Evaluating the Premises and Results of Four Metaphor Identification Systems

Jonathan Dunn

Purdue University
West Lafayette, IN USA
jonathan.edwin.dunn@gmail.com

Abstract. This study first examines the implicit and explicit premises of four systems for identifying metaphoric utterances from unannotated input text. All four systems are then evaluated on a common data set in order to see which premises are most successful. The goal is to see if these systems can find metaphors in a corpus that is mostly non-metaphoric without over-identifying literal and humorous utterances as metaphors. Three of the systems are distributional semantic systems, including a source-target mapping method [1–4]; a word abstractness measurement method [5], [6, 7]; and a semantic similarity measurement method [8, 9]. The fourth is a knowledge-based system which uses a domain interaction method based on the SUMO ontology [10, 11], implementing the hypothesis that metaphor is a product of the interactions among all of the concepts represented in an utterance [12, 13].

1 Introduction

This study evaluates four different approaches to metaphor identification. First, each approach's premises and view of metaphor, whether explicit or implicit, are examined in order to understand the differing claims made about metaphor. Second, all four systems are evaluated on a single data set in order to compare their effectiveness. This is important because it helps us to understand which premises are valid (i.e., successful) and which are not (i.e., unsuccessful). Each approach posits certain properties of metaphors that can be used to distinguish metaphors from non-metaphors. The goal of this study is to determine if these properties are essential properties of metaphor or accidental properties that can distinguish metaphors from non-metaphors only in limited data sets.

Humor is used as a counterfactual to metaphor because it contains many of the same properties as metaphor (i.e., connections between different domains) but is interpreted in a very different way. In metaphor, the domains are seen as similar and the interpretation of the utterance involves synthesizing aspects of the two domains. In humor, however, the domains are seen as incongruous and the interpretation of the utterance focuses on dissimilarities between the two domains. For this reason, humor is a useful counterfactual for testing the precision of metaphor identification systems: do the properties posited to be unique to metaphor also show up in humor?

A. Gelbukh (Ed.): CICLing 2013, Part I, LNCS 7816, pp. 471–486, 2013.

This study uses a new data set to provide comparable evaluations of these four systems. The evaluation corpus consists of 25% metaphoric (500), 25% humorous (500), and 50% literal (2,000) utterances taken from the Corpus of Contemporary American English [14]. The evaluation corpus is organized into four top-level domains (ABSTRACT, MENTAL, SOCIAL, PHYSICAL) each of which is represented by instances of five different verbs. This organization ensures wide coverage and allows the results to be examined according to domain membership.

2 Premises of Metaphor Identification Systems

2.1 Possible Choices

The discussion of each metaphor identification system will highlight seven theoretical choices or assumptions which each must make, whether that choice is implicit or explicit. First, is metaphor based on conceptual source-target mappings; if so, are these mappings directly available in the linguistic utterance or are they mediated and thus not directly available? Second, is metaphor a binary or a gradient phenomenon? Third, is lexical meaning best discovered using distributional profiles or using human intuitions; are the two methodologies incompatible? Fourth, are lexical items assumed to point or refer to concepts in the human conceptual system? Fifth, are these concepts organized into domains and, if so, do domains behave differently? Sixth, is metaphor a property possessed by instances of lexical items, by grammatical relations or phrases, or by utterances as a whole? Seventh, do metaphors and non-metaphors belong to distinct populations or do both represent different tails in the distribution of a single population? These questions are discussed in reference to each of the systems as relevant.

2.2 Source-Target Mapping System

This section looks at a verb-noun clustering approach to identifying metaphors [1–4]. The system relies on the view that metaphor consists of a source and a target and that the two metaphorically mapped concepts are directly represented in the surface utterance (e.g., are present in the input sentence, so that no distinction is made between concepts and lexical items). Thus, since metaphor in this view consists of source-target mappings, the metaphor identification task consists in discovering whether these mappings are present or not present (a binary task: an utterance either is or is not metaphoric). 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. This assumption is used to avoid the problem of determining which material in a metaphoric utterance is the “literal” material making up the source and which is the “metaphoric” material making up the target. This is a problem that must be faced under this view of metaphor because all that we see in the linguistic utterance is that some elements do not seem to match or go with other elements in literal language.

