Metaphor Detection in Discourse

Hyeju Jang, Seunghwan Moon, Yohan Jo, and Carolyn Penstein Rosé
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA 15213, USA
{hyejuj, seungwhm, yohanj, cprose}@cs.cmu.edu

Abstract

Understanding contextual information is key to detecting metaphors in discourse. Most current work aims at detecting metaphors given a single sentence, thus focusing mostly on *local* contextual cues within a short text. In this paper, we present a novel approach that explicitly leverages *global context* of a discourse to detect metaphors. In addition, we show that syntactic information such as dependency structures can help better describe local contextual information, thus improving detection results when combined. We apply our methods on a newly annotated online discussion forum, and show that our approach outperforms the state-of-the-art baselines in previous literature.

1 Introduction

Detecting metaphors in text is an active line of research which has attracted attention in recent years. To date, most of the previous literature has looked at lexical semantic features such as selectional restriction violations (Martin, 1996; Shutova and Teufel, 2010; Shutova et al., 2010; Shutova et al., 2013; Huang, 2014) or contrast in lexical concreteness and abstractness (Turney et al., 2011; Broadwell et al., 2013; Tsvetkov et al., 2013). While these approaches have been shown to be successful in detecting metaphors given a single sentence, metaphor detection in discourse brings a new dimension to the task. Consider the following excerpt from an online *Breast Cancer* discussion forum as an example:

> *welcome, glad for the company .... just sad to see that there are so many of us. Here is a thought that I have been thinking since I was diagnosed. This disease should be called the "Hurry up and Wait" illness. Since the day I heard the dreaded words "you need to have a biopsy", I feel like I am on a speeding train. It rushes into every station where you have to make instant decisions, while this ominous clock is ticking. Wait for test results, wait for appointments, wait for healing.*

In the example above, it is difficult to identify *"rushes into every station"* as a metaphorical expression using the previous approaches, because it does not violate selectional restrictions or have any notable contrast in lexical concreteness and abstractness. The reason for this is clear: the action of *rushing into stations* itself makes perfect sense literally when it is viewed *locally* as an isolated phrase, while the contextual cues for this metaphor are embedded globally throughout the discourse (*e.g. diagonsed*, *disease*, *biopsy* are semantically contrasted with *train*, *rushes*, and *station*). This clearly demonstrates the need for a new set of computational tools to represent context beyond a single sentence, in order to better detect metaphorical expressions that have contextual connections outside the sentence in which they are used.

Context for metaphor detection. Metaphor is a semantic phenomenon that describes objects often with a view borrowed from a different domain. As such, it is natural that metaphors inherently break the lexical coherence of a sentence or a discourse. Klebanov et al. (2009), for example, showed in their study that words related to the topic of discussion are less likely to be metaphorical than other words in text, implying that contextual incoherence might serve as a cue for detecting metaphors. Based on this observation, the idea of leveraging textual context to detect metaphors has been recently proposed by some researchers (Broadwell et al., 2013; Sporleder and Li, 2009).

Proceedings of the SIGDIAL 2015 Conference, pages 384–392,
Prague, Czech Republic, 2-4 September 2015.

Our contributions. We extend the previous approaches for detecting metaphors by explicitly addressing the *global* discourse context, as well as by representing the *local* context of a sentence in a more robust way. Our contribution is thus twofold: first, we propose several textual descriptors that can capture global contextual shifts among a discourse, such as semantic word category distribution obtained from a frame-semantic parser, homogeneity in topic distributions, and lexical chains. Second, we show that global and local contextual information are complimentary in detecting metaphors, and that leveraging syntactic features is crucial in better describing lexico-semantic information in a local context. Our method achieves higher performance on a metaphor disambiguation task than state-of-the-art systems from prior work (Klebanov et al., 2014; Tsvetkov et al., 2013) on our newly created dataset from an online discussion forum.

