Robust Extraction of Metaphors from Novel Data

Tomek Strzalkowski[1], George Aaron Broadwell[1], Sarah Taylor[2], Laurie Feldman[1], Boris Yamrom[1], Samira Shaikh[1], Ting Liu[1], Kit Cho[1], Umit Boz[1], Ignacio Cases[1] and Kyle Elliott[3]

[1]State University of New York
University at Albany
Albany NY USA 12222
tomek@albany.edu

[2]Sarah M. Taylor Consulting LLC
121 South Oak St.
Falls Church VA USA 22046
talymail59@gmail.com

[3]Plessas Experts
Network Inc.
Herndon VA 20171
kelliot@plessas.net

Abstract

This article describes our novel approach to the automated detection and analysis of metaphors in text. We employ robust, quantitative language processing to implement a system prototype combined with sound social science methods for validation. We show results in 4 different languages and discuss how our methods are a significant step forward from previously established techniques of metaphor identification. We use Topical Structure and Tracking, an Imageability score, and innovative methods to build an effective metaphor identification system that is fully automated and performs well over baseline.

1 Introduction

The goal of this research is to automatically identify metaphors in textual data. We have developed a prototype system that can identify metaphors in naturally occurring text and analyze their semantics, including the associated affect and force. Metaphors are mapping systems that allow the semantics of a familiar Source domain to be applied to a Target domain so that new frameworks of reasoning can emerge in the Target domain. Metaphors are pervasive in discourse, used to convey meanings indirectly. Thus, they provide critical insights into the preconceptions, assumptions and motivations underlying discourse, especially valuable when studied across cultures. When metaphors are thoroughly understood within the context of a culture, we gain substantial knowledge about cultural values. These insights can help better shape cross-cultural understanding and facilitate discussions and negotiations among different communities.

A longstanding challenge, however, is the large-scale, automated identification of metaphor in volumes of data, and especially the interpretation of their complex, underlying semantics.

We propose a data-driven computational approach that can be summarized as follows: Given textual input, we first identify any sentence that contains references to Target concepts in a given Target Domain (Target concepts are elements that belong to a particular domain; for instance "government bureaucracy" is a Target concept in the "Governance" domain). We then extract a passage of length 2N+1, where N is the number of sentences preceding (or succeeding) the sentence with Target Concept. We employ dependency parsing to determine the syntactic structure of each input sentence. Topical structure and imageability analysis are then combined with dependency parsing output to locate the candidate metaphorical expressions within a sentence. For this step, we identify nouns and verbs in the passage (of length 2N+1) and link their occurrences – including repetitions, pronominal references, synonyms and hyponyms. This linking uncovers the topical structure that holds the narrative together. We then locate content words that are outside the topical structure and compute their imageability scores. Any nouns or adjectives outside the main topical structure that also have high imageability scores and are dependency-linked in the parse structure to the Target Concept are identified as candidate *source relations*, i.e., expressions borrowed from a Source domain to describe the Target concept. In addition, any verbs that have a direct dependency on the Target Con-

Proceedings of the First Workshop on Metaphor in NLP, pages 67–76,
Atlanta, Georgia, 13 June 2013.

cept are considered as candidate relations. These candidate relations are then used to compute and rank proto-sources. We search for their arguments in a balanced corpus, assumed to represent standard use of the language, and cluster the results. Proto-source clusters and their ranks are exploited to determine whether the candidate relations are metaphorical or literal. Finally, we compute the affect and force associated with the metaphor.

Our approach is shown to work in four languages – American English, Mexican Spanish, Russian Russian and Iranian Farsi. We detail in this paper the application of our approach to detection of metaphors using specific examples from the "Governance" domain. However, our approach can be expanded to work on extracting metaphors in any domain, even unspecified ones. We shall briefly explain this in Section 5; we defer the details of the expanded version of the algorithm to a separate larger publication. In addition, we shall primarily present examples in English to illustrate details of our algorithms. However, modules for all four languages have the same implementation in our system.

The rest of the paper is organized as follows: in Section 2, we discuss related research in this field. Section 3 presents our approach in detail; Section 4 describes our evaluation and results. In Section 5 we discuss our conclusions and future directions.

