Metaphor Identification as Interpretation

Ekaterina Shutova
International Computer Science Institute and
Institute for Cognitive and Brain Sciences
University of California, Berkeley
katia@berkeley.edu

Abstract

Automatic metaphor identification and interpretation in text have been traditionally considered as two separate tasks in natural language processing (NLP) and addressed individually within computational frameworks. However, cognitive evidence suggests that humans are likely to perform these two tasks simultaneously, as part of a holistic metaphor comprehension process. We present a novel method that performs metaphor identification through its interpretation, being the first one in NLP to combine the two tasks in one step. It outperforms the previous approaches to metaphor identification both in terms of accuracy and coverage, as well as providing an interpretation for each identified expression.

1 Introduction

Metaphor undoubtedly gives our expression more vividness, distinction and artistry, however, it is also an important linguistic tool that has long become part of our every-day language. Metaphors arise when one concept or domain is viewed in terms of the properties of another (Lakoff and Johnson, 1980). Consider the examples in (1) and (2).

(1) My car *drinks* gasoline. (Wilks, 1978)

(2) This policy is *strangling* business.

The *car* in (1) and *business* in (2) are viewed as *living beings* and thus they can *drink* or *be strangled* respectively. The mapping between the *car* (the target concept) and *living being* (the source concept) is systematic and results in a number of metaphorical expressions (e.g. "This oil gives your car a *second life*", "this car has is very *temperamental*" etc.) Lakoff and Johnson call such generalisations a source–target domain mapping, or *conceptual metaphor*.

The ubiquity of metaphor in language has been established in a number of corpus studies (Cameron, 2003; Martin, 2006; Steen et al., 2010; Shutova and Teufel, 2010) and the role it plays in human reasoning has been confirmed in psychological experiments (Thibodeau and Boroditsky, 2011). This makes its automatic processing an important problem for NLP and its numerous applications (such as machine translation, information extraction, opinion mining and many others). For example, the use of the metaphorical verb *strangle* in (2) reflects the speaker's negative opinion regarding the government's tight business regulations, which would be an important fact for an opinion mining system to discover (Narayanan, 1999). Other experiments (Agerri, 2008) have investigated and confirmed the role of metaphor interpretation for textual entailment resolution (RTE).

The problem of metaphor modeling is rapidly gaining interest within NLP, with a growing number of approaches exploiting statistical techniques (Mason, 2004; Gedigian et al., 2006; Shutova, 2010; Shutova et al., 2010; Turney et al., 2011; Shutova et al., 2012a). Compared to more traditional approaches based on hand-coded knowledge (Fass, 1991; Martin, 1990; Narayanan, 1997; Narayanan, 1999; Feldman and Narayanan, 2004; Barnden and Lee, 2002; Agerri et al., 2007), these more recent methods tend to have a wider coverage, as well as be more efficient, accurate and robust. However, even the statistical metaphor processing approaches so far often focused on a limited domain or a subset of

*Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task*, pages 276–285, Atlanta, Georgia, June 13-14, 2013.

phenomena (Gedigian et al., 2006; Krishnakumaran and Zhu, 2007), and required training data (Shutova et al., 2010; Turney et al., 2011), often resulting in a limited coverage. The metaphor processing task itself has been most commonly addressed in NLP as two individual subtasks: metaphor identification and metaphor interpretation, with the systems focusing only on one of them at a time, or at best combing the two in a pipeline (Shutova et al., 2012a). Metaphor identification systems annotate metaphorical language in text, and metaphor interpretation systems discover literal meanings of the previously annotated expressions. However, cognitive evidence suggests that humans are likely to perform identification and interpretation simultaneously, as part of a holistic metaphor comprehension process (Coulson, 2008; Utsumi, 2011; Gibbs and Colston, 2012). In this paper, we also take this stance and present the first computational method that identifies metaphorical expressions in unrestricted text by means of their interpretation. Following Shutova (2010), we define metaphor interpretation as a task of finding a literal paraphrase for a metaphorically used word and introduce the concept of *symmetric reverse paraphrasing* as a criterion for metaphor identification. The main assumption behind our method is that the literal paraphrases of literally-used words should yield the original phrase when paraphrased in reverse. For example, when the expression "clean the house" is paraphrased as "tidy the house", the reverse paraphrasing of *tidy* would generate *clean*. Our expectation is that such a symmetry in paraphrasing is indicative of literal use. The metaphorically-used words are unlikely to exhibit this symmetry property when paraphrased in reverse. For example, the literal paraphrasing of the verb *stir* in "*stir* excitement" would yield "provoke excitement", but the reverse paraphrasing of *provoke* would not retrieve *stir*, indicating the non-literal use of *stir*.