The rest of the paper is organized as follows. Section 2 relates our work to prior work. Section 3 explains our method in detail, specifically in regards to how we use global context and local context for metaphor detection. Section 4 describes the *Breast Cancer* dataset annotated and used for our experiment. In Section 5, we present our experimental results and show the effectiveness of our method with the task of metaphor disambiguation. Section 6 analyzes the results and identifies potential areas of improvement, and we give our concluding remarks in Section 7.

2 Relation to Prior Work

The main approaches to computationally detecting metaphors can be categorized into work that considers the following three classes of features: selectional preferences, abstractness and concreteness, and lexical cohesion.

Selectional preferences relate to how semantically compatible predicates are with particular arguments. For example, the verb *drink* prefers *beer* as an object over *computer*. The idea behind using selectional preferences for metaphor detection is that metaphorical words tend to break selectional preferences. In the case of "*the clouds sailed across the sky*", for instance, *sailed* is determined to be metaphorically used because *clouds* as a subject violates its selectional restriction. The idea of using violation of selectional preferences as a cue for metaphors has been well studied in a variety of previous work (Martin, 1996; Shutova and Teufel, 2010; Shutova et al., 2010; Shutova et al., 2013; Huang, 2014) In general, this work can be further categorized into work that uses lexical resources and the work that uses corpus-based approaches to obtain selectional preferences

From the observations that metaphorical words (source domain) tend to use more concrete and imagination rich words than the target domain of metaphors, the **abstractness/concreteness** approaches computationally measure the degree of abstractness of words to detect metaphors. Take the following two phrases as examples that demonstrate this concept: *green idea* (metaphorical expression) and *green frog* (literal expression.) The former has a concrete word (*green*) modifying an abstract concept (*idea*), thus being more likely to be metaphorical. The idea of leveraging abstractness/concreteness in detecting metaphors has been proposed and studied by several groups of researchers (Turney et al., 2011; Broadwell et al., 2013; Tsvetkov et al., 2013; Assaf et al., 2013; Neuman et al., 2013). Note that most of this work uses datasets that comprise grammatically restricted sentences (*e.g.* ones with S+V+O or A+N structures) for their experiments, in order to test their hypothesis in a controlled way.

Another line of work considers **lexical coherence** of text as a cue for metaphor. The lexical coherence approach is motivated by the observation that metaphorical words are often semantically incoherent with context words. There have been several approaches proposed to compute lexical coherence. Broadwell et al. (2013), for instance, employed topic chaining to categorize metaphors, whereas Sporleder and Li (2009) have proposed to use lexical chains and semantic cohesion graphs to detect metaphors. Shutova and Sun (2013) and Shutova et al. (2013) have formulated the metaphor detection problem similar to outlier detection or anomaly detection tasks, and proposed to use topic signatures as lexical coherence features. Schulder and Hovy (2014) used TF-IDF to obtain domain term relevance, and applied this feature to detect metaphors.

Klebanov et al. (2014) propose to use various lexical features such as part-of-speech tags, concreteness ratings, and topic scores of target words to detect word-level metaphors in a running text. Our approach is different from theirs in that we explicitly gather global contextual information from

discourse to detect metaphors and that we leverage the syntactic structures to better represent local contextual information.

3 Our Method

In this section, we describe our method to measure nonliteralness of an expression in context. Specifically, we describe how we use contextual information as features for metaphor classification in discourse.

We first define *lexical cohesion* before we introduce our motivation and method for utilizing global contexts as features for detecting metaphor. A text is said to be lexically cohesive when the words in the text describe a single coherent topic. Specifically, lexical cohesion occurs when words are semantically related directly to a common topic or indirectly to the topic via another word. Figure 1 illustrates the lexical cohesion among words shown as a graph.

