Semantic classifications for detection of verb metaphors

Beata Beigman Klebanov[1] and **Chee Wee Leong**[1] and **Elkin Dario Gutierrez**[2]
and **Ekaterina Shutova**[3] and **Michael Flor**[1]
[1] Educational Testing Service
[2] University of California, San Diego
[3] University of Cambridge
{bbeigmanklebanov,cleong,mflor}@ets.org
edg@icsi.berkeley.edu, ekaterina.shutova@cl.cam.ac.uk

Abstract

We investigate the effectiveness of semantic generalizations/classifications for capturing the regularities of the behavior of verbs in terms of their metaphoricity. Starting from orthographic word unigrams, we experiment with various ways of defining semantic classes for verbs (grammatical, resource-based, distributional) and measure the effectiveness of these classes for classifying all verbs in a running text as metaphor or non metaphor.

1 Introduction

According to the Conceptual Metaphor theory (Lakoff and Johnson, 1980), metaphoricity is a property of concepts in a particular context of use, not of specific words. The notion of a concept is a fluid one, however. While *write* and *wrote* would likely constitute instances of the same concept according to any definition, it is less clear whether *eat* and *gobble* would. Furthermore, the Conceptual Metaphor theory typically operates with whole semantic domains that certainly generalize beyond narrowly-conceived concepts; thus, *save* and *waste* share a very general semantic feature of applying to finite resources – it is this meaning element that accounts for the observation that they tend to be used metaphorically in similar contexts.

In this paper, we investigate which kinds of generalizations are the most effective for capturing regularities of metaphor usage.

2 Related Work

Most previous supervised approaches to verb metaphor classification evaluated their systems on selected examples or in small-scale experiments (Tsvetkov et al., 2014; Heintz et al., 2013; Turney et al., 2011; Birke and Sarkar, 2007; Gedigan et al., 2006), rather than using naturally occurring continuous text, as done here. Beigman Klebanov et al. (2014) and Beigman Klebanov et al. (2015) are the exceptions, used as a baseline in the current paper.

Features that have been used so far in supervised metaphor classification address concreteness and abstractness, topic models, orthographic unigrams, sensorial features, semantic classifications using WordNet, among others (Beigman Klebanov et al., 2015; Tekiroglu et al., 2015; Tsvetkov et al., 2014; Dunn, 2014; Heintz et al., 2013; Turney et al., 2011). Of the feature sets presented in this paper, all but WordNet features are novel.

3 Semantic Classifications

In the following subsections, we describe the different types of semantic classifications; Table 1 summarizes the feature sets.

Name	Description	#Features
U	orthographic unigram	*varies*
UL	lemma unigram	*varies*
VN-Raw	VN frames	270
VN-Pred	VN predicate	145
VN-Role	VN thematic role	30
VN-RoRe	VN them. role filler	128
WordNet	WN lexicographer files	15
Corpus	distributional clustering	150

Table 1: Summary of feature sets. All features are binary features indicating class membership.

3.1 Grammar-based

The most minimal level of semantic generalization is that of putting together verbs that share the same lemma (lemma unigrams, **UL**). We use NLTK (Bird et al., 2009) for identifying verb lemmas.

Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pages 101–106,
Berlin, Germany, August 7-12, 2016.

3.2 Resource-based

VerbNet: The VerbNet database (Kipper et al., 2006) provides a classification of verbs according to their participation in *frames* – syntactic patterns with semantic components, based on Levin's classes (Levin, 1993). Each verb class is annotated with its member verb lemmas, syntactic constructions in which these participate (such as transitive, intransitive, diathesis alternations), semantic predicates expressed by the verbs in the class (such as motion or contact), thematic roles (such as agent, patient, instrument), and restrictions on the fillers of these semantic roles (such as pointed instrument).