We experimentally verify this hypothesis in a setting involving single-word metaphors expressed by a verb in verb-subject and verb-direct object relations. We apply the selectional preference-based metaphor paraphrasing method of Shutova (2010) to retrieve literal paraphrases of all input verbs and extend the method to perform metaphor identification. In summary, our system (1) determines the likelihood of a verb being metaphorical based on its selectional preference strength (Resnik, 1993); (2) identifies a set of literal paraphrases for verbs that may be used metaphorically using the algorithm of Shutova (2010); (3) performs reverse paraphrasing of each of the identified paraphrases, aiming to retrieve the original expression; and (4) if the original expression is retrieved then the verb is tagged as literal, otherwise it is tagged as metaphorical.

We evaluated the performance of the system using the manually annotated metaphor corpus of Shutova and Teufel (2010) in precision- and recall-oriented settings. In addition, we compared its performance to that of a baseline using selectional preference violation as an indicator of metaphor, as well as to two previous metaphor identification approaches of Shutova et al. (2010) and Turney et al. (2011).

2 Related Work

One of the first attempts to identify and interpret metaphorical expressions in text is the met* system of Fass (1991), that utilizes hand-coded knowledge and detects non-literalness via selectional preference violation. In case of a violation, the respective phrase is first tested for being metonymic using hand-coded patterns (e.g. CONTAINER-FOR-CONTENT). If this fails, the system searches the knowledge base for a relevant analogy in order to discriminate metaphorical relations from anomalous ones. The system of Krishnakumaran and Zhu (2007) uses WordNet (the hyponymy relation) and word bigram counts to predict verbal, nominal and adjectival metaphors at the sentence level. The authors discriminate between conventional metaphors (included in WordNet) and novel metaphors. Birke and Sarkar (2006) present a sentence clustering approach that employs a set of seed sentences annotated for literalness and computes similarity between the new input sentence and all of the seed sentences. The system then tags the sentence as literal or metaphorical according to the annotation in the most similar seeds, attaining an f-score of 53.8%.

The first system to discover source–target domain mappings automatically is CorMet (Mason, 2004). It does this by searching for systematic variations in domain-specific verb selectional preferences. For example, *pour* is a characteristic verb in both LAB and FINANCE domains. In the LAB domain it has

a strong preference for *liquids* and in the FINANCE domain for *money*. From this the system infers the domain mapping FINANCE – LAB and the concept mapping *money* – *liquid*. Gedigian et al. (2006) trained a maximum entropy classifier to discriminate between literal and metaphorical use. They annotated the sentences from PropBank (Kingsbury and Palmer, 2002) containing the verbs of MOTION and CURE for metaphoricity. They used PropBank annotation (arguments and their semantic types) as features for classification and report an accuracy of 95.12% (however, against a majority baseline of 92.90%). The metaphor identification system of Shutova et al. (2010) starts from a small seed set of metaphorical expressions, learns the analogies involved in their production and extends the set of analogies by means of verb and noun clustering. As a result, the system can recognize new metaphorical expressions in unrestricted text (e.g. from the seed "*stir* excitement" it infers that "*swallow* anger" is also a metaphor), achieving a precision of 79%.

Turney et al. (2011) classify verbs and adjectives as literal or metaphorical based on their level of concreteness or abstractness in relation to a noun they appear with. They learn concreteness rankings for words automatically (starting from a set of examples) and then search for expressions where a concrete adjective or verb is used with an abstract noun (e.g. "*dark* humour" is tagged as a metaphor and "dark hair" is not). They report an accuracy of 73%.