With this assumption in place, the system invokes the premise of distributional semantics, that the meaning of a lexical item is determined by (or at least described by) the patterns of its use as measured by the clustering of its surface arrangement in a large body of text. Thus, lexical items used in the same surface contexts have the same or similar meanings. We could perhaps posit a weaker distributional semantics premise, in which lexical items which occur in the same contexts have the same meaning and the same grammatical properties; other lexical items which have the same meaning are prevented from occurring in those contexts for syntactic or morphological reasons. The more similar the contexts, the more similar the meanings. The system combines this premise with the idea that there is a difference between the behavior of abstract and physical lexical items (requiring a sharp distinction between ABSTRACT and PHYSICAL domains). While physical lexical items cluster together (e.g., occur in the same contexts) according to their meaning, abstract lexical items cluster together according to their metaphoric association with particular source domains. In other words, these abstract lexical items derive their distributional properties from their metaphoric connection to particular source domains, so that lexical items which have very different meanings occur in the same contexts as a result of taking on the distributional properties of a single source domain.

This approach to metaphor identification is phrase-based, finding metaphors within grammatically related pairs (e.g., verb-object). This contrasts with word-level approaches (see the similarity and abstraction systems below) and utterance-level approaches (see the domain interaction system below). The focus on grammatical (e.g., syntactic) relations raises the problem of form-meaning mappings: are there metaphoric expressions in which the metaphoric mapping is represented by words that do not have a dependency relationship in the surface structure of the sentence? This lack of a dependency relationship can come about either because the concepts are not explicitly present in the linguistic utterance or because they are grammatically separated. The adoption of this operationalization of the source-target model brings with it the implicit premise that there is a one-to-one mapping between syntactic structure and semantic structure.

Further, because in Conceptual Metaphor Theory [15,16] metaphoric sources have many different targets and metaphoric targets have many different sources, this clustering approach requires that these overlapping connections always correspond: if RELATIONSHIPS and ARGUMENTS can be WARS, then they must also both be able to be JOURNEYS. But what happens when the members of a cluster which share a few source domain mappings diverge strongly in their preference for other source domains? How does this affect the identified clusters, which are then used to generalize mappings from one metaphor to another? In other words, if A and B are clustered together because both map to C, and if A also maps with D but B does not, then a seed metaphor with an A-D mapping will incorrectly predict B-D mappings. This issue may cause false positives but will not cause false negatives.

Source-Target Mapping System: Methods. First, the source-target mapping system parses the linguistic data into grammatical relations using the RASP parser [17]. At the same time, verb and noun clusters are formed by looking at the grammatical contexts in which they occur in a larger corpus [18, 19]. Although statistical methods provide candidate clusters of verbs and nouns, humans intervene in the final selection of clusters. The final clusters are trained on the seed metaphoric utterances in order to learn what source-clusters (e.g., verbs) map metaphorically with what target-clusters (e.g., nouns). Finally, a selectional preference filter is used to eliminate false positives from the identifications; this filter will not reduce but may increase the number of false negatives, the rate of which is not reported in the original study.

2.3 Word Abstractness System

The second metaphor identification system [7] is based on the claim that metaphors occur in abstract contexts, so that metaphor identification requires a measure of abstractness for lexical items and their contexts. Before turning to the identification system itself, previous work by one of the principles needs to be considered [5]. Neuman's related work is in many ways the reverse of the source-target mapping system: rather than use clusters of similarly behaving lexical items to identify metaphoric mappings, Neuman uses clusters of similar metaphoric mappings to determine which lexical items have a similar meaning. In other words, Neuman argues that if we collect a large number of metaphoric expressions and determine which lexical items / concepts are involved in metaphoric mappings, then we can find the meanings of the words if we assume that lexical items / concepts which exist in mappings to the same source domain have the same meaning. Thus, it is the reverse of the source-target mapping system. Neuman's point is that distributional, bag-of-words semantics is too simplistic; a better system is a distributional, bag-of-relations system in which the focus is on semantic relations between concepts. Metaphor, he argues, is one such relation: "our basic thesis is that by analyzing metaphors in which our target term is embedded we may uncover its meaning" (2720).

Neuman's approach here depends upon the premise that metaphoric mappings are (1) mediated and (2) themselves as basic as or more basic than the concepts which form the source and target domains. Shannon [20] points out that cognitive approaches to metaphor assume that the source and target involved in a metaphoric mapping have a fixed and already existing set of properties, only some of which will be activated during the mapping. Shannon's claim is that metaphor itself is more basic than the concepts involved, which means that there is not a source and target directly available in the linguistic utterance. This is what Neuman calls a mediated mapping: while the metaphor is present in the utterance, the source and target are not linked to directly from the linguistic expression. The metaphor itself, existing on its own, provides that link first. The practical implication of a mediated mapping is that approaches to metaphor identification which require finding explicitly and overtly a source concept and target concept represented by lexical items in the linguistic utterance will miss a great many metaphors: those entities are not directly present in this view and so cannot be used for metaphor identification. The problem of mediated mappings is distinct from the problem of metaphors with mappings within a single domain [21] which, however, still raises problems for metaphor identification.