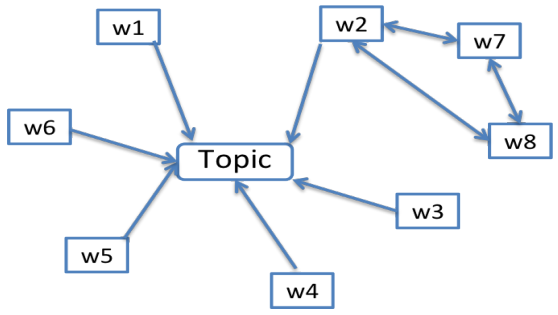

Figure 1: Graph representation depicting lexical cohesion among words in a given text. Edges represent lexical relatedness between a topic and a word or between words. For example, w_1 is directly related to the topic of discussion, whereas w_7 is only indirectly related to the topic through w_2.

The intuition for our main idea is that metaphorically-used words would often break lexical cohesion of text, while literal expressions would maintain a single connected graph of topically or semantically related words. Therefore, we identify that these incohesive words may serve as cues for nonliteral expressions. The following two examples illustrate the described phenomenon, both of which contain the same phrase *"break the ice"*.

> *... Meanwhile in Germany, the cold penetrated the vast interior of Cologne cathedral, where worshippers had to break the ice on holy water in the font. The death toll from the cold also increased ...*

> *... "Some of us may have have acted as critics at one point or another, but for the most part its just as filmgoers," he said. And, breaking the ice at a press conference, he praised his vice-president, French actress Catherine Deneuve ...*

The phrase *"break the ice"* in the first example is used with words such as *cold* and *water* which are semantically coherent with its literal meaning, whereas in the second example, the phrase is used with *press conference*, *praised*, and *vice-president*, which are far from the literal meaning of *break* and *ice*.

Note that this contextual information lies in different parts of a discourse, sometimes locally in the same sentence as the target word or globally throughout multiple surrounding sentences in a discourse. Given this observation, we categorize contextual information into two kinds depending on the scope of the context in text: *global* and *local*. Global contexts range over the whole document, whereas local contexts are limited to the sentence that contains the expression of interest. Section 3.1 explains how we represent global contexts. Section 3.2 describes the features we use for local contexts, and how we leverage syntactic information to make a more robust use of the semantic features in local context.

3.1 Global Contextual Features

We use the following features to represent global contexts of a given text.

Semantic Category: Lexico-semantic resources (*e.g.* `FrameNet`, `WordNet`) provide categorical information for much of the English lexicon. If a target word is used literally, the document may have a high proportion of words in the same semantic category. If the word is used metaphorically, the document may contain more words that share different semantic categories. To implement this intuition, we use `SEMAFOR` (Das et al., 2014) to assign each word to one of the categories provided by the `FrameNet 1.5` taxonomy (Baker et al., 1998). Then, we compute the relative proportion of the target word's category with regards to categories appearing in the document to measure the alignment of categories of the target word

and the surrounding contexts. Formally, we define the value of the *global word category feature* as

$$\frac{\sum_{w \in d} \mathbb{1}(c_w = c_{tw})}{N_d},$$

where c_w is the category of word w, c_{tw} is the category of the target word, and N_d is the number of words in document d. $\mathbb{1}(.)$ is an indicator function that equals 1 when the expression inside is true and 0 otherwise.

We have also used WordNet[1]'s 44 lexnames in our preliminary experiment to obtain word categories. However, we have found that its coarse categorization of words (44 categories as opposed to FrameNet's 1204) led to poorer performance, thus we have used FrameNet here instead.

Topic Distribution: Our intuition for using topic distributions is that non-literal words tend to have a considerably different topic distribution from that of the surrounding document (global context). To implement this idea, we run a topic model to obtain a word-topic distribution (= P(topic|word)) and document-topic distribution (= P(topic|document)). We use Latent Dirichlet Allocation (LDA) (Blei et al., 2003) to find 100 topics from the entire corpus, and calculate the topic distribution per document and the topic distribution per word from the trained topic model. Specifically, we begin by training our model for 2,000 iterations on a large data set. Then, for the estimation on test documents we apply this model to our test data set for 100 iterations of Gibbs sampling.

