, Weston Feely, Robert Frederking, Anatole Gershman, and Carlos Ramirez. 2014. Resources for the detection of conventionalized metaphors in four languages. In *Proc. of LREC*.

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 *Proc. of ENMLP*, pages 34–44.

Zachary J Mason. 2004. CorMet: a computational, corpus-based conventional metaphor extraction system. *Computational Linguistics*, 30(1):23–44.

Ryan McDonald, Slav Petrov, and Keith Hall. 2011. Multi-source transfer of delexicalized dependency parsers. In *Proc. of EMNLP*.

Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. 2013. Exploiting similarities among languages for Machine Translation. *CoRR*, abs/1309.4168.

Roberto Navigli. 2009. Word sense disambiguation: A survey. *ACM Comput. Surv.*, 41(2):10:1–10:69, February.

Yair Neuman, Dan Assaf, Yohai Cohen, Mark Last, Shlomo Argamon, Newton Howard, and Ophir Frieder. 2013. Metaphor identification in large texts corpora. *PloS one*, 8(4):e62343.

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. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Claudia Perlich, Foster Provost, and Jeffrey S. Simonoff. 2003. Tree induction vs. logistic regression: a learning-curve analysis. *Journal of Machine Learning Research*, 4:211–255.

Nathan Schneider, Behrang Mohit, Chris Dyer, Kemal Oflazer, and Noah A Smith. 2013. Supersense tagging for Arabic: the MT-in-the-middle attack. In *Proc. of NAACL-HLT*, pages 661–667.

Ekaterina Shutova and Lin Sun. 2013. Unsupervised metaphor identification using hierarchical graph factorization clustering. In *Proc. of NAACL-HLT*, pages 978–988.

Ekaterina Shutova and Simone Teufel. 2010. Metaphor corpus annotated for source-target domain mappings. In *Proc. of LREC*, pages 3255–3261.

Ekaterina Shutova, Lin Sun, and Anna Korhonen. 2010. Metaphor identification using verb and noun clustering. In *Proc. of COLING*, pages 1002–1010.

Ekaterina Shutova, Simone Teufel, and Anna Korhonen. 2013. Statistical metaphor processing. *Computational Linguistics*, 39(2):301–353.

Ekaterina Shutova. 2010. Models of metaphor in NLP. In *Proc. of ACL*, pages 688–697.

Gerard J Steen, Aletta G Dorst, J Berenike Herrmann, Anna A Kaal, and Tina Krennmayr. 2010. Metaphor in usage. *Cognitive Linguistics*, 21(4):765–796.

Tomek Strzalkowski, George Aaron Broadwell, Sarah Taylor, Laurie Feldman, Boris Yamrom, Samira Shaikh, Ting Liu, Kit Cho, Umit Boz, Ignacio Cases, et al. 2013. Robust extraction of metaphors from novel data. In *Proc. of the First Workshop on Metaphor in NLP*, page 67.

Oscar Täckström, Dipanjan Das, Slav Petrov, Ryan McDonald, and Joakim Nivre. 2013. Token and type constraints for cross-lingual part-of-speech tagging. *TACL*, 1:1–12.

Paul H Thibodeau and Lera Boroditsky. 2011. Metaphors we think with: The role of metaphor in reasoning. *PLoS One*, 6(2):e16782.

Yulia Tsvetkov, Elena Mukomel, and Anatole Gershman. 2013. Cross-lingual metaphor detection using common semantic features. In *The 1st Workshop on Metaphor in NLP 2013*, page 45.

Yulia Tsvetkov, Nathan Schneider, Dirk Hovy, Archna Bhatia, Manaal Faruqui, and Chris Dyer. 2014. Augmenting English adjective senses with supersenses. In *Proc. of LREC*.

Joseph Turian, Lev Ratinov, Yoshua Bengio, and Dan Roth. 2009. A preliminary evaluation of word representations for named-entity recognition. In *NIPS Workshop on Grammar Induction, Representation of Language and Language Learning*, pages 1–8.

Joseph Turian, Lev Ratinov, and Yoshua Bengio. 2010. Word representations: a simple and general method for semi-supervised learning. In *Proc. of ACL*, pages 384–394.

Peter D. Turney, Yair Neuman, Dan Assaf, and Yohai Cohen. 2011. Literal and metaphorical sense identification through concrete and abstract context. In *Proc. of EMNL*, pages 680–690.

Yorick Wilks. 1978. Making preferences more active. *Artificial Intelligence*, 11(3):197–223.

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