Both Neuman's related work and the domain interaction system below, which is based on Ontological Semantics [22], are explicit about the relationship between seen lexical items and the unseen concepts to which they refer. The idea is that natural language depends upon a human conceptual system of discrete, related concepts which can be modeled computationally using an ontology. Lexical items point or refer to concepts in this view; lexical items can point directly to a concept, they can point to a concept and specify or alter properties of that concept, and they can point to no concept but rather alter properties of the utterance as a whole [23]. Natural language meaning is also often under-specified in this view, so that concepts are present in the semantic structure of the utterance without being explicitly pointed to by lexical items. These relationships between lexical items and concepts affect all of the metaphor identification systems discussed in this paper, although they are only explicitly treated in the domain interaction system. At the same time, none of the systems here can deal with concepts that are not directly represented in the utterance by lexical items.

Word Abstractness System: Methods. In spite of Neuman's nuanced account of mediated metaphoric mappings, this metaphor identification system [7] does not rely on a principled view of metaphor: "Therefore we hypothesize that the degree of abstractness in a word's context is correlated with the likelihood that the word is used metaphorically" (680). The system focuses on the identification of metaphoric senses of a lexical item; certainly for some lexical items an abstract context will signal a metaphoric usage. But how well does this system transfer to new lexical items? And what about metaphors that occur in a non-abstract context: many metaphoric expressions describe physical scenes.

The word abstractness system relies, like the source-target mapping system, on a distinction between ABSTRACT and PHYSICAL domains, although the distinction is assumed to be gradient: lexical items are assigned a value that represents their relative abstractness. Metaphors are assumed to have a unique pattern of abstractness values, most likely patterns in which a few non-abstract lexical items occur in a highly abstract context. Unlike the source-target mapping system, this system can allow the implementation of a fuzzy or gradient threshold for metaphor identification.

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 [6]; 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. For each sentence a feature vector is created that consists of five different combinations of abstractness ratings: (1) average of all non-proper nouns; (2) average of all proper nouns; (3) average of all verbs excluding target verb; (4)

average of all adjectives; (5) average of all adverbs. This vector is trained with a number of tokens of different verbs that are used metaphorically using a logistic regression learning algorithm. This is then applied to new instances of the same verbs as well as to new verbs.

2.4 Semantic Distance / Similarity System

This section looks at a metaphor identification system [8,9] that depends on the hypothesis that metaphoric material comes from a different origin (distribution) than non-metaphoric material. In other words, metaphor and non-metaphor are entirely separate, belonging to different populations with different properties. The system claims that metaphor can be identified by looking at semantic similarity measures within and between the metaphoric and non-metaphoric material in an utterance. Thus, literal and non-literal sentences or word usages are from two different categories and some mixture of properties will be able to determine which population or category a particular sentence or word usage belongs to.

The main property of non-literal language is that it does not exhibit semantic similarity with its context. The non-literal language does not fit, or exhibits a mismatch, with the semantic context in which it occurs. Thus, the task of metaphor identification is a matter of measuring semantic similarity. The distributional properties of lexical items are used as a representation of the meaning of the lexical items, so that lexical items which occur in the same contexts have the same meanings. As a result of this premise, the system claims that metaphors can be detected by finding unusual contexts; this is because semantic similarity or distance is measured using contexts in the first place (lexical items that do not frequently occur together will be measured as semantically dissimilar, so that this system detects infrequent co-occurrences). One result of this premise is that if a particular lexical item occurs often enough in a certain metaphoric mapping, this mapping will become part of the literal meaning of that lexical item: frequent metaphors will not be detected as metaphors. This is an interesting side-effect and may represent how speakers actually process metaphoric utterances. At the same time, the system seems to ignore the fact that unusual patterns (as measured using distributional semantic methods) have many possible sources, of which metaphor is only one. Humor, used as a counterfactual in the evaluation below, is another possible source.

Semantic Distance / Similarity System: Methods. The system adopts the distributional semantic premises and uses Normalized Google Distance [24] as the instrument for measuring semantic similarity or cohesion. Each sentence is represented using a feature vector with five different similarity measures: (1) the semantic similarity between the target expression and its context; (2) the average semantic similarity of the sentence as a whole; (3) the difference between the first and second measures; (4) a binary distinction between cases with a low or high difference between average and expression-specific semantic similarity; (5) the highest degree of similarity between the target expression and its context. A Bayes decision rule is used to determine which population the sentence is more likely to belong to, metaphoric or non-metaphoric, based on similarity of a sentence's feature vector with the feature vector of seed metaphors.