The original LDA computes P(word|topic) instead of P(topic|word). In order to compute P(topic|word), the first 20 iterations out of 100 are used as a burn-in phase, and then we collect sample topic assignments for each word in every other iteration. This process results in a total of 40 topic assignments for a word in a document, and we use these topic assignments to estimate the topic distributions per word as in (Remus and Biemann, 2013). We used the GibbsC++ toolkit (Phan and Nguyen, 2007) with default parameters to train the model.

Finally, we use the cosine similarity between P(topic|document) and P(topic|word) as features that represent the global alignment of topics between the target word and the document.

Lexical Chain: We use *lexical chains* (Morris and Hirst, 1991) to obtain multiple sequences of semantically related words in a text. From the intuition that metaphorical words would not belong to dominant lexical chains of the given text, we use the lexical chain membership of a target word as a cue for its non-literalness. Because each discourse instance of our dataset tends to be short and thus does not produce many lexical chains, we use a binary feature of whether a target word belongs to the longest chain of the given text. In our implementation, we use the ELKB toolkit (Jarmasz and Szpakowicz, 2003) to detect lexical chains in text which is built on Roget's thesaurus (Roget, 1911). Note that a similar approach has been used by Sporleder and Li (2009) to grasp topical words in a text.

Context Tokens: In addition, we use unigram features to represent the global context. Specifically, we use binary features to indicate whether the context words appeared anywhere in a given discourse.

3.2 Local Contextual Features

The local contextual information within a sentence is limited because it often contains fewer words, but the information could be more direct and richer because it reflects the immediate context of an expression of interest. We represent local contextual information using the semantic features listed below, combined with grammatical dependencies to induce relational connections between a target word and its contextual information.

Semantic Category: We follow the same intuition as using semantic categories to represent global features (Section 3.1), and thus compare the target word's semantic category and that of other words in the same sentence to induce local contextual information. However, since a sentence often has only a small number of words, the proportion of the target word's category in one sentence depends too much on the sentence length. Therefore, we instead look at the words that have dependency relations with the target word, and create nominal features by pairing word categories of lexical items with their dependency relations. The paired dependency-word category features specifies *how* local contextual words are used in relation to the target word, thus providing richer information. We also specify the target word's category as a categorical feature, expecting that the interplay between the target word's category and other words' categories is indicative of the non-literalness of the

[1] https://wordnet.princeton.edu/man/lexnames.5WN.html

target word.

Semantic Relatedness: If the semantic relatedness between a target word and the context words is low, the target word is likely to be metaphorically used. From the observation that the words that are in grammatical relation to the target word are more informative than other words, we use the dependency relations of a target word to pick out the words to compute semantic relatedness with. To represent the semantic relatedness between two words, we compute the cosine similarity of their topic distributions.

We use the semantic relatedness information in two different ways. One way is to compute average semantic relatedness over the words that have dependency relations with a target word, and use it as a feature (AvgSR). The other is to use semantic relatedness of the words in grammatical relations to the target word as multiple features (DepSR).

We use the same techniques as in Section 3.1 to compute topic distribution using an LDA topic model.

Lexical Abstractness/Concreteness: People often use metaphors to convey a complex or abstract thought by borrowing a word or phrase having a concrete concept that is easy to grasp. With this intuition, Turney et al. (2011) showed that the word abstractness/concreteness measure is a useful clue for detecting metaphors.

To represent the concreteness of a word, we used Brysbaert's database of concreteness ratings for about 40,000 English words (Brysbaert et al., 2014). We use the mean ratings in the database as a numerical feature for the target word. In addition, we also use the concreteness ratings of the words in grammatical relations to the target word as local context features.

