What metaphor identification systems can tell us about metaphor-in-language

Jonathan Dunn
Purdue University
West Lafayette, Indiana USA
jonathan.edwin.dunn@gmail.com

Abstract

This paper evaluates four metaphor identification systems on the 200,000 word VU Amsterdam Metaphor Corpus, comparing results by genre and by sub-class of metaphor. The paper then compares the rate of agreement between the systems for each genre and sub-class. Each of the identification systems is based, explicitly or implicitly, on a theory of metaphor which hypothesizes that certain properties are essential to metaphor-in-language. The goal of this paper is to see what the success or failure of these systems can tell us about the essential properties of metaphor-in-language. The success of the identification systems varies significantly across genres and sub-classes of metaphor. At the same time, the different systems achieve similar success rates on each even though they show low agreement among themselves. This is taken to be evidence that there are several sub-types of metaphor-in-language and that the ideal metaphor identification system will first define these sub-types and then model the linguistic properties which can distinguish these sub-types from one another and from non-metaphors.

1 Introduction

The purpose of this paper is to evaluate four systems for identifying metaphor-in-language on the large and representative VU Amsterdam Metaphor Corpus (Steen, et al., 2010) and then to analyze the correct and incorrect identifications in order to see what they can tell us about the linguistic properties of metaphor-in-language. The four metaphor identification systems include a word-level semantic similarity measurement method (Sporleder and Li, 2009; Li and Sporleder, 2010), a word-level abstractness measurement method (Turney and Littmann, 2003; Turney, et al., 2011), a grammatical-relation-level source-target mapping method (Shutova, 2010; Shutova and Teufel, 2010; Shutova, Sun, and Korhonen, 2010; Shutova, Teufel, and Korhonen, 2013), and an utterance-level domain interaction method (Dunn, 2013b).

2 The VU Amsterdam Metaphor Corpus

The VU Amsterdam Metaphor Corpus (Steen, et al., 2010) consists of approximately 200,000 words taken from the British National Corpus's Baby Corpus and divided into four genres: academic, news, fiction, and conversation. It was manually annotated for metaphoric uses of words by five analysts using a version of the MIP method (Pragglejaz Group, 2007). For the purposes of this study, the corpus was divided into sentences, under the assumption that each sentence represents an utterance. There are 16,202 sentences in the corpus. Sentences which contain at least one metaphoric use of a word are labeled as metaphoric sentences. This is done because a metaphorically used word is not metaphoric except in relation to its linguistic context; thus, a larger linguistic unit like the sentence is necessary for revealing metaphorically used words.

The VU Amsterdam Corpus is annotated with several sub-classes of metaphor-in-language. The sub-classes included in this evaluation are MRW-Met (a metaphoric use of a metaphor related word);

Proceedings of the First Workshop on Metaphor in NLP, pages 1–10,
Atlanta, Georgia, 13 June 2013.

Table 1: Number of sentences with sufficient representation in each system.

System	Non-Metaphor	MRW-Met	MRW-Lit	PP	Double	WIDLII
Total	7,979	5,977	126	754	180	1,186
Similarity	4,300	4,274	104	612	153	855
Abstractness	6,851	5,497	118	723	174	1,090
Source-Target	6,256	5,391	121	719	178	1,070
Domain Interaction	6,770	5,588	122	729	178	1,115

MRW-Lit (a literal use of a metaphor related word); PP (a possible personification resulting in a metaphor related word); Double (a metaphor related word which is involved in a double metaphor; for example, personification and a conceptual metaphor); WIDLII (possible metaphor related words which were considered ambiguous between metaphoric and non-metaphoric use).

Table 1 shows a break-down of the number of sentences in each of these sub-classes in the corpus as a whole and as represented by each of the metaphor identification systems. Because each system uses different linguistic properties to identify metaphor-in-language and uses different methods to represent those properties, the systems differ in how many of the sentences are sufficiently represented. For example, the semantic similarity measurement system looks at pairwise similarity values while the abstractness measurement system looks at values for individual words. Thus, the abstractness system could potentially have twice as many data points as the similarity system. The numbers in Table 1 include only the sentences with a minimum number of data points. The evaluation results below do not take into account sentences for which a system has insufficient representation. However, it is important to note that the systems differ in how many sentences they adequately represent, which means that some (for example, the similarity system) are less able to identify metaphor-in-language because they have a less robust representation of the linguistic utterance.