3 Method

3.1 Selectional Preference Strength Filtering

One of the early influential ideas in the field of computational metaphor processing is that metaphor represents a violation of selectional preferences (SP) of a word in a given context (Wilks, 1975; Wilks, 1978). However, applied directly as an identification criterion, violation of SPs is also indicative of many other linguistic phenomena (e.g. metonymy), and not only metaphor, which is problematic. We modify this view and apply it to measure the potential of a word to be used metaphorically based on its selectional preference strength (SPS). The main intuition behind SPS filtering is that not all verbs have an equal potential of being a metaphor. For example, verbs such as *choose*, *remember*, *describe* or *like* do not have a strong preference for their direct objects and are equally likely to appear with many argument classes. If metaphor represents a violation of SPs, then the verbs with weak SPS are unlikely to be used metaphorically in any context. For every verb in the input text, the filter determines their likelihood of being a metaphor based on their SPS and discards the weak ones. The SPS filter is context-free, and the reverse paraphrasing method is then applied in the next steps to determine if the remaining verbs are indeed used metaphorically in the given context.

We automatically acquired selectional preference distributions for verb-subject and verb-direct object relations from the British National Corpus (BNC) (Burnard, 2007) that was parsed using the RASP parser (Briscoe et al., 2006; Andersen et al., 2008). We applied the noun clustering method of Sun and Korhonen (2009) to 2000 most frequent nouns in the BNC to obtain 200 common selectional preference classes. To quantify selectional preferences, we adopted the SPS measure of Resnik (1993). Resnik defines SPS of a verb as the difference between the posterior distribution of noun classes in a particular relation with the verb and their prior distribution in that syntactic position (regardless of the verb). He quantifies this difference using the Kullback-Leibler divergence:

$$S_R(v) = D(P(c|v)||P(c)) = \sum_c P(c|v) \log \frac{P(c|v)}{P(c)}, \quad (1)$$

where $P(c)$ is the prior probability of the noun class, $P(c|v)$ is the posterior probability of the noun class given the verb and R is the grammatical relation.

We calculated SPS for verb-subject and verb-direct object grammatical relations. The optimal selectional preference strength thresholds were set experimentally on a small heldout dataset at 0.30 for verb-subject and 0.70 for verb-direct object relations (via qualitative analysis of the data). The system excludes expressions containing the verbs with preference strength below these thresholds from the set of candidate metaphors. Examples of verbs with weak direct object SPs include e.g. *imagine, avoid, contain, dislike, make, admire, separate, remember* and the strong SPs are exhibited by e.g. *sip, hobble, roar, hoover, slam, skim, drink* etc.

3.2 Literal Paraphrasing

The verbs that can be used metaphorically according to the SPS filter are then paraphrased using the context-based literal paraphrasing method of Shutova (2010). While Shutova only used the method to paraphrase manually annotated metaphors, we extend and apply the method to paraphrasing of literally used terms and metaphor identification, eliminating the need for manual annotation of metaphorical expressions.

The system takes verbs and their context in the form of subject and direct-object relations as input. It generates a list of possible paraphrases of the verb that can occur in the same context and ranks them according to their likelihood, as derived from the corpus. It then identifies shared features of the paraphrases and the verb using the WordNet (Fellbaum, 1998) hierarchy and removes unrelated concepts. It then identifies literal paraphrases among the remaining candidates based on the verb's automatically induced selectional preferences and the properties of the context.

3.2.1 Context-based Paraphrase Ranking

Following Shutova (2010), we compute the likelihood L of a particular paraphrase of the verb v as a joint probability of the paraphrase i co-occurring with the other lexical items from its context $w_1, ..., w_N$ in syntactic relations $r_1, ..., r_N$.

$$L_i = P(i, (w_1, r_1), (w_2, r_2), ..., (w_N, r_N)). \quad (2)$$

Assuming statistical independence between the relations of the terms in a phrase, we obtain:

$$\begin{aligned} &P(i, (w_1, r_1), (w_2, r_2), ..., (w_N, r_N)) = \\ &P(i) \cdot P((w_1, r_1)|i) \cdot ... \cdot P((w_N, r_N)|i). \end{aligned} \quad (3)$$

The probabilities can be calculated using maximum likelihood estimation as $P(i) = \frac{f(i)}{\sum_k f(i_k)}$ and $P(w_n, r_n|i) = \frac{f(w_n, r_n, i)}{f(i)}$, where $f(i)$ is the frequency of the interpretation irrespective of its arguments, $\sum_k f(i_k)$ is the number of times its part of speech class is attested in the corpus and $f(w_n, r_n, i)$ is the number of times the interpretation co-occurs with context word w_n in relation r_n. By performing appropriate substitutions into (3), we obtain:

$$\begin{aligned} &P(i, (w_1, r_1), (w_2, r_2), ..., (w_N, r_N)) = \\ &\frac{f(i)}{\sum_k f(i_k)} \cdot \frac{f(w_1, r_1, i)}{f(i)} \cdot ... \cdot \frac{f(w_N, r_N, i)}{f(i)} = \\ &\frac{\prod_{n=1}^{N} f(w_n, r_n, i)}{(f(i))^{N-1} \cdot \sum_k f(i_k)}. \end{aligned} \quad (4)$$