Grammatical Dependencies: We use the `stanford-corenlp` toolkit (Manning et al., 2014) to parse dependency relations of our data and apply grammatical dependencies as described above for each semantic feature. We use grammatical dependencies only between content words (*e.g.* words with syntactic categories of noun, verb, adjective, and adverb).

4 Data

We conduct experiments on data acquired from discussion forums for an online breast cancer support group. The data contains all the public posts, users, and profiles on the discussion boards from October 2001 to January 2011. The dataset consists of 1,562,459 messages and 90,242 registered members. 31,307 users have at least one post, and the average number of posts per user is 24.

We built an annotated dataset for our experiments as follows. We first picked seven metaphor candidates that appear either metaphorically or literally in the *Breast Cancer* corpus: *boat*, *candle*, *light*, *ride*, *road*, *spice*, and *train*. We then retrieved all the posts in the corpus that contain these candidate words, and annotated each post as to whether the candidate word in the post is used metaphorically. When the candidate word occurs more than once in a single post, all occurrences within a post were assumed to have the same usage (either metaphorical or literal).

Note that our annotation scheme is different from the VU Amsterdam metaphor-annotated dataset (Steen et al., 2010) or the essay data used in (Klebanov et al., 2014), where every word in the corpus is individually labeled as a metaphor or a literal word. Our approach of pre-defining a set of metaphor candidate words and annotating each post as opposed to every word has several practical and fundamental benefits. First, metaphors often have a wide spectrum of "literalness" depending on how frequently they are used in everyday text, and there is a continuing debate as to how to operationalize metaphor in a binary decision (Jang et al., 2014). In our work, we can circumvent this metaphor decision issue by annotating a set of metaphor candidate words that have a clear distinction between metaphorical and literal usages. Second, our annotation only for ambiguous words ensures to focus on how well a model distinguishes between metaphorical and literal usage of the same word.

We employed Amazon Mechanical Turk (MTurk) workers to annotate metaphor use for candidate words. A candidate word was highlighted in the full post it originated from. MTurkers were asked to copy and paste the sentence where a highlighted word is included to a given text box to make sure that MTurkers do not give a random answer. We gave a simple definition of metaphor from Wikipedia along with a few examples to instruct them. Then, they were asked whether the highlighted word is used metaphorically or literally. Five different MTurk workers annotated each candidate word, and they were paid $0.03 for annotating each word. For

candidate	#		%	
	N	**L**	**N**	**L**
boat*	54	281	16.12	83.88
candle*	4	18	18.18	81.82
light	503	179	73.75	26.25
ride	234	185	55.85	44.15
road	924	129	87.75	12.25
spice*	3	21	12.50	87.50
train	94	41	69.63	30.37
all	1816	854	68.01	31.99

Table 1: Metaphor use statistics of data used for MTurk (* indicates metaphor candidates for which the literal usage is more common than the non-literal one, **N**: nonliteral use **L**: literal use).

annotation quality control, we requested that all workers have a United States location and have 98% or more successful submissions. We excluded annotations for which the first task of copy and paste failed. 18 out of 13,348 annotations were filtered out in this way.

To evaluate the reliability of the annotations by MTurkers, we calculated Fleiss's kappa (Fleiss, 1971), which is widely used to evaluate inter-annotators reliability. Using a value of 1 if the MTurker coded a word as a metaphorical use, and a value of 0 otherwise, we find kappa value of 0.81, suggesting strong inter-annotator agreement.

We split the data randomly into two subsets, one as a development set for observation and analysis, and the other as a cross-validation set for classification. The development set contains 800 instances, and the cross-validation set contains 1,870 instances. Table 1 shows the metaphor use statistics of the annotated data.

5 Evaluation

We evaluate our method on a metaphor disambiguation task detailed in Section 5.1. Section 5.2 lists the metrics we used for the evaluation on this test set. Section 5.3 describes the baselines we compare our method against on these metrics. We detail our classification settings in Section 5.4 and report our results in Section 5.5.