2 Related Work

Most current research on metaphor falls into three groups: (1) theoretical linguistic approaches (as defined by Lakoff & Johnson, 1980; and their followers) that generally look at metaphors as abstract language constructs with complex semantic properties; (2) quantitative linguistic approaches (e.g., Charteris-Black, 2002; O'Halloran, 2007) that attempt to correlate metaphor semantics with their usage in naturally occurring text but generally lack robust tools to do so; and (3) social science approaches, particularly in psychology and anthropology that seek to explain how people deploy and understand metaphors in interaction, but which lack the necessary computational tools to work with anything other than relatively isolated examples.

Metaphor study in yet other disciplines has included cognitive psychologists (e.g., Allbritton, McKoon & Gerrig, 1995) who have focused on the way metaphors may signify structures in human memory and human language processing. Cultural anthropologists, such as Malkki in her work on refugees (1992), see metaphor as a tool to help outsiders interpret the feelings and mindsets of the groups they study, an approach also reflective of available metaphor case studies, often with a Political Science underpinning (Musolff, 2008; Lakoff, 2001).

In computational investigations of metaphor, knowledge-based approaches include MetaBank (Martin, 1994), a large knowledge base of metaphors empirically collected. Krishnakumaran and Zhu (2007) use WordNet (Felbaum, 1998) knowledge to differentiate between metaphors and literal usage. Such approaches entail the existence of lexical resources that may not always be present or satisfactorily robust in different languages. Gedigan et al (2006) identify a system that can recognize metaphor. However their approach is only shown to work in a narrow domain (Wall Street Journal, for example).

Computational approaches to metaphor (largely AI research) to date have yielded only limited scale, often hand designed systems (Wilks, 1975; Fass, 1991; Martin, 1994; Carbonell, 1980; Feldman & Narayan, 2004; Shutova & Teufel, 2010; inter alia, also Shutova, 2010b for an overview). Baumer et al (2010) used semantic role labels and typed dependency parsing in an attempt towards computational metaphor identification. However they self-report their work to be an initial exploration and hence, inconclusive. Shutova et al (2010a) employ an unsupervised method of metaphor identification using nouns and verb clustering to automatically impute metaphoricity in a large corpus using an annotated training corpus of metaphors as seeds. Their method relies on annotated training data, which is difficult to produce in large quantities and may not be easily generated in different languages.

By contrast, we propose an approach that is fully automated and can be validated using empirical social science methods. Details of our algorithm follow next.

3 Our Approach

In this section, we walk through the steps of metaphor identification in detail. Our overall algorithm

consists of five main steps from obtaining textual input to classification of input as metaphorical or literal.

3.1 Passage Identification

The input to our prototype system is a piece of text. This text may be taken from any genre – news articles, blogs, magazines, official announcements, broadcast transcripts etc.

Given the text, we first identify sentences that contain Target concepts in the domain we are interested in. Target concepts are certain keywords that occur within the given domain and represent concepts that may be targets of metaphor. For instance, in the "Governance" domain, concepts such as "federal bureaucracy" and "state mandates" serve as Target concepts. We keep a list of Target concepts to search through when analyzing given input. This list can be automatically created by mining Target Concepts from resource such as Wikipedia, given the Target domain, or manually constructed. Space limits the discussion of how such lists may be automatically created; a separate larger publication addresses our approach to this task in greater detail.

In Figure 1, we show a piece of text drawn from a 2008 news article. The sentence in italics contains one of our Target concepts: "federal bureaucracy". We extract the sentence containing Target concepts that match any of those in our list, including N sentences before and N sentences after the sentence if they exist, to yield a passage of at most 2N+1 sentences. For the example shown in Figure 1, the Target concept is "federal bureaucracy". In current system prototype, N=2. Hence, we extract two sentences prior to the sentence containing "federal bureaucracy" (in Figure 1 example, these are omitted for ease of presentation) and two sentences following the given sentence.

Once this passage is extracted, we need to determine whether a metaphor is present in the middle sentence. To accomplish that, we follow the steps as described in the next section.

These qualities[1] have helped him[4] ***navigate the labyrinthine*** *federal bureaucracy in his demanding $191,300-a-year job as the top federal official[3] responsible for bolstering airline, border[2], port and rail security against a second catastrophic terrorist attack.*

But those same personal qualities[1] also explain why the 55-year-old Cabinet officer[3] has alienated so many Texans along the U.S.-Mexico border[2] with his[4] relentless implementation of the Bush administration's hard-nosed approach to immigration enforcement - led by his unyielding push to construct 670 miles of border[2] fencing by the end of the year.

Some Texas officials are so exasperated that they say they'll just await the arrival of the next president before revisiting border enforcement with the federal government.