This model is then used to rank the candidate substitutes of the verb v in the fixed context according to the data. The parameters of the model were estimated from the RASP-parsed BNC using the grammatical relations output created by Andersen et al. (2008). The goal of this model is to emphasize the paraphrases that match the context of the verb in the sentence best.

3.2.2 WordNet Filter

After obtaining the initial list of possible substitutes for the verb v, the system filters out the terms whose meanings do not share any common properties with that of the verb. This overlap of properties is identified using the hyponymy relation in WordNet. Within the initial list of paraphrases, the system selects the terms that are hypernyms of the verb v, or share a common hypernym with it. Following Shutova, we restrict the hypernym search to a depth of three levels in the taxonomy. Table 1 shows the filtered lists of paraphrases for the expressions "*stir* excitement" and "campaign *surged*". The goal of the filter is to discard unrelated paraphrases and thus ensure the meaning retention during paraphrasing. Note, however, that we define meaning retention broadly, as sharing a set of similar basic properties. Such a broad definition distinguishes our system from other WordNet-based approaches to lexical substitution (McCarthy and Navigli, 2007) and allows for a transition from metaphorical to literal language, while preserving the original meaning.

3.2.3 SP-based Re-ranking

The lists of paraphrases which were generated as described above contain some irrelevant paraphrases (e.g. "campaign lifted" for "campaign *surged*") and some metaphorically-used paraphrases (e.g. "campaign *soared*"). However, our aim is to identify literal paraphrases among the candidates. Shutova's method uses selectional preferences of the candi-

Log-likelihood	Paraphrase
Verb-DirectObject	
stir excitement:	
-14.28	create
-14.84	provoke
-15.53	make
-15.53	elicit
-15.53	arouse
-16.23	stimulate
-16.23	raise
-16.23	excite
-16.23	conjure
Subject-Verb	
campaign ***surge***:	
-13.01	run
-15.53	improve
-16.23	soar
-16.23	lift

Table 1: The list of paraphrases with the initial ranking

Association	Paraphrase
Verb-DirectObject	
stir excitement:	
0.0696	provoke
0.0245	elicit
0.0194	arouse
0.0061	conjure
0.0028	create
0.0001	stimulate
≈ 0	raise
≈ 0	make
≈ 0	excite
Subject-Verb	
campaign ***surge***:	
0.0086	improve
0.0009	run
≈ 0	soar
≈ 0	lift

Table 2: The list of paraphrases re-ranked using SPs

dates for this purpose. Candidates used metaphorically are likely to demonstrate semantic preference for the source domain, e.g. *soar* would select for *birds* or *flying devices* as its subject rather than *campaigns* (the target domain), whereas the ones used literally would have a higher preference for the target domain. This is yet another modification of Wilks' SP violation view of metaphor. Shutova (2010) has previously shown that selecting the paraphrases whose preferences the noun in the context matches best allows to filter out non-literalness, as well as unrelated terms.

As in case of the SPS filter, we automatically acquired selectional preference distributions of the verbs in the paraphrase lists (for verb-subject and verb-direct object relations) from the RASP-parsed BNC. In order to quantify how well a particular argument class fits the verb, we adopted the selectional association measure proposed by Resnik (1993). Selectional association is defined as follows:

$$A_R(v,c) = \frac{1}{S_R(v)} P(c|v) \log \frac{P(c|v)}{P(c)}, \quad (5)$$

where $P(c)$ is the prior probability of the noun class, $P(c|v)$ is the posterior probability of the noun class given the verb and S_R is the overall selectional preference strength of the verb in the grammatical relation R.