5.1 Task

The task for our experiment is metaphor disambiguation: given a candidate word, decide whether the word is used as a metaphor or as a literal word in a post. For example, *boat* in (1) is used metaphorically, whereas *boat* in (2) is used literally. The task is thus to classify each of the seven candidate metaphors defined in Section 4 into either a metaphor or a literal word.

(1) *Just diagnosed late November. Stage I and with good prognosis. ... Now I am having to consider a hysterectomy and am really scared and don't know what to do. I have no children and don't really think I want to. I really want to do what is best for me but it is so hard to know. Anyone else been in the same boat with the endometriosis?*

(2) *Good Morn Girls, It is 52 this morn. WOW! there is a bad storm rolling in at this time and tornado watches but those are pretty common. ... Hubby started his truck driving school today. We use to have ski boats so he and I could both drive a semi. Backing is the hardest part cause the trailer goes opposite of the direction you turn but once you get use to it, it's not hard. ...*

5.2 Evaluation Metrics

We report four evaluation metrics: accuracy, precision, recall, and F-score.

Accuracy: Accuracy is the percentage of correctly classified instances among all instances.

Precision: Precision is the percentage of correctly classified instances among instances assigned to a particular class (metaphor or literal) by the model.

Recall: Recall is the percentage of correctly classified instances among all nonliteral or literal instances. Precision and recall are recorded for both metaphorical and literal labels.

F-score: F-score is the harmonic mean of precision and recall.

5.3 Baselines

We compare our method to a context unigram model as well as two other baselines from recent work on metaphor detection: Klebanov et al. (2014), and Tsvetkov et al. (2013).

Context unigram model uses all the context words including the target word in a post as features.

Type	Model	A	P-M	R-M	P-L	R-L	F1
Baseline	Tsvetkov et al. (2013)	0.245	0.857	0.168	0.236	0.991	0.207
	Klebanov et al. (2014)	0.833	0.830	0.984	0.866	0.340	0.694
	U	0.836	0.867	0.929	0.697	0.535	0.751
Global	U+GWC	0.842	0.869	0.934	0.716	0.541	0.759
	U+GT*	0.843	0.873	0.931	0.711	0.557	0.763
	U+LC	0.839	0.866	0.934	0.709	0.530	0.753
	U+GWC+GT+LC*	0.845	0.871	0.936	0.724	0.546	0.762
Local	U+LWC	0.849	0.874	0.939	0.735	0.557	0.634
	U+SR(AvgSR)	0.852	0.873	0.965	0.563	0.243	0.628
	U+SR(DepSR)	0.858	0.880	0.943	0.756	0.580	0.783
	U+AC	0.853	0.880	0.936	0.735	0.582	0.778
	U+LWC+SR+AC*	0.862	0.885	0.942	0.759	0.598	0.791
Global+Local	ALL*	0.860	0.882	0.943	0.761	0.589	0.788
	ALL-LC*	0.863	0.886	0.941	0.759	0.605	0.793

Table 2: Performance on metaphor disambiguation evaluation. (**Models**) U: context unigram, GWC: global word category, GT: global topic dist., LC: lexical chain, LWC: local word category, SR: semantic relatedness, AC: abstractness/concreteness. (**Metrics**) A: accuracy, P-M: precision on metaphors, R-M: recall on metaphors, P-L: precision on literal words, R-L: recall on literal words, F1: Average F1 score over M/L., *: statistically significant improvement over baselines

Tsvetkov et al. (2013) use local contextual features (such as abstractness and imageability, supersenses, and vector space word representations), and targets for two syntactic constructions: subject-verb-object (`SVO`) and adjective-noun (`AN`) tuples. Note that the output of this system is a sentence level label rather than a word (*e.g.* they output a binary label that indicates whether the target sentence contains any metaphorical phrase). Thus, we take the output of their sentence level label on the sentence that contains our target word, and treat their output as a label for our target word disambiguation task. Although it is therefore not a fair comparison, we included this system as a baseline because this is a state-of-the-art system for metaphor detection tasks. In addition, we can make this comparison to contextualize results with regards to how a state-of-the-art non-discourse model (i.e. not using global context) will perform in more general discourse contexts.