2.5 Domain Interaction System

The knowledge-based system, a domain interaction system called MIMIL (Measuring and Identifying Metaphor-in-Language), identifies metaphoric utterances using properties of all 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 first stage will be discussed below under methods.

The domain interaction system assumes, as discussed above involving Neuman's related work, that lexical items refer or point to concepts in the human conceptual system. These concepts have many properties and relations that connect them. Following Ontological Semantics [22], concepts are represented in part using two ontological properties: domain (for example, PHYSICAL VS. MENTAL), which is a product of the hierarchy of concepts; and event-status (for example, OBJECT VS. PROCESS / EVENT), which is independent of the hierarchy of concepts. The system assumes that every concept possesses these two properties and that these properties are sufficient for distinguishing metaphor from non-metaphor. In this way, the system takes concepts and their properties as identified by human intuitions (present in a knowledge-base) rather than as identified using distributional semantics.

The domain interaction system further assumes that metaphoricity is an utterance-level property that is not possessed by individual lexical items or individual grammatical relations. Thus, the system takes utterances as input (more specifically, the concepts referred to by the lexical items in the linguistic utterance). This means that all the concepts in the utterance, whether or not they are found in certain grammatical configurations, can interact metaphorically. The approach is somewhat similar to the source-target mapping system, except that it is formulated so that the source and target do not have to be identified as such. The advantage to this approach is that it does not assume a one-to-one mapping between syntactic structure and semantic structure and that it does not assume that nouns and verbs always represent the metaphoric target and source, respectively. A further advantage is that the system deals with concepts directly, rather than assuming that lexical items and concepts are equivalent. Finally, the domain interaction system covers both mediated (i.e., indirectly present) and unmediated metaphoric conceptual mappings, while the source-target mapping system covers only unmediated mappings (as discussed above).

On the other hand, these assumptions bring with them several weaknesses. The first is that the system removes all grammatical information from consideration: all concepts are assumed to interact equally with all other concepts. Even though there is not a one-to-one mapping between syntactic structure and semantic structure, there is a good deal of mapping between the two and the system ignores this. The second is that, while the system does not arbitrarily limit its scope to noun-verb relations (and thus includes more concepts in the utterance that are relevant for metaphor detection), it dilutes the influence of the relevant concepts by including irrelevant concepts as well. In other words, the source-target mapping system is limited because it takes a narrow approach to deciding what lexical items in the utterance are relevant for metaphor; but the domain interaction approach is limited because it takes a broad approach that avoids the issue altogether. The empirical question, to be tested below, is which simplifying assumption has fewer side-effects. Ultimately, both systems are implementations of the same underlying view of metaphor, that it involves the interaction between cognitive domains. They differ, however, in what properties of the interaction between domains are considered relevant to metaphor.

The domain interaction system differs from the semantic similarity / distance system because it assumes that metaphors and non-metaphors belong to different extremes of the same population. Thus, the system views metaphoricity as a continuous property which utterances possess to greater or lesser degrees. On one side of the distribution of this population are proto-typical metaphors, which speakers of a language would intuitively identify as metaphoric. On the other side of the distribution are proto-typical non-metaphors, which speakers of a language would intuitively identify as literal. In the middle, however, a majority of utterances have some amount of metaphoricity but are not clearly metaphoric or non-metaphoric (see [25] for further discussion). The implementation of the domain interaction system evaluated here is binary, so that utterances are taken to be metaphoric or non-metaphoric. However, like the word abstractness system and the semantic similarity / distance system, it can be converted into a gradient system that identifies different levels or degrees of metaphor (e.g., moderately metaphoric utterances vs. highly metaphoric utterances). The source-target mapping system is alone in not being easily converted into a gradient system (although it could be converted; for example, by manually assigning different weights to the seed metaphors used).

Domain Interaction System: Methods. The domain interaction system takes as input unrestricted and unannotated English text and uses existing resources to pre-process that text; the pre-processing constitutes the first stage discussed above which identifies the concepts referred to by the lexical items in the utterance. First, the system relies on OpenNLP [26] for tokenization, named entity recognition, and part of speech tagging. Second, the system relies on Morpha [27] for lemmatizing words. At this point, the lemmatized words are mapped to their WordNet synsets [28] 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 [10] using the mappings provided [11]. 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 makes use of 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 [22] as a knowledge-base specific to the domain interaction system.