We use selectional association as a measure of semantic fitness of the paraphrases into the context, which stands for their literalness. The paraphrases are re-ranked based on their selectional association with the noun in the context. The incorrect or metaphorical paraphrases are de-emphasized within this ranking. The new ranking is shown in Table 2. While the model in 3.2.1 selected the candidate paraphrases that match the context better than all other candidates, the SP model emphasizes the paraphrases that match this particular context better than any other context they may appear in. Shutova's experiments have shown that the paraphrase in rank 1 (i.e. the verb with which the noun in the context has the highest selectional association) represents a literal interpretation in 81% of all cases. Such a level of accuracy makes Shutova's method state-of-the-art in metaphor paraphrasing. We now apply it to the task of metaphor identification.

3.3 Reverse Paraphrasing

At the heart of our approach to metaphor identification is the concept of reverse paraphrasing. The main intuition behind it is that when literally-used words are paraphrased with their literal substitutes, the reverse literal paraphrasing of that substitute should yield the original expression as one of the candidates. This is, however, not the case for metaphor, since its literal paraphrase would yield another literal expression via literal paraphrasing. We ran the above paraphrasing method on every verb in the input text and then again on the top

Original expression	Lit. paraphrase	Reverse paraphrase
Verb-DirectObject		
stir excitement	provoke:	elicit, arouse, cause, create, stimulate, raise, make
	elicit:	provoke, arouse, see, derive, create, raise, make
buy a dress		
	get:	change, find, **buy**, purchase, take, hit, alter, ...
	purchase:	get, **buy**
Subject-Verb		
campaign ***surge***	improve:	change, turn
	run:	succeed, direct, continue, lead, last, win, extend, ...
prisoner **escape**	flee:	**escape**, run
	get:	drive, go, turn, transfer, arrive, bring, come, ...

Table 3: The list of top two literal paraphrases and their reverse paraphrases, as identified by the system

two paraphrases it produces. If this process resulted in retrieving the original expression then the latter was tagged as literal, otherwise it was tagged as metaphorical. Some examples of reverse paraphrasing results are given in Table 3. One can see from the table that when the metaphorical verb *stir* in "*stir* excitement" is paraphrased as the literal "provoke", the subsequent paraphrasing of "provoke" does not produce "stir". In contrast, when the literal expression "buy a dress" is paraphrased as "purchase", the reverse paraphrasing generates "buy" as one of the candidates, indicating the literalness of the original expression. The same is true for the metaphorical *surge* in "campaign *surged*" and the literal *escape* in "the prisoner escaped".

4 Evaluation and Discussion

4.1 Baseline

The baseline system is the implementation of the selectional preference violation view of Wilks (1978) using automatically induced SPs. Such a choice of a baseline allows us to compare our own modifications of the SP violation view to the original approach of Wilks in a computational setting, as well as evaluate the latter on real-world data. Another motivation behind this choice is that the symmetry of reverse paraphrasing can be seen as a kind of "normality" test, in a similar way as the satisfied selectional preferences are in Wilk's approach. However, we believe that the SP-based reverse paraphrasing method captures significantly more information than SP violations do and thus compare the performance of the two methods in an experimental setting.

The baseline SP classes were created as described above and the preferences were quantified using selectional association as a measure. The baseline system then classified the instances where selectional association of the verb and the noun in the phrase were below a certain threshold, as metaphorical. We determined the optimal threshold by qualitative analysis of the selectional preference distributions of 50 verbs of different frequency and SPS (through the analysis of literally and metaphorically-used arguments). The threshold was averaged over individual verbs' thresholds and equals 0.07 for direct object relations, and 0.09 for subject relations.

4.2 Evaluation Corpus

We evaluated the system and the baseline against the corpus of Shutova and Teufel (2010), that was manually annotated for metaphorical expressions. The corpus is a 14,000-word subset of the BNC, with the texts selected to retain the original balance of genre in the BNC itself. The corpus contains extracts from fiction, newspaper text, radio broadcast (transcribed speech), essays and journal articles on politics, social science and literature. Shutova and Teufel (2010) identified 241 metaphorical expressions in the corpus, out of which 164 were verbal metaphors.

We parsed the corpus using the RASP parser and extracted subject and direct object relations from its output. Among the direct object relations there were 310 literal phrases and 79 metaphorical ones; and among the subject relations 206 were literal and 67 metaphorical. This constitutes a dataset of 662 relations for the systems to classify.