Klebanov et al. (2014) use target word lexical features such as part-of-speech tags, concreteness rating, and topic score. Their approach does not use any contextual information as our method does. As a result, the same words are most likely to obtain the same features. Note that Klebanov et al. (2014) evaluated their approach for each content word in a given text, but in our paper we evaluate how their method performs on ambiguous words in particular.

5.4 Classification

We used the `LightSIDE` toolkit (Mayfield and Rosé, 2010) for extracting features and performing classification. For the machine learning algorithm, we used the logistic regression classifier provided by `LightSIDE` with L_1 regularization. We used basic unigram features extracted by `LightSIDE`, and performed 10-fold cross validation for the following experiments. Instances for each fold were randomly chosen.

5.5 Results

The classification results on the *Breast Cancer* corpus are shown in Tables 2 and in 3.

Note that both our global context features (*e.g.*U+GWC+GT+LC, U+GT) and local context features (*e.g.*U+LWC+SR+AC) perform significantly better than all of the baselines ($p < 0.05$). This indicates that our contextual features successfully capture additional information from discourse both locally and globally. In general, it can be seen that local features are more powerful indicators of metaphors than global features. Note also that Tsvetkov et al. (2013) performs poorly on this task, probably due to the reasons mentioned in Section 5.3. It is interesting to note that

Target word	A	P-M	R-M	P-L	R-L	F1
boat	0.843	0.886	0.935	0.500	0.351	0.843
light	0.831	0.857	0.920	0.738	0.594	0.773
ride	0.843	0.847	0.888	0.836	0.782	0.838
road	0.926	0.936	0.983	0.823	0.543	0.806
train	0.711	0.759	0.887	0.429	0.231	0.559

Table 3: Performance on metaphor disambiguation task per target word with the best setting ALL-LC. Note that the performance results on target words *candle* and *spice* are not reported because of their small number of instances.

Klebanov et al. (2014) performs poorly at recall on literal words. We conclude that our methods significantly outperform the baselines in detecting metaphors in discourse.

6 Discussion

The results of our methods on the metaphor disambiguation task are promising, indicating that both global features and local features can serve as strong indicators of metaphor.

Note that the combined global+local features did not show significant improvement over the local features on this task in Table 2. We had believed that local and global features (aside from unigram features) would provide synergistic predictions, however we found that the local features provided stronger predictions and drowned out the effect of the global features.

We identify the following possible sources of errors of our method: first of all, the low performance of lexican chain (LC) features is noticeable. This might be due to errors originating from the output of the `ELKB` toolkit which we employ to obtain lexical chains. More specifically, `ELKB` builds lexical chains using a standard thesaurus, which is extremely vulnerable to noisy text such as our online discussion forum (which contains typos, abbreviations, medical terms, etc.).

Secondly, the semantic relatedness scores obtained from LDA gives high scores to frequently co-occurring words, thus inevitably reducing effectiveness in disambiguating frequently used metaphors. While this is an issue inherent in any distributional semantics approach, we find that our LDA-based features do improve overall performance.

7 Conclusion

We summarize our contributions as follows: we identified that both global and local contextual features can serve as powerful indicators of metaphor, and proposed several methods to represent contextual features in discourse. We also extended previous literature that considers local contextual information by explicitly incorporating the syntactic information, such as dependency relations, into local contextual features, resulting in an improved performance. The performance was evaluated on our newly built *Breast Cancer* dataset, which provides examples of metaphors in a discourse setting. We showed that our method significantly outperforms the systems from recent literature on a metaphor disambiguation task in discourse. Our method can be easily applied to disambiguate all the content words in text once we have correspondingly labeled data.

Acknowledgments

This research was supported in part by NSF Grant IIS-1302522.