The system claims that the interaction between the properties of the concepts referred to in the utterance can be used to identify metaphors. The implementation of the system evaluated here creates a feature vector using variables based on the properties of the concepts (discussed below).

3 Evaluation

This study replicates the methods of the systems in question in their most important details, adding new distinctions in order to test the explicit and implicit premises of the approaches. The unifying factor is the data set, which uses humorous utterances as a counterfactual for testing for the over-identification of metaphor.

First, the systems are evaluated using different classes: a three-way distinction between metaphor, humor, and literal language; a two-way distinction between metaphor and non-metaphor (a) with humor included in non-metaphor and (b) with humor excluded. These conditions allow us to test whether humor interferes with metaphor identification methods.

Second, the systems are evaluated using a four-way distinction between domains and without any distinction between domains. This allows us to test whether domain membership influences the behavior of metaphors (as revealed in the success rate of the identification systems).

3.1 Evaluation Methods for Source-Target Mapping System

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 [19]; [6] 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 [4] is to include the frequency of grammatical relations (subject, object, indirect object), as annotated by the RASP parser, in a feature vector used to cluster nouns. 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 [29], 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 [30] implementation of the k means algorithm.

There are advantages and disadvantages to relying on the semantic relatedness metric rather than frequency of grammatical relations. On the one hand, the similarity measure is less sensitive to arbitrary patterns of object restrictions. In other words, many objects and indirect objects cannot occur with certain verbs, not because of their meaning but because of verb valency. This interferes with using grammatical relations as a substitute for meaning. The clusters put together using the similarity measure, however, will not all share the same valency but should have a related meaning. On the other hand, because the system detects similar combinations of a verb cluster and a noun cluster, valency is a salient property even though it is not directly related to meaning. Unlike the original system, no manual intervention was used in preparing the noun clusters. Finally, the evaluation did not need to filter out sentences with loose-valency verbs, those that accept a large variety of arguments, because the test corpus was designed around certain verbs chosen, in part, to avoid this property.

The search for metaphors was performed on the RASP-parsed version of the evaluation corpus; all verb-noun relations were included in the search. For each verb, 5 out of 25 metaphoric instances were used as seed cases, for a total of 105 seed metaphors. The seed metaphors were searched for across all verbs, not restricted to the verb they were taken from. 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. No clear procedure was provided for choosing from among the candidate relations; in this evaluation we have erred on the side of inclusion by searching for all possible candidates. A total of 478 grammatical relations between clusters were identified in the 105 seed sentences; no manual intervention was used to trim this number down.

3.2 Evaluation Methods for Word Abstractness System

In replicating this study, we used the abstractness ratings from the authors. The corpus sentences were tagged using OpenNLP POSTagger [26] and all function words were removed. All words not found on the list of abstractness ratings (after reduced to their lemmatized form using Morpha [27]) were removed; empty slots in the feature vector (e.g., if there were no adjectives) were filled with a value of .5 for abstractness, following the original system. We started with the five attributes given by [7] and discussed above; we then augmented the feature vector with four additional variables: (6) average abstractness of all words, eliminating the grammatical distinction between them; (7) average abstractness of all words except for the target word; (8) the difference between the average abstractness of the sentence with the target word and without it; (9) the standard deviation of all the words. We tested these additional attributes because of the hypothesis that metaphors will cause mismatches between the total abstractness and the target word's abstractness.

3.3 Evaluation Methods for Semantic Distance / Similarity System

The evaluation of this approach tested a different distributional method for determining semantic similarity, Iosif's SemSim system [29]. There were two main reasons for not using the NGD [24] measure: (1) the test corpus had function words removed and other words reduced to their stems; the NGD results would not have taken this into account; (2) SemSim is more transparent in terms of its methodology and in terms of the corpus used. In this case, we used the American National Corpus (henceforth, OANC [31]), which consists of 14 million words taken from spoken and written contemporary American English. Thus, it has sources comparable to those used to create COCA, from which the test utterances were drawn (but COCA is not available to run SemSim on). The corpus was made comparable to the evaluation data by removing the most frequent functions words and running the Morpha analyzer to retrieve the lemmatized forms. SemSim's lexical classification system was then run on the entire OANC corpus for every word present in the evaluation data (H, the contextual window, was set at 2), creating an 8,690x8,690 matrix of similarity scores.