4.3 Results and Discussion

The system and baseline performance was evaluated against the corpus in terms of precision and recall. Precision, P, measures the proportion of metaphorical expressions that were tagged correctly among

Relation	Bsln P	System P	Bsln R	System R
Verb-DObj	0.20	0.69	0.52	0.63
Verb-Subj	0.13	0.66	0.59	0.70
Average	0.17	0.68	0.55	0.66

Table 4: Baseline and system performance by relation

the ones that were tagged by the system. Recall, R, measures the proportion of metaphorical expressions that were identified out of all metaphorical expressions in the gold standard corpus. The system $P = 0.68$ and $R = 0.66$, whereas the baseline only attains $P = 0.17$ and $R = 0.55$. System performance by relation is shown in Table 4. The human ceiling for this task, according to the annotation experiments of Shutova and Teufel (2010) approximates to $P = 0.80$. Figure 1 shows example sentences with metaphors identified and paraphrased by the system. Table 5 provides a breakdown of the annotated instances into true / false positives and true / false negatives. As one can see from the table, the systems can accurately annotate both metaphorical and literal expressions, providing a balance between precision and recall.

The system outperforms the baseline for both verb-subject and verb-direct object constructions. Its performance is also close to the previous metaphor identification systems of Turney et al. (2011) (accuracy of 0.73) and Shutova et al. (2010) (precision of 0.79), however, the results are not directly comparable due to different experimental settings. Our method has a strong advantage over the system of Shutova et al. (2010) in terms of coverage: the latter system heavily relied on manually annotated seed metaphors which limited its applicability in unrestricted text to the set of topics covered by the seeds. As opposed to this, our method is domain-independent and can be applied to any data. Shutova et al. (2010) have not measured the recall of their system, however indicated its possible coverage limitations.

In addition, our system produces paraphrases for the identified metaphorical expressions. Since the identification is directly dependent on the quality of literal paraphrasing, the majority of the interpretations the system provided for the identified metaphors appear to be correct. However, we found a few instances where, despite the correct initial paraphrasing, the system was not able to identify

FYT Gorbachev **inherited** a Soviet state which was, in a celebrated Stalinist formulation, national in form but socialist in content.
Paraphrase: Gorbachev received a Soviet state which was, in a celebrated Stalinist formulation, national in form but socialist in content.

CEK The Clinton campaign **surged** again and he easily won the Democratic nomination.
Paraphrase: The Clinton campaign improved again and he easily won the Democratic nomination.

CEK Their views **reflect** a lack of enthusiasm among the British people at large for John Major 's idea of European unity.
Paraphrase: Their views show a lack of enthusiasm among the British people at large for John Major 's idea of European unity.

J85 [..] the reasons for this superiority are never **spelled out**.
Paraphrase [..] the reasons for this superiority are never specified.

J85 Anyone who has introduced speech act theory to students will know that these technical terms are not at all easy to **grasp**.
Paraphrase: Anyone who has introduced speech act theory to students will know that these technical terms are not at all easy to understand.

G0N The man's voice **cut in** .
Paraphrase: The man's voice interrupted.

Figure 1: Metaphors tagged by the system (in bold) and their paraphrases

the metaphor, usually in case of highly conventionalized metaphorical expressions. Overall, the most frequent system errors fall into the following categories:

Errors due to incorrect parsing: The system failed to discover some of the metaphorical expressions in the corpus since their grammatical relations were missed by the parser. In addition, some of the instances were misclassified, e.g. "pounds paid to [...]" or "change was greatly *accelerated*" were labeled as subject relations. Overall, the parser missed 9 metaphorical expressions.

Errors due to incorrect paraphrasing: The most common type of error that leads to false positives is the incorrect paraphrasing (resulting in a change of meaning). This makes it nearly impossible for the system to retrieve the original term. There were also

	Positives	Negatives	Total
True	99	464	563
False	47	52	99
Total	146	516	

Table 5: System tagging statistics

cases where the system could not generate any paraphrase (usually for literal expressions, e.g. "play an anthem").

Errors due to metaphorical paraphrasing: Some of the system errors are due to metaphorical paraphrasing. For example, the metaphorical expression "*mend* marriage" was paraphrased as "*repair* marriage", which is also used metaphorically. And *repair* in return generated *mend*, when paraphrased in reverse. Errors of this type have been mainly triggered by the WordNet filter, and the fact that some metaphorical senses are included in WordNet.