VerbNet can thus be thought of as providing a number of different classifications over the same set of nearly 4,000 English verb lemmas. The main classification is based on syntactic frames, as enacted in VerbNet classes. We will refer to them as **VN-Raw** classes. An alternative classification is based on the predicative meaning of the verbs; for example, the verbs *assemble* and *introduce* are in different classes based on their syntactic behavior, but both have the meaning component of *together*, marked in VerbNet as a possible value of the Predicate variable. Similarly, *shiver* and *faint* belong to different VerbNet classes in terms of syntactic behavior, but both have the meaning element of describing an *involuntary* action. Using the different values of the Predicate variable, we created a set of **VN-Pred** classes. We note that the same verb lemma can occur in multiple classes, since different senses of the same lemma can have different meanings, and even a single sense can express more than one predicate. For example, the verb *stew* participates in the following classes of various degrees of granularity: *cause* (shared with 2,912 other verbs), *use* (with 700 other verbs), *apply heat* (with 49 other verbs), *cooked* (with 49 other verbs).

Each VerbNet class is marked with the thematic roles its members take, such as *agent* or *beneficiary*. Here again, verbs that differ in syntactic behavior and in the predicate they express could share thematic roles. For example, *stew* and *prick* belong to different VerbNet classes and share only the most general predicative meanings of *cause* and *use*, yet both share a thematic role of *instrument*. We create a class for each thematic role (**VN-Role**).

Finally, VerbNet provides annotations of the restrictions that apply to fillers of various thematic roles. For example, verbs that have a thematic role of *instrument* can have the filler restricted to being inanimate, body part, concrete, pointy, solid, and others. Across the various VerbNet classes, there are 128 restricted roles (such as *instrument_pointy*). We used those to generate **VN-RoRe** classes.

WordNet: We use lexicographer files to classify verbs into 15 classes based on their general meaning, such as verbs of communication, consumption, weather, and so on.

3.3 Corpus-based

We also experimented with automatically-generated verb clusters as semantic classes. We clustered VerbNet verbs using a spectral clustering algorithm and lexico-syntactic features. We selected the verbs that occur more than 150 times in the British National Corpus, 1,610 in total, and clustered them into 150 clusters (**Corpus**).

We used verb subcategorization frames (SCF) and the verb's nominal arguments as features for clustering, as they have proved successful in previous verb classification experiments (Shutova et al., 2010). We extracted our features from the Gigaword corpus (Graff et al., 2003) using the SCF classification system of Preiss et al. (2007) to identify verb SCFs and the RASP parser (Briscoe et al., 2006) to extract the verb's nominal arguments.

Spectral clustering partitions the data relying on a similarity matrix that records similarities between all pairs of data points. We use *Jensen-Shannon divergence* (d_{JS}) to measure similarity between feature vectors for two verbs, v_i and v_j, and construct a similarity matrix S_{ij}:

$$S_{ij} = \exp(-d_{\mathrm{JS}}(v_i, v_j)) \quad (1)$$

The matrix S encodes a similarity graph G over our verbs. The clustering problem can then be defined as identifying the optimal partition, or *cut*, of the graph into clusters. We use the multiway normalized cut (MNCut) algorithm of Meila and Shi (2001) for this purpose. The algorithm transforms S into a stochastic matrix P containing transition probabilities between the vertices in the graph as $P = D^{-1}S$, where the degree matrix D is a diagonal matrix with $D_{ii} = \sum_{j=1}^{N} S_{ij}$. It then computes the K leading eigenvectors of P, where K is the desired number of clusters. The graph is partitioned by finding approximately equal elements

in the eigenvectors using a simpler clustering algorithm, such as *k-means*. Meila and Shi (2001) have shown that the partition I derived in this way minimizes the MNCut criterion:

$$\text{MNCut}(I) = \sum_{k=1}^{K}[1 - P(I_k \rightarrow I_k|I_k)], \quad (2)$$

which is the sum of transition probabilities across different clusters. Since k-means starts from a random cluster assignment, we ran the algorithm multiple times and used the partition that minimizes the cluster distortion, that is, distances to cluster centroid.