For the purposes of this study, metaphor identification was conceptualized as a sentence-level task. For example, the systems evaluated here could be used within a larger computational semantic system to separate metaphoric and non-metaphoric sentences for purposes of reasoning. One result of this choice is that some of the original systems need to be slightly reconceptualized; thus, it is better to say that these systems are inspired by the cited systems, rather than strict reimplementations of those systems. The similarity and abstractness systems originally were meant to decide which uses of a given verb are metaphoric and which are not metaphoric. In the present study, however, metaphor is not limited to verbs and the systems do not know which words in the sentence may be metaphoric (e.g., it could be any noun or any verb, etc.). Thus, these systems have been altered to determine whether there are any metaphorically used words anywhere in the sentence. Further, all of the reconceptualized systems compared here involve training or seed metaphors, even those which were originally unsupervised systems.

3 Identifying Metaphor-in-Language Using Semantic Similarity

The semantic similarity system (Sporleder and Li, 2009; Li and Sporleder, 2010) uses pairwise semantic similarity to detect metaphoric uses of words. As conceptualized in this study, the system is designed to detect whether any of the words in the sentence are used metaphorically without knowing in advance which words are candidates for metaphoric use.

While the original system used Normalized Google Distance (Cilibrasi and Vitanyi, 2007) to measure semantic similarity, the evaluation in this study used Iosif's SemSim system (Iosif and Potamianos, 2012). There were two main reasons for not using the NGD measure: (1) SemSim offers more control because the corpus used to determine pairwise similarity is known and can be made similar to the test corpus; (2) SemSim is more transparent in terms of its methodology and its results are more stable over time. For this evaluation we used the Open American National Corpus (henceforth,

OANC (Ide and Suderman, 2004)), which consists of 14 million words taken from spoken and written contemporary American English, to determine the pairwise similarity values. Both the test corpus and OANC were lemmatized and had common function words removed. Pairwise similarities were determined for all words in the test corpus which occurred 10 or more times, for a total of 1,691 words. SemSim's contextual window was set at 2. As with all systems discussed below, Morpha (Guido, Carroll, and Pearce, 2001) was used for lemmatization and OpenNLP (Apache, 2011) was used for named entity recognition.

The variables used in the original system had to be changed slightly because no particular word in the sentence is given a special focus. The following variables were used: (1) the number of similarity measurements for a given sentence; (2) the average similarity; (3) the standard deviation of similarity, in order to see how much divergence there was from the average; (4) the highest pairwise similarity; (5) the lowest pairwise similarity; (6) the difference between the highest and lowest pairwise similarity. One of the weaknesses of this particular implementation of the system is that it only considers words that are adjacent to one another (with function words removed). While the original system also used the average pairwise similarity between the candidate word and all other words, this was not possible here given that there were no words starting as candidates.

4 Identifying Metaphor-in-Language Using Word Abstractness

The word abstractness system uses a measurement of word abstractness to identify highly abstract contexts which are posited to be more likely to contain metaphors. In the reconceptualization of the system evaluated here there is also a focus on disparities in abstractness ratings within a given sentence, so that the mixture of abstract and concrete words can be used to detect possible metaphors.

The system first rates lexical items according to how abstract they are, on a scale from 0 to 1, with 1 being the most abstract. The approach to rating abstraction is taken from (Turney, et al., 2011); a list of rated lexical items is available from the authors. The system tags the words in the sentence with their parts of speech and finds the abstractness rating for each; if an abstractness rating is not available for a particular word form, the system attempts to find a match for its lemmatized form. All words not found on the list of abstractness ratings after these searches were removed.