The pairwise similarity between words, comparable to NGD, was used to compute the 5 variables used in Sporleder and Li's system. To this we added four additional variables to test additional hypotheses: (6) the standard deviation of the similarity between the target word and the context; (7) the standard deviation of the similarity within the context. These were added to test the hypothesis that metaphor comes from a different source from the literal context, causing a mismatch in their similarity/distance. These caused two further variables to be included: (8) the difference between the standard deviations in similarity scores within the context and between target and context; and (9) the marker for negative differences in standard deviations that corresponds with variable (4) from the original study.

3.4 Evaluation Methods for Domain Interaction System

The domain interaction system has been implemented for the purposes of this study using a feature vector of variables 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. This feature vector was evaluated using the same learning algorithms as the abstraction and similarity systems.

In creating this feature vector, four knowledge-bases (three existing and one new) were used: (1) the SUMO ontology; (2) WordNet synsets; (3) mappings between WordNet and SUMO; (4) domain and event-status properties of the SUMO concepts. The knowledge-bases used in the evaluation which are not available elsewhere can be found at http://www.jdunn.name.

4 Results

This section presents the results of the evaluations. Note that the different class comparisons (e.g., three-way vs. non-metaphor) will influence only the systems based on feature vectors. The "Joint" system takes variables from all the systems which use feature vectors (the abstractness, similarity, and domain interaction systems). To make the comparison as consistent as possible, evaluation of the semantic similarity / distance, word abstractness, joint, and domain interaction (MIMIL in the tables) systems was done, following [7], using Weka's implementation of the logistic regression learning algorithm. All instances were normalized before training and testing; the evaluations were performed using cross-validation (100 folds). The F-measures reported here are for metaphor classification only (i.e., precision for non-metaphor is not directly considered because this inflates the performance of the systems). This is done because some of the systems greatly over-identify literal utterances; however, because literal utterances dominate the evaluation data set, the over-identification of literal utterances would disproportionately raise the average F-measure for all classes in these systems. The feature vectors and other material used in the evaluation can be found at http://www.jdunn.name.

Table 1. Three-way distinction between metaphor, humor, and literal in all domains

System	True Pos.	False Pos.	True Neg.	False Neg.	F-Meas.
Similarity	1	0	2,482	504	0.004
Abstractness	1	2	2,482	505	0.004
Joint	67	44	2,446	444	0.215
MIMIL	133	382	2,437	63	0.374
Source-Tar.	113	461	2,038	300	0.229

As shown in Table 1, when tested on the three-way distinction between metaphor, humor, and literal utterances, the similarity and abstractness systems performed very poorly, essentially identifying no metaphors. The joint system performed worse than the domain interaction system, showing that the abstractness and similarity features reduce performance. As shown in later tests, the measurements of abstractness and semantic similarity, both at the word-level, simply do not distinguish between metaphor and non-metaphor in a realistic data set. The domain interaction and source-target mapping systems performed much better. Both systems identified a similar number of metaphors (133 and 113), but the domain interaction system had somewhat fewer false positives (382 vs. 461). More importantly, the source-target mapping system had a significantly higher number of false negatives (300 vs. 63). Using a higher number of seed metaphors would have lowered the source-target mapping system's false negative rate, but at the same time that would likely have raised the already high false positive rate.

Table 2. Two-way distinction with and without humor present in all domains

System	Data	True Pos.	False Pos.	True Neg.	False Neg.	F-Meas.
Similarity	+Humor	0	0	2,489	506	0.000
Abstractness	+Humor	1	3	2,486	505	0.004
Joint	+Humor	33	15	2,475	478	0.118
MIMIL	+Humor	90	31	2,469	425	0.283
Source-Tar.	+Humor	113	461	2,038	300	0.229
Similarity	-Humor	0	0	1,989	506	0.000
Abstractness	-Humor	2	5	1,984	504	0.008
Joint	-Humor	62	28	1,964	449	0.206
MIMIL	-Humor	125	46	1,954	390	0.364
Source-Tar.	-Humor	113	373	1,625	300	0.251

Table 2 shows that when tested using a two-way distinction that conflates literal and humorous utterances into a single non-metaphoric class (+Humor), the performance of MIMIL drops significantly, showing that humor is distinct from both metaphor and non-humor. The similarity and abstractness systems continue to perform very poorly; the joint system continues to perform more poorly than the domain interaction system on its own. One advantage of the source-target mapping system over the implementation of MIMIL evaluated here is that its identifications do not depend on the make up of the data set (only on the seed metaphors). Thus, its performance remains constant while MIMIL has more false negatives when humor and literal utterances are conflated into a single class. With humor removed altogether (-Humor), results similar to the three-way evaluation are achieved. Similarity and abstractness continue to perform poorly. MIMIL and the source-target mapping system identify a comparable number of metaphors (125 and 113). In this evaluation, however, MIMIL produces more false negatives (390 vs. 300) while the source-target mapping system produces more false positives (373 vs. 46).