Figure 1. Excerpt from news article. Passage containing target concept highlighted in italics. The callouts [1], [2] etc., indicate topic chains (see next section).

3.2 Topical Structure and Imageability Analysis

Our hypothesis is that metaphorically used terms are typically found outside the topical structure of the text. This is an entirely novel method of effectively selecting candidate relations. It draws on Broadwell et al. (2012), who proposed a method to establish the topic chains in discourse as a means of modeling associated socio-linguistic phenomena such as topic control and discourse cohesiveness. We adapted this method to identify and exclude any words that serve to structure the core discussion, since the metaphorical words, except in the cases of extended and highly elaborated metaphors, are not the main subject, and thus unlikely to be repeated or referenced in the context surrounding the sentence.

We link the occurrences of each noun and verb in the passage (5 sentence length). Repetitions via synonyms, hyponyms, lexical variants and pronoun references are linked together. These words, as elements of the several topic chains in a text, are then excluded from further consideration. WordNet (Felbaum, 1998) is used to look up synonyms and hyponyms of the remaining content words. We

illustrate this in Figure 1. We show the two sentences that form the latter context in the example passage. We show four of the topic chains discovered in this passage. These have been labeled via superscripts in Figure 1. [1] and [2] are the repetitions of word "qualities" and "border". The [3] identifies repetition via lexical variants "officer" and "official" and [4] identifies the pronoun co-references "him" and "his". We shall exclude these words from consideration when searching for candidate metaphorical relations in the middle sentence of the passage.

To further narrow the pool of candidate relations in this sentence, we compute the imageability scores of the remaining words. The hypothesis is metaphors use highly imageable words to convey their meaning. The use of imageability scores for the primary purpose of metaphor detection distinguishes our approach from other research on this problem. While Turney et al. (2011) explored the use of word concreteness (a concept related but not identical to imageability) in an attempt to disambiguate between abstract and concrete verb senses, their method was not specifically applied to detection of metaphors; rather it was used to classify verb senses for the purpose of resolving textual entailment. Broadwell et al. (2013) present a detailed description of our approach and how we use imageability scores to detect metaphors.

Our assertion is that any highly imageable word is more likely to be a metaphorical relation. We use the MRCPD (Coltheart 1981, Wilson 1988) expanded lexicon to look up the imageability scores of words not excluded via the topic chains. Although the MRCPD contains data for over 150,000 words, a major limitation of the database for our purposes is that the MRCPD has imageability ratings (i.e., how easily and quickly the word evokes a mental image) for only ~9,240 (6%) of the total words in its database. To fill this gap, we expanded the MRCPD database by adding imagery ratings for an further 59,989 words. This was done by taking the words for which the MRCPD database has an imageability rating and using that word as an index to synsets determined using WordNet (Miller, 1995). The expansion and validation of the expanded MRCPD imageability rating is presented in a separate, future publication.

Words that have an imageability rating lower than an experimentally determined threshold are further excluded from consideration. In the example shown in Figure 1, words that have sufficiently high imageability scores are "labyrinthine", "port", "rail" and "airline". We shall consider them as candidate relations, to be further investigated, as explained in the dependency parsing step described next.

3.3 Relation Extraction

Dependency parsing reveals the syntactic structure of the sentence with the Target concept. We use the Stanford parser (Klein and Manning, 2003) for English language data. We identify candidate metaphorical relations to be any verbs that have the Target concept in direct dependency path (other than auxiliary and modal verbs). We exclude verbs of attitude ("think", "say", "consider"), since these have been found to be more indicative of metonymy than of metaphor. This list of attitude verbs is automatically derived from WordNet.

From the example shown in Figure 1, one of the candidate relations extracted would be the verb "navigate".

In addition, we have a list of candidate relations from Step 3.2, which are the highly imageable nouns and adjectives that remain after topical structure analysis. Since "port", "rail" and "airline" do not have a direct dependency path to our Target concept of "federal bureaucracy", we drop these from further consideration. The highly imageable word remaining in this list is "labyrinthine".

Thus, two candidate relations are extracted from this passage – "navigate" and "labyrinthine". We shall now show how we use these to discover proto-sources for the potential metaphor.