For each sentence a feature vector was created that consisted of twelve different combinations of abstractness ratings: (1) the number of abstractness ratings available for the sentence; (2) the average abstractness for all words; (3) the standard deviation of the abstractness for all words; (3)-(4) the average and standard deviation for the abstractness of nouns; (5)-(6) the average and standard deviation for the abstractness of verbs; (7)-(8) the average and standard deviation for the abstractness of adjectives and adverbs; (9)-(10) the highest and lowest abstractness in the sentence; (11) the difference between the highest and lowest abstractness; (12) the difference between the average abstractness for nouns and for verbs. Empty slots in the feature vector (e.g., if there were no adjectives) were filled with a value of 0.5 for abstractness, following the original system.

5 Identifying Metaphor-in-Language Using Source-Target Mappings

The source-target mapping system clusters verbs and nouns using their distributional properties and argues that abstract nouns will cluster according to the metaphoric source domains to which they are connected. The system moves from the linguistic utterance to the underlying conceptual mapping by assuming that the verb directly represents the source domain in the metaphoric mapping and that nouns (functioning as the subject and/or object of the verb) directly represent the target. Thus, the system looks at grammatical relations containing a verb and a noun and generalizes from seed metaphors to other metaphors involving words from the same clusters.

The first part of evaluating the source-target mapping approach to metaphor identification was to cluster lexical items. The method for clustering verbs is described in (Sun and Korhonen, 2009); (Sun, Korhonen, and Krymolowski, 2008) provide a resource of the most frequent 1,510 English verbs in the Gigaword corpus divided into 170 clusters.

These clusters were used in the evaluation. The procedure used for clustering nouns in (Shutova, Teufel, and Korhonen, 2013) is to include the frequency of grammatical relations (subject, object, indirect object), as annotated by the RASP parser, in a feature vector. In evaluating the source-target system, we took a different approach to obtaining noun clusters. Starting with 8,752 nouns examined by Iosif's SemSim system (Iosif and Potamianos, 2012), we used a pairwise similarity matrix (measured using the Google-based Semantic Relatedness metric, as computed by Iosif) for the feature vector used for clustering nouns. The nouns were divided into 200 clusters using Weka's (Witten and Frank, 2005) implementation of the k-means algorithm.

The search for metaphors was performed on the RASP-parsed version of the evaluation corpus. A total of 1,000 randomly selected metaphoric sentences were used as seed metaphors; any relation between two different clusters was accepted as a candidate. Many of the seed metaphoric utterances contained multiple grammatically related clusters (e.g., verb-object) which were candidates for the metaphoric material in the utterance. In this evaluation we have erred on the side of inclusion by searching for all possible candidates. A total of 903 grammatical relations between clusters were identified in the seed sentences; no attempt was made to trim this number down. While the original system removed verbs which have loose selectional restrictions, such verbs were not removed from the clusters here; the original system focuses on preventing false positives, but in the evaluation here the focus is on preventing false negatives, which such a reduction would necessarily create.

6 Identifying Metaphor-in-Language Using Domain Interactions

The domain interaction system (Dunn, 2013b) is a knowledge-based system unlike the previous distributional-semantic systems. It identifies metaphoric utterances using properties of the concepts pointed to by lexical items in the utterance. The system has two stages: first, determining what concepts are present in an utterance and what their properties are; second, using these properties to model metaphor.

The system maps lexical items to their WordNet synsets (WordNet, 2011) using the part of speech tags to maintain a four-way distinction between nouns, verbs, adjectives, and adverbs. The system then maps the WordNet synsets onto concepts in the SUMO ontology (Niles and Pease, 2001) using the mappings provided (Niles and Pease, 2003). This is done using the assumption that each lexical item is used in its default sense, so that no disambiguation takes place. Once the concepts present in the utterance have been identified in this manner, using the concepts present in the SUMO ontology, the system uses domain (ABSTRACT, PHYSICAL, SOCIAL, MENTAL) and event-status (PROCESS, STATE, OBJECT) properties of each concept present in the utterance. These are not present as such in the SUMO ontology, but were developed following Ontological Semantics (Nirenburg and Raskin, 2004) as a knowledge-base specific to the system.