References

Dan Assaf, Yair Neuman, Yohai Cohen, Shlomo Argamon, Newton Howard, Mark Last, Ophir Frieder, and Moshe Koppel. 2013. Why dark thoughts aren't really dark: A novel algorithm for metaphor identification. In *Computational Intelligence, Cognitive Algorithms, Mind, and Brain (CCMB), 2013 IEEE Symposium on*, pages 60–65. IEEE.

Collin F Baker, Charles J Fillmore, and John B Lowe. 1998. The berkeley framenet project. In *Proceedings of the 17th international conference on Computational linguistics-Volume 1*, pages 86–90. Association for Computational Linguistics.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March.

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.

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior research methods*, 46(3):904–911.

Dipanjan Das, Desai Chen, André FT Martins, Nathan Schneider, and Noah A Smith. 2014. Frame-semantic parsing. *Computational Linguistics*, 40(1):9–56.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378–382.

Ting-Hao Kenneth Huang. 2014. Social metaphor detection via topical analysis. In *Sixth International Joint Conference on Natural Language Processing*, page 14.

Hyeju Jang, Mario Piergallini, Miaomiao Wen, and Carolyn Penstein Rosé. 2014. Conversational metaphors in use: Exploring the contrast between technical and everyday notions of metaphor. *ACL 2014*, page 1.

Mario Jarmasz and Stan Szpakowicz. 2003. Not as easy as it seems: Automating the construction of lexical chains using rogets thesaurus. In *Advances in Artificial Intelligence*, pages 544–549. Springer.

Beata Beigman Klebanov, Eyal Beigman, and Daniel Diermeier. 2009. Discourse topics and metaphors. In *Proceedings of the Workshop on Computational Approaches to Linguistic Creativity*, pages 1–8. Association for Computational Linguistics.

Beata Beigman Klebanov, Chee Wee Leong, Michael Heilman, and Michael Flor. 2014. Different texts, same metaphors: Unigrams and beyond. *ACL 2014*, page 11.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60.

James H Martin. 1996. Computational approaches to figurative language. *Metaphor and Symbol*, 11(1):85–100.

Elijah Mayfield and Carolyn Rosé. 2010. An interactive tool for supporting error analysis for text mining. In *Proceedings of the NAACL HLT 2010 Demonstration Session*, pages 25–28. Association for Computational Linguistics.

Jane Morris and Graeme Hirst. 1991. Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Comput. Linguist.*, 17(1):21–48, March.

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.

Xuan-Hieu Phan and Cam-Tu Nguyen. 2007. Gibbslda++: Ac/c++ implementation of latent dirichlet allocation (lda).

Steffen Remus and Chris Biemann. 2013. Three knowledge-free methods for automatic lexical chain extraction. In *HLT-NAACL*, pages 989–999.

Peter Mark Roget. 1911. *Roget's Thesaurus of English Words and Phrases...* TY Crowell Company.

Marc Schulder and Eduard Hovy. 2014. Metaphor detection through term relevance. *ACL 2014*, page 18.

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

Ekaterina Shutova and Simone Teufel. 2010. Metaphor corpus annotated for source-target domain mappings. In *LREC*.

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. Association for Computational Linguistics.

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

Caroline Sporleder and Linlin Li. 2009. Unsupervised recognition of literal and non-literal use of idiomatic expressions. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 754–762. Association for Computational Linguistics.

Gerard J Steen, Aletta G Dorst, J Berenike Herrmann, Anna Kaal, Tina Krennmayr, and Trijntje Pasma. 2010. *A method for linguistic metaphor identification: From MIP to MIPVU*, volume 14. John Benjamins Publishing.

Yulia Tsvetkov, Elena Mukomel, and Anatole Gershman. 2013. Cross-lingual metaphor detection using common semantic features.

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 2011 Conference on the Empirical Methods in Natural Language Processing*, pages 680–690.