As shown in Table 3, the source-target mapping system and the domain interaction system perform similarly within the ABSTRACT domain (the other systems are not shown here because their performance was too low). However, the performance of MIMIL is significantly less than on the data set as a whole (0.276 vs. 0.374 F-measure) while the source-target mapping system performs at the same level (0.239 vs. 0.229 F-measure). Within the MENTAL domain the source-target mapping system identifies more metaphors than MIMIL, but continues to have more false positives. MIMIL has more false negatives. Within the PHYSICAL domain, MIMIL greatly out-performs the source-target mapping system (0.629 vs. 0.268 F-measure). Further, within this domain both systems perform better than they did in any other domain. On the other hand, in the SOCIAL domain both systems perform more poorly than in any other domain. Here, also, the roles are reversed: the source-target mapping system significantly out performs MIMIL, which identifies almost no metaphors.

Table 3. Two-way distinction with humor present, by domain

System	Domain	True Pos.	False Pos.	True Neg.	False Neg.	F-Meas.
MIMIL	Abstract	25	16	609	115	0.276
Source-Tar.	Abstract	29	99	526	85	0.239
MIMIL	Mental	23	9	617	102	0.293
Source-Tar.	Mental	33	130	496	67	0.251
MIMIL	Physical	66	19	606	59	0.629
Source-Tar.	Physical	32	107	517	68	0.268
MIMIL	Social	4	1	623	121	0.062
Source-Tar.	Social	19	125	499	80	0.156

5 Conclusions

We can draw several interesting and useful conclusions from this evaluation. First, we see the importance of a justified theory underlying metaphor identification systems. Both the source-target mapping system and the domain interaction system (MIMIL) are concerned with explaining and justifying their choices; both greatly out-perform the systems which are not as firmly grounded in theory. Second, we see that domain membership has a significant influence on the performance of the systems. Third, we see that the two top systems have their best performance on different domains and classes.

This suggests that there are multiple types of metaphor and that each system is stronger at identifying one type over another type. In other words, some of the assumptions about metaphor discussed in the first part of this study are mutually exclusive, but others are not. For example, some metaphor identification systems assume that source-target mappings are explicitly present, some assume that they are indirectly present, and others assume that they are not relevant for metaphor identification. It is likely that some, but not all, metaphors have a mediated or unmediated mapping and that other metaphors have no mapping at all [13]. If this is the case, a synthesis of approaches to metaphor identification might allow stronger coverage overall. In other words, if there are multiple types of metaphor, and if the systems perform better on some types of metaphor as a result of their assumptions about metaphor, then a full-coverage metaphor identification system should include multiple synthesized methods.

For example, the source-target mapping system and the domain interaction system could be synthesized by ordering the application of the methods within a single system. The source-target mapping system could be run first and search for explicitly present mappings. This would miss many metaphors that do not have an explicit source-target mapping (e.g., metaphors whose mapping was mediated and thus not directly present, metaphors whose mapping was not present in a noun-verb relation, or metaphors which did not have an underlying conceptual mapping to begin with). The domain interaction system could then be run second in order to identify the metaphors which did not fall into the rather narrow scope of the source-target mapping system.