The domain interaction system was implemented with a feature vector created using the properties of the concepts referred to by lexical items in the utterance. The feature vector uses the following variables: (1) number of concepts in the utterance; (2-5) number of instances of each type of domain (ABSTRACT, PHYSICAL, SOCIAL, MENTAL); (6-8) number of instances of each type of event status (PROCESS, STATE, OBJECT); (9) number of instances of the domain with the highest number of instances; (10) number of instances of event-status with the highest number of instances; (11) sum of the individual domain variables minus (9); (12) sum of individual event-status variables minus (10); (13) number of domain types present at least once in the utterance; (14) number of event-status types present at least once in the utterance; (15) number of instances of the main domain divided by the number of concepts; (16) number of other domain instances divided by the number of concepts; (17) number of main event-status instances divided by the number of concepts; (18) number of other event-status instances divided by the number of concepts.

7 Evaluation Results

The evaluation results discussed in this section consider only the sentences for which each system has the minimum representation; for example, the se-

Table 2: Results for each system across all genres and sub-classes.

System	True Positive	False Positive	True Negative	False Negative	F-Measure
Similarity	5,936	4,214	86	62	0.444
Abstractness	4,627	3,049	3,752	2,954	0.582
Source-Target	1,063	785	5,470	5,496	0.440
Domain Interaction	5,446	3,664	3,106	2,286	0.583

Table 3: Results for each system across all genres and sub-classes without Named Entity Recognition.

System	True Pos.	False Pos.	True Neg.	False Neg.	F-Meas.	Represented
Similarity	5,658	3,973	63	56	0.444	9,750
Abstractness	5,882	4,205	441	354	0.482	10,883
Source-Target	1,725	1,342	2,171	2,677	0.487	8,547
Domain Interaction	6,561	4,205	1,462	676	0.573	12,904

mantic similarity system had a minimum representation for many fewer sentences than does the abstractness system, but those unrepresented sentences are not held against the system. Three of the systems use feature vectors: the semantic similarity, word abstractness, and domain interaction systems. To make the evaluation comparable all three systems are evaluated using Weka's (Witten and Frank, 2005) implementation of the logistic regression algorithm, following (Turney, et al., 2011), using cross-validation (100 folds) and a ridge estimator value of 0.2. The evaluation of the source-target system searched for the 903 seed relations in the RASP-parsed test corpus. The sentences used as seeds were removed from the test corpus before searching. For each evaluation, the reported F-Measure is the weighted average of the F-Measures for metaphors and non-metaphors.

Table 2 shows the evaluation results for the four systems on the entire corpus. The similarity system has the highest number of true positives (5,936), but also the highest number of false positives (4,214). In fact, the similarity system identifies very few utterances as non-metaphors and this makes the results rather unhelpful. The abstractness and domain interaction systems have similar F-measures (0.582 and 0.583, respectively); both make a large number of predictions for both metaphor and non-metaphor, so that they attempt to distinguish between the two, but these predictions are not particularly accurate. The source-target system stands out here, as it does below, with a significantly smaller number of false positives than the other systems (785). At the same time, it also has a significantly higher number of false negatives (5,496). The similarity and source-target systems are on opposite ends of the spectrum in terms of over-identifying and under-identifying metaphor-in-language, and both have similar F-measures (0.444 and 0.440, respectively) which are lower than the abstractness and domain interaction systems.

In Table 3 the same results across all genres and sub-types are presented for implementations without Named Entity Recognition. The only system which performs significantly differently is the abstractness system, with an F-Measure of 0.482 without vs. 0.582 with NER. This decline goes hand-in-hand with the fact that the system with NER has sufficient representation for a total of 14,454 sentences, while without NER it has sufficient representation for only 10,883 sentences.

Table 4 starts to break these results down further by genre, in order to find out if the systems perform differently on different sorts of texts. Every system except for the similarity system (with F-measures of 0.444 and then 0.463) performs more poorly on fiction than on the corpus as a whole. More interestingly, within the fiction genre the similarity and abstractness systems do not predict that any utterances are non-metaphors, which makes their F-measures largely meaningless. The source-target system continues to make a distinction between metaphor and

Table 4: Results for each system in the Fiction genre.

System	True Positive	False Positive	True Negative	False Negative	F-Measure
Similarity	1,778	1,135	0	0	0.463
Abstractness	2,074	1,375	0	0	0.452
Source-Target	293	244	1,151	1,567	0.379
Domain Interaction	2,067	1,349	75	67	0.485

Table 5: Results for each system in the News genre.