3.4 Discovery of Proto-sources

Once candidate relations are identified, we examine whether the usage of these relations is metaphorical or literal. To determine this, we search for all uses of these relations in a balanced corpus and examine in which contexts the candidate relations occur. To demonstrate this via our example, we shall consider one of the candidate relations identified in Figure 1 – "navigate"; the search method is the same for all candidate relations identified. In the case of the verb "navigate" we search a balanced corpus for the collocated words, that is, those that occur within a 4-word window following the verb, with high mutual information (>3) and occurring together in the corpus with a frequency

at least 3. This search returns a list of words, mostly nouns in this case, that are the objects of the verb "navigate", just as "federal bureaucracy" is the object in the given example. However, since the search occurs in a balanced corpus, given the parameters we search for, we discover words where the objects are literally navigated. Given these search parameters, the top results we get are generally literal uses of the word "navigate". We cluster the resulting literal uses as semantically related words using WordNet and corpus statistics. Each such cluster is an emerging prototype source domain, or a proto-source, for the potential metaphor.

In Figure 2, we show three of the clusters obtained when searching for the literal usage of the verb "navigate". We use elements of the clusters to give names or label the proto-source domains. WordNet hypernyms or synonyms are used in most cases. The clusters shown in Figure 2 represent three potential source domains for the given example, the labels "MAZE", "WAY" and "COURSE" are derived from WordNet.

1. Proto-source Name: **MAZE**
Proto-source Elements: [mazes, system, networks]
IMG Score: 0.74
2. Proto-source Name: **WAY**
Proto-source Elements: [way, tools]
IMG Score: 0.60
3. Proto-source Name: **COURSE**
Proto-source Elements: [course, streams]
IMG: 0.55

Figure 2. Three of several clusters obtained from balanced corpus search for objects of verb "navigate".

We rank the clusters according to the combined frequency of cluster elements in the balanced corpus. In a similar fashion, clusters are obtained for the candidate relation "labyrinthine"; however here we search for the nouns modified by the adjective "labyrinthine".

3.5 Estimation of Linguistic Metaphor

A ranked list of proto-sources from the previous step serves as evidence for the presence of a metaphor.

If any Target domain elements are found in the top two ranked clusters, we consider the phrase being investigated to be literal. This eliminates examples where one of the most frequently encountered sources is within the target domain.

If neither of the top two most frequent clusters contains any elements from the target domain, we then compute the average imageability scores for each cluster from the mean imageability score of the cluster elements. If no cluster has a sufficiently high imageability score (experimentally determined to be >.50 in the current prototype), we again consider the given input to be literal. This step reinforces the claim that metaphors use highly imageable language to convey their meaning. If a proto-source cluster is found to meet both criteria, we consider the given phrase to be metaphorical. For the example shown in Figure 1, our system finds "navigate the …federal bureaucracy" to be metaphorical. One of the top Source domains identified for this metaphor is "MAZE". Hence the conceptual metaphor output for this example can be:

"FEDERAL BUREAUCRACY IS A MAZE".

Our system can thus classify input sentences as metaphorical or literal by the series of steps outlined above. In addition, we have modules that can determine a more complex conceptual metaphor, based upon evidence of one or more metaphorical passages as identified above. We do not discuss those modules in this article. Once a metaphor is identified, we compute associated Mappings, Affect and Force.

3.6 Mappings

In the current prototype system, we assign metaphors to one of three types of mappings. Propertive mappings – which state what the domain objects

Relation type	**Type 1 (property)** $T \rightarrow Rel$	**Type 2 (agentive)** $T \rightarrow Rel \rightarrow X$		**Type 3 (patientive)** $X \rightarrow Rel \rightarrow T$	
Relation/X		$X \geq$ *neutral*	$X <$ *neutral*	$X \geq$ *neutral*	$X <$ *neutral*
Rel > Positive	POSITIVE	POSITIVE	$\leq$ UNSYMP	POSITIVE	$\leq$ SYMPAT
Rel < Negative	NEGATIVE	$\leq$ UNSYMP	$\geq$ SYMPAT	$\geq$ SYMPAT	$\geq$ SYMPAT
Rel = Neutral	NEUTRAL	NEUTRAL	$\leq$ NEUTRAL	NEUTRAL	$\leq$ NEUTRAL

Table 1. Algorithm assigns affect of metaphor based upon mappings.

are and descriptive features; Agentive mappings – which describe what the domain elements do to other objects in the same or different domains; and Patientive mappings – which describe what is done to the objects in these domains. These are broad categories to which relations can, with some exceptions be assigned at the linguistic metaphor level by the parse tag of the relation. Relations that take Target concepts as objects are usually Patientive relations. Similarly, relations that are Agentive take Target concepts as subjects. Propertive relations are usually determined by adjectival relations.