We tried expanding the coverage of VerbNet verbs and the number of clusters using grid search on the training data, with coverage grid ={2,500; 3,000; 4,000} and #clusters grid = {200; 250; 300; 350; 400}, but obtained no improvement in performance over our original setting.

4 Experiment setup

4.1 Data

We use the VU Amsterdam Metaphor Corpus (Steen et al., 2010).[1] The corpus contains annotations of all tokens in running text as metaphor or non metaphor, according to a protocol similar to MIP (Pragglejaz, 2007). The data come from the BNC, across 4 genres: news (**N**), academic writing (**A**), fiction (**F**), and conversation (**C**). We address each genre separately. We consider all verbs apart from *have*, *be*, and *do*.

We use the same training and testing partitions as Beigman Klebanov et al. (2015). Table 2 summarizes the data. [2]

Data	Training			Testing	
	#T	#I	%M	#T	#I
News	49	3,513	42%	14	1,230
Fict.	11	4,651	25%	3	1,386
Acad.	12	4,905	31%	6	1,260
Conv.	18	4,181	15%	4	2,002

Table 2: Summary of the data. #T = # of texts; #I = # of instances; %M = percentage of metaphors.

4.2 Machine Learning Methods

Our setting is that of supervised machine learning for binary classification. We experimented with a number of classifiers using VU-News training data, including those used in relevant prior work: Logistic Regression (Beigman Klebanov et al., 2015), Random Forest (Tsvetkov et al., 2014), Linear Support Vector Classifier. We found that Logistic Regression was better for unigram features, Random Forest was better for features using WordNet and VerbNet classifications, whereas the corpus-based features yielded similar performance across classifiers. We therefore ran all evaluations with both Logistic Regression and Random Forest classifiers. We use the skll and scikit-learn toolkits (Blanchard et al., 2013; Pedregosa et al., 2011). During training, each class is weighted in inverse proportion to its frequency. The optimization function is F1 (metaphor).

5 Results

We first consider the performance of each type of semantic classification separately as well as various combinations using cross-validation on the training set. Table 3 shows the results with the classifier that yields the best performance for the given feature set.

Name	N	F	A	C	Av.
U	.64	.51	.55	.39	.52
UL	.65	.51	.61	.39	.54
VN-Raw	.64	.49	.60	.38	.53
VN-Pred	.62	.47	.58	.39	.52
VN-Role	.61	.46	.55	.40	.50
VN-RoRe	.59	.47	.54	.36	.49
WN	.64	.50	.60	.38	.53
Corpus	.59	.49	.53	.36	.49
VN-RawToCorpus	.63	.49	.59	.38	.53
UL+WN	.67	.52	.63	.40	.56
UL+Corpus	.66	.53	.62	.39	.55

Table 3: Performance (F1) of each of the feature sets, xval on training data. U = unigram baseline.

Of all types of semantic classification, only the grammatical one (lemma unigrams, UL) shows an overall improvement over the unigram baseline with no detriment for any of the genres.VN-Raw and WordNet show improved performance for Academic but lower performance on Fiction than the unigram baseline. Other versions of VerbNet-based semantic classifications are generally worse than VN-Raw, with some exceptions for the Conversation genre. Distributional clusters (Corpus) generally perform worse than the resource-based classifications, even when the resource is restricted to the exact same set of verbs as that covered in the Corpus clusters (compare Corpus to VN-RawToCorpus).

The distributional features are, however, about as effective as WordNet features when combined

[1] available at http://metaphorlab.org/metcor/search/

[2] Data and features will be made available at https://github.com/EducationalTestingService/metaphor.

with the lemma unigrams (UL); the combinations improve the performance over UL alone for every genre. We also note that the better performance for these combinations is generally attained by the Logistic Regression classifier. We experimented with additional combinations of feature sets, but observed no further improvements.