System	True Positive	False Positive	True Negative	False Negative	F-Measure
Similarity	1,806	292	0	0	0.796
Abstractness	1,940	321	0	0	0.792
Source-Target	348	61	262	1,352	0.321
Domain Interaction	1,956	324	0	0	0.792

non-metaphor within this genre, although the true and false positives (293 and 243, respectively) are much closer to one another than when looking at the corpus as a whole.

Table 5 looks at the systems' performance within the News genre. The similarity system, which above made few predictions for non-metaphor continues to predict only metaphors; the abstractness and domain interaction systems join it, predicting only metaphors. The source-target system, on the other hand, maintains a small number of false positives (61), although continuing to show a large number of false negatives (1,352). In terms of practical applications, the F-measures here do not adequately reflect the fact that three of the four systems essentially fail on this genre. One of the difficulties is the fact that the News genre contains 1,708 metaphoric sentences and 325 non-metaphoric sentences according to the manual annotations in the VU Amsterdam Metaphor Corpus; that means that 84% of the sentences are annotated as metaphoric.

Table 6 looks at the results within the Academic genre. Here all systems make a distinction between metaphor and non-metaphor; this is the first set on which the similarity system has predicted a meaningful number of non-metaphors. The source-target system misses the most metaphors (1,321) but also makes significantly fewer false positives (146 vs. the next lowest 590 by the similarity system). The F-measures do not adequately reflect the performance of the systems for this genre.

Table 7 shows the results within the Conversation genre. This is the reverse of the News genre: three of the four systems make no predictions of metaphors. This genre contains 1,958 utterances with at least one metaphorically used word and 5,262 without. Further, this genre contains many more short and/or fragmentary sentences than the others. Even the source-target system, which is the only system to identify any metaphors, has more than twice as many false positives as true positives (334 vs. 136, respectively), which reverses its performance on the three previous genres.

The initial conclusions we can draw from the genre break-down is that (1) the F-measure does not always reflect meaningful performance and thus that the numbers of true and false positives and negatives should be reported as well; and (2) that the performance on the corpus as a whole disguises a large amount of variation according to genre.

Table 8 shows the results for only the MRW-Met sub-class in the corpus. This is the basic metaphor sub-class in the corpus and the most common. The systems perform better on this sub-class than on any other. Interestingly, the source-target system makes more false than true positives here (785 vs. 749) and is the only system to make more false than true positives for this sub-class. It also makes more false negatives than the other systems, although the abstractness, source-target, and domain interaction systems make a comparable number (3,971 and 3,990 and 3,386, respectively). The domain interaction system

Table 6: Results for each system in the Academic genre.

System	True Positive	False Positive	True Negative	False Negative	F-Measure
Similarity	1,287	590	289	214	0.635
Abstractness	1,604	667	273	204	0.649
Source-Target	286	146	786	1,321	0.367
Domain Interaction	1,720	720	232	154	0.646

Table 7: Results for each system in the Conversation genre.

System	True Positive	False Positive	True Negative	False Negative	F-Measure
Similarity	0	0	1,994	913	0.558
Abstractness	0	0	4,165	1,759	0.580
Source-Target	136	334	3,271	1,256	0.621
Domain Interaction	0	0	4,070	1,768	0.573

makes the most true positives, although all the F-measures are comparable (the lowest is only 0.062 below the highest).

Table 9 shows the results for the ambiguous metaphors, under the label WIDLII, and the results are comparable to the results for all other sub-classes except for the MRW-Met sub-class (thus, the other sub-classes will not be discussed individually). The similarity, abstractness, and domain interaction systems do not detect any of these sentences as containing metaphorically used words. In some ways this failure is acceptable because the original analysts were not convinced that these utterances contained metaphors in the first place. The source-target system has a very uncharacteristic performance on this sub-class, with 5-times as many false positives as true positives (785 vs. 157, respectively).