Once mappings are assigned, we can use them to group linguistic metaphors. A set of linguistic metaphors on the same or semantically equivalent Target concepts can be grouped together if the relations are all agentive, patientive or propertive. The mapping assigned to set of examples in Figure 3 is Patientive.

One immediate consequence of the proposed approach is the simplicity with which we can represent domains, their elements, and the metaphoric mappings between domains. Regardless of what specific relations may operate within a domain (be it Source or Target), they can be classified into just 3 categories. We are further expanding this module to include semantically richer distinctions within the mappings. This includes the determination of the sub-dimensions of mappings i.e. assigning groups of relations to a semantic category.

3.7 Affect and Force

Affect of a metaphor may be positive, negative or neutral. Our affect estimation module computes an affect score taking into account the relation, Target concept and the subject or object of the relation based on the dependency between relation and Target concept. The algorithm is applied according to the categories shown in Table 1.

The expanded ANEW lexicon (Bradley and Lang, 2010) is used to look up affect scores of words. ANEW assigns scores from 0 (highly negative) to 9 (highly positive); 5 being neutral. We compute the affect of a metaphorical phrase within a sentence by summing the affect scores of the relation and its object or subject. If the relation is agentive, we then look at the object in source domain that the Target concept is acting upon. If the object (denoted in above table as X) has an affect score that is greater than neutral, and the relation itself has an affect score that is greater than neutral, then a POSITIVE affect is assigned to the metaphor. This is denoted by the cell at the intersection of the row labeled "Rel > Positive" and the 3rd column in Table 1. Similarly affect for the other mapping categories can be assigned.

1. His attorney described him as a family man who was lied to by a friend and who got **tangled in** federal bureaucracy he knew nothing about.
2. The chart, composed of 207 boxes illustrates the **maze of** federal bureaucracy that would have been created by then-President Bill Clinton's relation health reform plan in the early 1990s.
3. "Helping my constituents **navigate** the federal bureaucracy is one of the most important things I can do," said Owens.
4. A Virginia couple has donated $1 million to help start a center at Arkansas State University meant to help wounded veterans **navigate** the federal bureaucracy as they return to civilian life.

Figure 3. Four metaphors for the Target concept "federal bureaucracy".

We also seek to determine the impact of metaphor on the reader. This is explored using the concept of Force in our system. The force of a metaphor is estimated currently by the commonness of the expression in the given Target domain. We compute the frequency of the relation co-occurring with Target concept in a corpus of documents in the given Target domain. This frequency represents the commonness of expression, which is the inverse of Force. The more common a metaphorical expression is, the lesser its force.

For the example shown below in Figure 4, the affect is computed to be positive ("navigate" and "veterans" are both found to have positive affect scores, the relation is patientive). The force of this expression is low, since its commonness is 742 (commonness score > 100 is high commonness, determined experimentally).

A Virginia couple has donated $1 million to help start a center at Arkansas State University meant to help wounded *veterans* **navigate** the federal bureaucracy as they return to civilian life.

Figure 4. Example of metaphor with positive affect and low force.

The focus of this article is the automatic identification of metaphorical sentences in naturally occurring text. Affect and force modules are utilized to understand metaphors in context and contrast them across cultures, if feasible. We defer more detailed discussion of affect and force and their implications to a future, larger article.

4 Evaluation and Results

In order to determine the efficacy of our system in classifying metaphors as well as to validate various system modules such as affect and force, we performed a series of experiments to collect human validation of metaphors in a large set of examples.

4.1 Experimental Setup

We constructed validation tasks that aimed at performing evaluation of linguistic metaphor extraction accuracy. The first task – Task 1, consists of a series of examples, typically 50, split more or less equally between those proposed by the system to be metaphorical and those proposed to be literal. This task was designed to elicit subject and expert judgments on several aspects related to the presence or absence of linguistic metaphors in text. Subjects are presented with brief passages where a Target concept and a relation are highlighted. They are asked to rank their responses on a 7-point scale for the following questions:

Q1: To what degree does the above passage use metaphor to describe the highlighted concept?
Q2: To what degree does this passage convey an idea that is either positive or negative?
Q3: To what degree is it a common way to express this idea?

There are additional questions that ask subjects to judge the imageability and arousal of a given passage, which we do not discuss in this article. Q1 deals with assessing the metaphoricity of the example, Q2 deals with affect and Q3 deals with force.