Errors due to metaphor conventionality: a number of conventional metaphors were missed by the system, since the original verb was retrieved due to its conventionality. Such examples include "*impose* a decision", "*put* the issue forward", "*lead* a life". Such cases suggest that the system is better suited to identify more creative, novel metaphors.

Cases of metonymy: a few cases of general metonymy were annotated by the system as metaphorical, e.g. "shout support", which stands for "shout the words of support", and "humiliate a moment", that is likely to mean "humiliate the event of the moment". However, there were only 4 errors of this type in the data.

Baseline Errors: The output of the baseline exhibited two main types of error. The first stemmed from the conventionality of many metaphorical expressions, which resulted in their literal annotation. Conventionality leads to high selectional association for verbs with their metaphorical arguments, e.g. *embrace* has {*view, ideology, conception* etc.} class as its top ranked direct object argument with the selectional association of 0.18. The second type of error was the system selecting many language anomalies that violate selectional preferences and tagging these as metaphors. This resulted in a high number of false positives.

5 Conclusions and Future Directions

Previous research on metaphor addressed a number of its aspects using both symbolic and statistical techniques. While some of this work met with success with respect to precision in metaphor annotation, the methods often focused on a limited domain and needed manually-labeled training data. Their dependence on manually annotated training data made the systems hard to scale. As a result, many of these systems are not directly applicable to aid real-world NLP due to their limited coverage. In contrast, our method does not require any manually-labeled data, which makes it more robust and applicable to a wide range of genres. It is also the first one to perform accurate metaphor identification and interpretation in one step, as opposed to the previous systems focusing on one part of the task only. It identifies metaphor with a precision of 68% and a recall of 66%, which is a very encouraging result. We believe that this work has important implications for computational modeling of metaphor, and is relevant to a range of other semantic tasks within NLP.

Although we have so far tested our system on verb-subject and verb-object metaphors only, we believe that the described identification and paraphrasing techniques can be similarly applied to a wider range of syntactic constructions. Extending the system to deal with more parts of speech and types of phrases (e.g. nominal and adjectival metaphors) is part of our future work.

Another promising future research avenue is integrating the techniques with unsupervised paraphrasing and lexical substitution methods, using e.g. distributional similarity measures (Pucci et al., 2009; McCarthy et al., 2010) or vector space models of word meaning (Erk and Padó, 2008; Erk and Padó, 2009; De Cao and Basili, 2009; Shutova et al., 2012b). These methods could fully or partly replace the WordNet filter in the detection of similar basic features of the concepts, or add useful information to it. Fully replacing the WordNet filter by an unsupervised method would make the system more robust and more easily portable across domains and genres. This may also eliminate some of the system errors that arise from the inconsistent sense annotation and the inclusion of some metaphorical senses in WordNet.

Acknowledgments

This work was supported by the ICSI MetaNet project (grant number W911NF-12-C-0022). Many thanks to Srini Narayanan, Eve Sweetser and Jerry Feldman for their advice and feedback.

References

Rodrigo Agerri, John Barnden, Mark Lee, and Alan Wallington. 2007. Metaphor, inference and domain-independent mappings. In *Proceedings of RANLP-2007*, pages 17–23, Borovets, Bulgaria.

Rodrigo Agerri. 2008. Metaphor in textual entailment. In *Proceedings of COLING 2008*, pages 3–6, Manchester, UK.

Oistein Andersen, Julien Nioche, Ted Briscoe, and John Carroll. 2008. The BNC parsed with RASP4UIMA. In *Proceedings of LREC 2008*, pages 865–869, Marrakech, Morocco.

John Barnden and Mark Lee. 2002. An artificial intelligence approach to metaphor understanding. *Theoria et Historia Scientiarum*, 6(1):399–412.

Julia Birke and Anoop Sarkar. 2006. A clustering approach for the nearly unsupervised recognition of nonliteral language. In *In Proceedings of EACL-06*, pages 329–336.

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

Lou Burnard. 2007. *Reference Guide for the British National Corpus (XML Edition)*.

Lynne Cameron. 2003. *Metaphor in Educational Discourse*. Continuum, London.

Seana Coulson. 2008. Metaphor comprehension and the brain. In R.W. Gibbs, editor, *Metaphor and Thought*, Cambridge. Cambridge University Press.

Diego De Cao and Roberto Basili. 2009. Combining distributional and paradigmatic information in a lexical substitution task. In *Proceedings of EVALITA workshop, 11th Congress of Italian Association for Artificial Intelligence*.