This is interesting because it is exactly the opposite of the other systems, which do not predict any sentences to be metaphors at all. This difference is likely a result of the fact that the other three systems rely on feature vectors that were trained on the WIDLII / Non-Metaphor distinction, while the source-target system uses seed grammatical relations from other sub-classes as well (it shouldn't matter because the relations are hypothesized to represent conceptual metaphors for which the sub-class distinction is not relevant; more seed metaphors were not used because this would have removed them from the evaluation). In other words, the sub-class comparisons try to distinguish between WIDLII metaphors and non-metaphors in the corpus. The source-target system was trained on one and only one set of seed metaphors; in other cases this fact increased the system's performance, but in this case it had the opposite effect. It also shows that non-metaphors are more likely to contain the seed clusters than are ambiguous metaphors.

8 Error Analysis

The next question to ask is whether these four systems succeed and fail on the same metaphors. Each system makes different assumptions and is based on a different theory of what linguistic properties are essential to metaphor-in-language, and thus can be used to distinguish metaphor from non-metaphor.

Table 10: Agreement among the four metaphor identification systems using Fleiss' Kappa.

Sub-set	Full	Reduced
Fiction	0.293	0.301
News	0.279	0.277
Academic	0.282	0.286
Conversation	0.259	0.286
MRW-Met	0.280	0.291
MRW-Lit	0.285	0.298
PP	0.293	0.290
Double	0.346	0.369
WIDLII	0.278	0.292

Table 10 shows the agreement between the four

Table 8: Results for each system in the MRW-Met Sub-Class.

System	True Positive	False Positive	True Negative	False Negative	F-Measure
Similarity	2,141	1,841	2,459	2,133	0.536
Abstractness	1,505	1,287	5,514	3,971	0.537
Source-Target	749	785	5,470	3,990	0.499
Domain Interaction	2,202	1,895	4,875	3,386	0.561

Table 9: Results for each system in the WIDLII Sub-Class.

System	True Positive	False Positive	True Negative	False Negative	F-Measure
Similarity	0	0	4,300	855	0.759
Abstractness	0	2	6,799	1,090	0.798
Source-Target	157	785	5,470	768	0.785
Domain Interaction	0	0	6,770	1,115	0.793

systems as measured by Fleiss' Kappa. In the first column, under "Full," the predictions used to determine agreement differ slightly from the earlier predictions because all sentences were included, even those for which a particular system lacked sufficient representation. This was done in order to make a comparison of the four systems possible (sentences without representation could not be identified as metaphors and thus defaulted to non-metaphors). The sentences used as seeds for the source-target system were removed for all systems. A possible cause for low agreement between the systems is that if one system lacks sufficient representation for a sentence, it will cause disagreement by its lack of representation. The second column, under "Reduced," shows the agreement between the four systems for only those sentences for which all systems had an adequate representation and which were not used for seed metaphors (a total of 8,887 sentences rather than the full 16,202). The results are similar, showing that the low agreement is not caused by lack of sufficient representation.

All of the divisions, whether by genre or by subclass, have a similarly low level of agreement, with a range from 0.259 to 0.293. The sub-class of Double metaphors has a higher agreement of 0.346. This low agreement is the case even though the systems have similar overall performance on these particular genres and sub-classes. In other words, even though the systems make similar numbers of correct predictions, the particular utterances for which metaphor is correctly or incorrectly predicted are not the same.

This is an important point because if all four systems succeeded and failed on the same utterances then we could say that those particular utterances were the cause of the failure and try to model the properties of those utterances. What seems to be happening is quite the opposite: each system implements a particular model of metaphor-in-language which makes specific explicit and implicit assumptions about what metaphor-in-language is and what properties are essential for distinguishing metaphoric language from non-metaphoric language. These different models seem to be succeeding on those metaphors which fall within their scope and failing on all others, which leads to disagreement in the predictions of the systems.

9 Synthesizing the Systems

Several meta-systems were constructed using the results of the four systems on the sub-set of the corpus for which each system had adequate representation (8,887 sentences). The first meta-system identified as metaphor only those sentences which the two top-performing systems, the source-target mapping and the domain interaction systems, agreed were metaphoric; the second only those sentences which all four systems agreed were metaphoric; the third only those sentences which a majority of systems agreed were metaphoric; the fourth those sentences for which either the domain interaction or

Table 11: Results for meta-systems across all sentences with sufficient representation for all systems.