To assess the consistency of metaphoricity behavior of semantic classes across genres, we calculated correlations between the weights assigned by the UL+WN model to the 15 WordNet features. All pairwise correlations between News, Academic, and Fiction were strong ($r > 0.7$), while Conversation had low to negative correlation with other genres. The low correlations with Conversation was largely due to a highly discrepant behavior of verbs of weather[3] – these are consistently used metaphorically in all genres apart from Conversation. This discrepancy, however, is not so much due to genre-specific differences in behavior of the same verbs as to the difference in the identity of the weather verbs that occur in the data from the different genres. While *burn*, *pour*, *reflect*, *fall* are common in the other genres, the most common weather verb in Conversation is *rain*, and none of its occurrences is metaphoric; its single occurrence in the other genres is likewise not metaphoric. More than a difference across genres, this case underscores the complementarity of lemma-based and semantic class-based information – it is possible for weather verbs to tend towards metaphoricity as a class, yet some verbs might not share the tendency – verb-specific information can help correct the class-based pattern.

5.1 Blind Test Benchmark

To compare the results against state-of-art, we show the performance of Beigman Klebanov et al. (2015) system (**SOA'15**) on the test data (see Table 2 for the sizes of the test sets per genre). Their system uses Logistic Regression classifier and a set of features that includes orthographic unigrams, part of speech tags, concreteness, and difference in concreteness between the verb and its direct object. Against this benchmark, we evaluate the performance of the best combination identified during the cross-validation runs, namely, UL+WN feature set using Logistic Regression classifier. We also show the performance of the resource-lean model, UL+Corpus. The top three rows of Table 4 show the results. The UL+WN model outperforms the state of art for every genre; the improvement is statistically significant ($p<0.05$).[4] The improvement of UL+Corpus over SOA'15 is not significant.

Following the observation of the similarity between weights of semantic class features across genres, we also trained the three systems on all the available training data across all genres (all data in the Train column in Table 2), and tested on test data for the specific genre. This resulted in performance improvements for all systems in all genres, including Conversation (see the bottom 3 rows in Table 4). The significance of the improvement of UL+WN over SOA'15 was preserved; UL+Corpus now significantly outperformed SOA'15.

	Feature Set	N	F	A	C	Av.
Train	SOA'15	.64	.47	.71	.43	.56
in	UL+WN	.68	.49	.72	.44	.58
genre	UL+Corpus	.65	.49	.71	.43	.57
Train	SOA'15	.66	.48	.74	.44	.58
on all	UL+WN	.69	.50	.77	.45	.60
genres	UL+Corpus	.67	.51	.76	.45	.60

Table 4: Benchmark performance, F1 score.

6 Conclusion

The goal of this paper was to investigate the effectiveness of semantic generalizations/classifications for metaphoricity classification of verbs. We found that generalization from orthographic unigrams to lemmas is effective. Further, lemma unigrams and semantic class features based on WordNet combine effectively, producing a significant improvement over the state of the art. We observed that semantic class features were weighted largely consistently across genres; adding training data from other genres is helpful. Finally, we found that a resource-lean model where lemma unigram features were combined with clusters generated automatically using a large corpus yielded a competitive performance. This latter result is encouraging, as the knowledge-lean system is relatively easy to adapt to a new domain or language.

[3] Removing verbs of weather propelled the correlations with Conversation to a moderate range, r = 0.25-0.45 across genres.

[4] We used McNemar's test of significance of difference between correlated proportions (McNemar, 1947), 2-tailed. We combined data from all genres into on a 2X2 matrix: both SOA'15 and UL+WN correct in (1,1), both wrong (0,0), SOA'15 correct UL+WN wrong (0,1), UL+WN correct SOA'15 wrong (1,0)).

Acknowledgment

We are grateful to the ACL reviewers for their helpful feedback. Ekaterina Shutova's research is supported by the Leverhulme Trust Early Career Fellowship.

References

Beata Beigman Klebanov, Chee Wee Leong, Michael Heilman, and Michael Flor. 2014. Different texts, same metaphors: Unigrams and beyond. In *Proceedings of the Second Workshop on Metaphor in NLP*, pages 11–17, Baltimore, MD, June. Association for Computational Linguistics.