Katrin Erk and Sebastian Padó. 2008. A structured vector space model for word meaning in context. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 897–906, Waikiki, Hawaii, USA.

Katrin Erk and Sebastian Padó. 2009. Paraphrase assessment in structured vector space: exploring parameters and datasets. In *Proceedings of the Workshop on Geometrical Models of Natural Language Semantics*, pages 57–65. Association for Computational Linguistics.

Dan Fass. 1991. met*: A method for discriminating metonymy and metaphor by computer. *Computational Linguistics*, 17(1):49–90.

Jerome Feldman and Srini Narayanan. 2004. Embodied meaning in a neural theory of language. *Brain and Language*, 89(2):385–392.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database (ISBN: 0-262-06197-X)*. MIT Press, first edition.

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

Raymond W. Gibbs and Herbert L. Colston. 2012. *Interpreting Figurative Meaning*. Cambridge University Press.

Paul Kingsbury and Martha Palmer. 2002. From TreeBank to PropBank. In *Proceedings of LREC-2002*, pages 1989–1993, Gran Canaria, Canary Islands, Spain.

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

George Lakoff and Mark Johnson. 1980. *Metaphors We Live By*. University of Chicago Press, Chicago.

James Martin. 1990. *A Computational Model of Metaphor Interpretation*. Academic Press Professional, Inc., San Diego, CA, USA.

James Martin. 2006. A corpus-based analysis of context effects on metaphor comprehension. In A. Stefanowitsch and S. T. Gries, editors, *Corpus-Based Approaches to Metaphor and Metonymy*, Berlin. Mouton de Gruyter.

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

Diana McCarthy and Roberto Navigli. 2007. Semeval-2007 task 10: English lexical substitution task. In *Proceedings of the 4th workshop on Semantic Evaluations (SemEval-2007)*, pages 48–53.

Diana McCarthy, Bill Keller, and Roberto Navigli. 2010. Getting synonym candidates from raw data in the english lexical substitution task. In *Proceedings of the 14th EURALEX International Congress*, Leeuwarden, The Netherlands.

Srini Narayanan. 1997. Knowledge-based Action Representations for Metaphor and Aspect (KARMA). Technical report, PhD thesis, University of California at Berkeley.

Srini Narayanan. 1999. Moving right along: A computational model of metaphoric reasoning about events. In *Proceedings of AAAI 99)*, pages 121–128, Orlando, Florida.

Dario Pucci, Marco Baroni, Franco Cutugno, and Alessandro Lenci. 2009. Unsupervised lexical substitution with a word space model. In *Proceedings of EVALITA workshop, 11th Congress of Italian Association for Artificial Intelligence*.

Philip Resnik. 1993. *Selection and Information: A Class-based Approach to Lexical Relationships*. Ph.D. thesis, Philadelphia, PA, USA.

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

Ekaterina Shutova, Lin Sun, and Anna Korhonen. 2010. Metaphor identification using verb and noun clustering. In *Proceedings of Coling 2010*, pages 1002–1010, Beijing, China.

Ekaterina Shutova, Simone Teufel, and Anna Korhonen. 2012a. Statistical Metaphor Processing. *Computational Linguistics*, 39(2).

Ekaterina Shutova, Tim Van de Cruys, and Anna Korhonen. 2012b. Unsupervised metaphor paraphrasing using a vector space model. In *Proceedings of COLING 2012*, Mumbai, India.

Ekaterina Shutova. 2010. Automatic metaphor interpretation as a paraphrasing task. In *Proceedings of NAACL 2010*, pages 1029–1037, Los Angeles, USA.

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

Lin Sun and Anna Korhonen. 2009. Improving verb clustering with automatically acquired selectional preferences. In *Proceedings of EMNLP 2009*, pages 638–647, Singapore, August.

Paul H. Thibodeau and Lera Boroditsky. 2011. Metaphors we think with: The role of metaphor in reasoning. *PLoS ONE*, 6(2):e16782, 02.

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

Akira Utsumi. 2011. Computational exploration of metaphor comprehension processes using a semantic space model. *Cognitive Science*, 35(2):251–296.

Yorick Wilks. 1975. A preferential pattern-seeking semantics for natural language inference. *Artificial Intelligence*, 6:53–74.

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