System	True Positive	False Positive	True Negative	False Negative	F-Measure
Only top two agree	520	360	3,558	4,449	0.362
Only all agree	374	244	3,674	4,595	0.341
Majority vote	1,513	1,655	2,263	2,921	0.445
Top two inclusive	3,200	2,552	1,366	1,769	0.505
Top two, settled inc	2,689	2,164	1,754	2,280	0.501
Top two, settled exc	2,086	1,688	2,230	2,883	0.485

the source-target system identified as metaphor; the fifth all sentences which the domain interaction and source-target systems agreed were metaphoric, using the similarity and abstractness systems to resolve disagreement. There are two versions of this last meta-system: the inclusive version identifies disputed sentences as metaphoric if either the similarity or abstractness system does, and the exclusive version only if the two agree.

Table 11 shows the results of the evaluations of these meta-systems. The system with the fewest false positives is the one which requires four-way agreement before an utterance is identified as metaphor; however, this also has the fewest true positives. The performance of the exclusive meta-system for the top two systems has a better proportion of true to false positives, but also has an unfortunately high number of false negatives. The majority vote meta-system has more false than true positives and, thus, is not successful. The last three meta-systems differ in how they resolve disagreements between the top two systems; there is a consistent trade-off between more true positives and fewer false positives and all three have comparable F-measures.

10 What This Tells Us About Metaphor-in-Language

What can we learn about metaphor-in-language from the successes and failures of these four metaphor identification systems? First, there is a significant difference between genres. The linguistic properties which can distinguish metaphors in one genre may not apply to other genres. Or, looked at another way, different genres are more likely to contain different types of metaphors (the types of metaphor referred to here involve different sources of metaphoric meaning and are not comparable to the corpus's sub-classes).

Second, the predictions of the four systems, regardless of their accuracy, have a relatively low level of agreement. This low level of agreement is consistent across genres and sub-classes. This means that the systems are succeeding and failing on different metaphors. Each of the systems is based on a different theory of metaphor-in-language. The combination of these two facts suggests that different types of metaphor have different linguistic properties.

Most theories of metaphor conceive of it as a single and coherent phenomenon, so that the predictions of competing theories are mutually exclusive. The lack of agreement coupled with similar success rates, however, suggests that these theories of metaphor-in-language are not mutually exclusive but rather apply to different types of metaphor-in-language. If this is the case, then a more accurate model of metaphor-in-language will start by positing a number of different types of metaphor-in-language, which differ in the source of their metaphoric meaning, and then predicting what linguistic properties can be used to distinguish among these types and between them and non-metaphors.

Metaphor identification systems can be improved by focusing on two important properties of metaphor-in-language: First, metaphors are gradient, with some being much more metaphoric than others (Dunn, 2011). One problem with the systems described in this paper is that they are forced to draw an arbitrary line between two classes to represent a gradient phenomenon. Second, metaphoric expressions receive their metaphoric meaning from different sources (Dunn, 2013a). These different types of metaphor-in-language have different properties and should be modeled individually.

References

Apache. 2011. OpenNLP

Briscoe, E., Carroll, J., Watson, R. "The Second Release of the RASP System." Curran, J. (ed.) *Proceedings of COLING/ACL 2006 Interactive Presentation Sessions* 77-80 Association for Computational Linguistics Stroudsburg, PA 2006

Cilibrasi, R. and Vitanyi, P. "The Google similarity distance." *Knowledge and Data Engineering, IEEE Transactions on* 19(3): 370–383 2007

Dunn, J. "Gradient semantic intuitions of metaphoric expressions" *Metaphor & Symbol* 26(1): 53-67 2011

Dunn, J. "How linguistic structure influences and helps to predict metaphoric meaning" *Cognitive Linguistics* 24(1): 33-66 2013

Dunn, J. "Evaluating the premises and results of four metaphor identification systems." Gelbukh, A. (ed.) *Proceedings of CICLing 2013, LNCS 7816* 471-486 Springer Heidelberg 2013

Guido, M., Carroll, J., Pearce, D. "Applied morphological processing of English." *Natural Language Engineering* 7(3): 207-223 2001