Beata Beigman Klebanov, Chee Wee Leong, and Michael Flor. 2015. Supervised word-level metaphor detection: Experiments with concreteness and reweighting of examples. In *Proceedings of the Third Workshop on Metaphor in NLP*, pages 11–20, Denver, Colorado, June. Association for Computational Linguistics.

Steven Bird, Edward Loper, and Ewan Klein. 2009. *Natural Language Processing with Python*. OReilly Media Inc.

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*, pages 21–28, Rochester, New York.

Daniel Blanchard, Michael Heilman, and Nitin Madnani. 2013. *SciKit-Learn Laboratory*. GitHub repository, https://github.com/EducationalTestingService/skll.

E. Briscoe, J. Carroll, and R. Watson. 2006. The second release of the rasp system. In *Proceedings of the COLING/ACL on Interactive presentation sessions*, pages 77–80.

Jonathan Dunn. 2014. Multi-dimensional abstractness in cross-domain mappings. In *Proceedings of the Second Workshop on Metaphor in NLP*, pages 27–32, Baltimore, MD, June. Association for Computational Linguistics.

M. Gedigan, J. Bryant, S. Narayanan, and B. Ciric. 2006. Catching metaphors. In *PProceedings of the 3rd Workshop on Scalable Natural Language Understanding*, pages 41–48, New York.

D. Graff, J. Kong, K. Chen, and K. Maeda. 2003. English Gigaword. *Linguistic Data Consortium, Philadelphia*.

Ilana Heintz, Ryan Gabbard, Mahesh Srivastava, Dave Barner, Donald Black, Majorie Friedman, and Ralph Weischedel. 2013. Automatic Extraction of Linguistic Metaphors with LDA Topic Modeling. In *Proceedings of the First Workshop on Metaphor in NLP*, pages 58–66, Atlanta, Georgia, June. Association for Computational Linguistics.

Karin Kipper, Anna Korhonen, Neville Ryant, and Martha Palmer. 2006. Extensive classifications of english verbs. In *Proceedings of the 12th EURALEX International Congress*, Turin, Italy, September.

George Lakoff and Mark Johnson. 1980. *Metaphors we live by*. University of Chicago Press, Chicago.

Beth Levin. 1993. *English Verb Classes and Alternations: A Preliminary Investigation*. Chicago, IL: University of Chicago Press.

Quinn McNemar. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2).

M. Meila and J. Shi. 2001. A random walks view of spectral segmentation. In *Proceedings of AISTATS*.

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.

Group Pragglejaz. 2007. MIP: A Method for Identifying Metaphorically Used Words in Discourse. *Metaphor and Symbol*, 22(1):1–39.

Judita Preiss, Ted Briscoe, and Anna Korhonen. 2007. A system for large-scale acquisition of verbal, nominal and adjectival subcategorization frames from corpora. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 912–919, Prague, Czech Republic, June. Association for Computational Linguistics.

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 (COLING)*, pages 1002–1010.

Gerard Steen, Aletta Dorst, Berenike Herrmann, Anna Kaal, Tina Krennmayr, and Trijntje Pasma. 2010. *A Method for Linguistic Metaphor Identification*. Amsterdam: John Benjamins.

Serra Sinem Tekiroglu, Gözde Özbal, and Carlo Strapparava. 2015. Exploring sensorial features for metaphor identification. In *Proceedings of the Third Workshop on Metaphor in NLP*, pages 31–39, Denver, Colorado, June. Association for Computational Linguistics.

Yulia Tsvetkov, Leonid Boytsov, Anatole Gershman, Eric Nyberg, and Chris Dyer. 2014. Metaphor detection with cross-lingual model transfer. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 248–258, Baltimore, Maryland, June. Association for Computational Linguistics.

Peter 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*, pages 680–690, Stroudsburg, PA, USA. Association for Computational Linguistics.