References

1. Shutova, E.: Models of metaphor in NLP. In: Hajiv, J., Carberry, S., Clark, S., Nivre, J. (eds.) Proceedings of ACL 2010, pp. 688–697. Association for Computational Linguistics, Stroudsburg (2010)
2. Shutova, E., Teufel, S.: Metaphor corpus annotated for source – target domain mappings. In: Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Rosner, M., Tapias, D. (eds.) Proceedings of LREC 2010, pp. 3255–3261. European Language Resources Association, Paris (2010)
3. Shutova, E., Sun, L., Korhonen, A.: Metaphor identification using verb and noun clustering. In: Huang, C., Jurafsky, D. (eds.) Proceedings of COLING 2010, pp. 1002–1010. Tsinghua University Press, Beijing (2010)
4. Shutova, E., Teufel, S., Korhonen, A.: Statistical metaphor processing. Computational Linguistics 39 (2013) (forthcoming)
5. Neuman, Y., Nave, O.: Metaphor-based meaning excavation. Information Sciences 179, 2719–2728 (2009)
6. Turney, P., Littman, M.: Measuring praise and criticism: Inference of semantic orientation from association. ACM Transactions on Information Systems 21, 315–346 (2003)
7. Turney, P., Neuman, Y., Assaf, D., Cohen, Y.: Literal and metaphorical sense identification through concrete and abstract context. In: Barzilay, R., Johnson, M. (eds.) Proceedings of EMNLP 2011, pp. 680–690. Association for Computational Linguistics, Stroudsburg (2011)
8. Li, L., Sporleder, C.: Using Gaussian Mixture Models to detect figurative language in context. In: Kaplan, R., Burstein, J., Harper, M., Penn, G. (eds.) Proceedings of HLT-NAACL 2010, pp. 297–300. Association for Computational Linguistics, Stroudsburg (2010)
9. Sporleder, C., Li, L.: Contextual idiom detection without labelled data. In: Koehn, P., Mihalcea, R. (eds.) Proceedings of EMNLP 2009, pp. 315–323. Association for Computational Linguistics, Stroudsburg (2009)
10. Niles, I., Pease, A.: Towards a Standard Upper Ontology. In: Welty, C., Barry, C. (eds.) Proceedings of FOIS 2001, pp. 2–9. Association for Computational Linguistics, Stroudsburg (2001)
11. Niles, I., Pease, A.: Linking lexicons and ontologies: Mapping WordNet to the Suggested Upper Merged Ontology. In: Arabnia, H. (ed.) Proceedings of IEEE Intl. Conf. on Inf. and Knowl. Eng. (IKE 2003), pp. 412–416. IEEE Press, New York (2003)
12. Dunn, J.: Gradient semantic intuitions of metaphoric expressions. Metaphor and Symbol 26, 53–67 (2011)
13. Dunn, J.: How linguistic structure influences and helps to predict metaphoric meaning. Cognitive Linguistics 24, 33–66 (2013)
14. Davies, M.: The 385+ million word Corpus of Contemporary American English (1990–2008+): Design, architecture, and linguistic insights. International Journal of Corpus Linguistics 14, 159–190 (2009)
15. Lakoff, G., Johnson, M.: Metaphors We Live By. University of Chicago Press, Chicago (1980)
16. Lakoff, G., Johnson, M.: Philosophy in the Flesh: The embodied mind and its challenge to western thought. Basic Books, New York (1999)
17. Briscoe, E., Carroll, J., Watson, R.: The second release of the RASP system, pp. 77–80 (2006)

18. Sun, L., Korhonen, A.: Improving verb clustering with automatically acquired selectional preferences. In: Koehn, P., Mihalcea, R. (eds.) Proceedings of EMNLP 2009, pp. 638–647. Association for Computational Linguistics, Stroudsburg (2009)
19. Sun, L., Korhonen, A., Krymolowski, Y.: Verb Class Discovery from Rich Syntactic Data. In: Gelbukh, A. (ed.) CICLing 2008. LNCS, vol. 4919, pp. 16–27. Springer, Heidelberg (2008)
20. Shannon, B.: Metaphor: From fixedness and selection to differentiation and creation. Poetics Today 13, 659–685 (1992)
21. Barnden, J.: Metaphor and metonymy: Making their connections more slippery. Cognitive Linguistics 21, 1–34 (2010)
22. Nirenburg, S., Raskin, V.: Ontological Semantics. MIT Press, Cambridge (2004)
23. Raskin, V., Nirenburg, S.: An applied ontological semantic microtheory of adjective meaning for natural language processing. Machine Translation 13, 135–227 (1998)
24. Cilibrasi, R., Vitanyi, P.: The Google similarity distance. IEEE Transactions on Knowledge and Data Engineering 19, 370–383 (2007)
25. Gibbs, R.: Literal meaning and psychological theory. Cognitive Science 8, 275–304 (1984)
26. Apache: OpenNLP (2011), http://opennlp.apache.org
27. Guido, M., Carroll, J., Pearce, D.: Applied morphological processing of English. Natural Language Engineering 7, 207–223 (2001)
28. Princeton, U.: WordNet (2012), http://wordnet.princeton.edu/
29. Iosif, E., Potamianos, A.: SemSim: Resources for normalized semantic similarity computation using lexical networks. In: Calzolari, N., Choukri, K., Declerck, T., Doğan, M., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S. (eds.) Proceedings of LREC 2012, pp. 3499–3504. European Language Resources Association, Paris (2012)
30. Witten, I., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kaufmann, San Francisco (2005)
31. Ide, N., Suderman, K.: The American National Corpus first release. In: Lino, M., Xavier, M., Ferreira, F., Costa, R., Silva, R. (eds.) Proceedings of LREC 2004, pp. 1681–1684. European Language Resources Association, Paris (2004)