Ide, N. and Suderman, K. "The American National Corpus First Release." Lino, M. Xavier, M., Ferreira, F., Costa, R., and Silva, R. (eds.) *Proceedings of LREC-2004* 1681-1684 European Language Resources Association Paris 2004

Iosif, E. and Potamianos, A. "SemSim: Resources for Normalized Semantic Similarity Computation Using Lexical Networks." Calzolari, N., Choukri, K., Declerck, T., Doan, M., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S. (eds.) *Proceedings of LREC-2012* 3499-3504 European Language Resources Association Paris 2012

Li, L. and Sporleder, C. "Using Gaussian Mixture Models to Detect Figurative Language in Context." Kaplan, R., Burstein, J., Harper, M., and Penn, G. (eds.) *Proceedings of HLT-NAACL-2010* 297–300 Association for Computational Linguistics Stroudsburg, PA 2010

Niles, I. and Pease, A. "Towards a Standard Upper Ontology" Welty, C. and Barry, C. (eds.) *Proceedings of FOIS-2001* 2-9 Association for Computational Linguistics Stroudsburg, PA 2001

Niles, I. and Pease, A. "Linking Lexicons and Ontologies: Mapping WordNet to the Suggested Upper Merged Ontology." Arabnia, H. (ed) *Proceedings of IEEE Intl Conf on Inf. and Knowl. Eng. (IKE 03)* 412-416 IEEE Press New York 2003

Nirenburg, S. and Raskin, V. *Ontological Semantics* Cambridge, MA MIT Press 2004

Pragglejaz Group "MIP: A method for identifying metaphorically used words in discourse." *Metaphor and Symbol* 22(1): 139 2007

Princeton University *WordNet* 2012

Shutova, E. "Models of Metaphor in NLP." Hajiv, J., Carberry, S., Clark, S. and Nivre, J. (eds.) *Proceedings of ACL-2010* 688–697 Association for Computational Linguistics Stroudsburg, PA 2010

Shutova, E. and Teufel, S. "Metaphor corpus annotated for source – target domain mappings." Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Rosner, M. and Tapias, D. (eds.) *Proceedings of LREC 2010* 3255–3261 European Language Resources Association Paris 2010

Shutova, E., Sun, L,. and Korhonen, A. "Metaphor identification using verb and noun clustering." Huang, C. and Jurafsky, D. (eds.) *Proceedings of COLING 2010* 1002–1010 Tsinghua University Press Beijing 2010

Shutova, E., Teufel, S., and Korhonen, A. "Statistical Metaphor Processing." *Computational Linguistics* 39 2013

Sporleder, C. and Li, L. "Contextual idiom detection without labelled data." Koehn, P. and Mihalcea, R. (eds.) *Proceedings of EMNLP-09* 315-323 Association for Computational Linguistics Stroudsburg, PA 2009

Steen, G., Dorst, A., Herrmann, J., Kaal, A., and Krennmayr, T. "Metaphor in usage." *Cognitive Linguistics* 21(4): 765-796 2010

Sun, L. and Korhonen, A. "Improving verb clustering with automatically acquired selectional preferences." Koehn, P. and Mihalcea, R. (eds.) *Proceedings of EMNLP-2009* 638–647 Association for Computational Linguistics Stroudsburg, PA 2009

Sun, L., Korhonen, A., and Krymolowski, Y. "Verb Class Discovery from Rich Syntactic Data." Gelbukh, A. (ed) *Proceedings of CICLING-2008, LNCS, vol. 4919* 16-27 Springer Heidelberg 2008

Turney, P. and Littman, M. "Measuring praise and criticism: Inference of semantic orientation from association." *ACM Transactions on Information Systems* 21(4): 315–346 2003

Turney, P., Neuman, Y, Assaf, D., and Cohen, Y. "Literal and Metaphorical Sense Identification through Concrete and Abstract Context." Barzilay, R. and Johnson, M. (eds.) *Proceedings of EMNLP-2011* 680–690 Association for Computational Linguistics Stroudsburg, PA 2011

Witten, I. and Frank, E. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations* Morgan Kaufmann San Francisco 2005