Each instance of Task 1 consists of a set of instructions, training examples, and a series of passages to be judged. Instructions provide training examples whose ratings fall at each end the rating continuum. Following the task, participants take a grammar test to demonstrate native language proficiency in the target language. All task instances are then posted on Amazon's Mechanical Turk. The goal is to collect at least 30 valid judgments per task instance. We typically collect ~50 judgments from Mechanical Turkers, so that after filtering for invalid data which includes turkers selecting items at random, taking too little time to complete the task, grammar test failures, and other inconsistent data, we would still retain 30 valid judgments per passage. In addition to grammar test and time filter, we also inserted instance of known metaphors and known literal passages randomly within the Task. Any turker judgments that classify these known instance incorrectly more than 30% of the total known instance size are discarded.

The valid turker judgments are then converted to a binary judgment for the questions we presented. For example, for question Q1, the anchors to 7-point scale are 0 (none at all i.e. literal) to 7 (highly i.e metaphorical). We take [0, 2] as a literal judgment and [4, 6] as metaphorical and take a majority vote. If the majority vote is 3, we discard that passage from our test set, since it is undetermined whether the passage is literal or metaphorical.

We have collected human judgments on hundreds of metaphors in all four languages of interest. In Section 4.3, we explain our performance and compare our results to baseline where appropriate.

4.2 Test Reliability

The judgments collected from subjects are tested for reliability and validity. Reliability among the raters is computed by measuring intra-class correlation (ICC) (McGraw & Wong, 1996; Shrout & Fleiss, 1979). A coefficient value above 0.7 indicates strong reliability.
Table 3 shows the current reliability coefficients established for the selected Task 1 questions in all 4 languages. In general, our analyses have shown that with approximately 30 or more subjects we obtain a reliability coefficient of at least 0.7. We note that Russian and Farsi reliability scores are low in some categories, primarily due to lack of sufficient subject rating data. However, reliability of subject ratings for metaphor question (Q1) is sufficiently high in three of the four languages we are interested in.

Dimension	English	Spanish	Russian	Farsi
Metaphor	.908	.882	.838	.606
Affect	.831	.776	.318	.798
Commonness	.744	.753	.753	.618

Table 3. Intraclass correlations for linguistic metaphor assessment by Mechanical Turk subjects (Task 1)

4.3 Results

In Table 4, we show our performance at classifying metaphors across four different languages. The baseline in this table assigns all given examples in the test set to be metaphorical. We note that performance of the system at the linguistic metaphor level when compared to human gold standard is significantly over baseline for all four languages. The system performances cited in Table 4 validate the system against test sets that contain the distribution of metaphorical vs. literal examples as outlined in Table 5.

	English	Spanish	Russian	Farsi
Baseline	45.8%	41.7%	56.4%	50%
System	71.3%	80%	69.2%	78%

Table 4. Performance accuracy of system when compared to baseline for linguistic metaphor classification.

	English	Spanish	Russian	Farsi
Metaphor	50	50	22	25
Literal	59	70	17	25
Total	109	120	39	50

Table 5. Number of metaphorical and literal examples in test sets across all four languages.

Table 6 shows the accuracy in classification by the Affect and Force modules. We note that the low performance of affect and force for languages other than English. Our focus has been on improving NLP tools for Spanish, Russian and Farsi, so that a similar robust performance for those language can be achieved as we can demonstrate in English.

Accuracy	English	Spanish	Russian	Farsi
Affect	72%	54%	51%	40%
Force	67%	50%	33%	66%

Table 6. Affect and force performance of system on linguistic metaphor level.

5 Discussion and Future Work

In this article, we described in detail our approach to detecting metaphors in text. We have developed an automated system that does not require the existence of annotated training data or a knowledge base of predefined metaphors. We have described the various steps for detecting metaphors from receiving an input, to selecting candidate relations, to the discovery of prototypical source domains, and leading to the identification of a metaphor as well as the discovery of the potential source domain being applied in the metaphor. We presented two novel concepts that have heretofore not been fully explored in computational metaphor identification systems. The first is the exclusion of words that form the thread of the discussion in the text, by the application of a Topic Tracking module. The second is the application of Imageability scores in the selection of salient candidate relations.

Our evaluation consists first of validating the evaluation task itself. Once we ensure that sufficient reliability has been established on the various dimensions we seek to evaluate – metaphoricity, affect and force – we compare our system performance to the human gold standard. The performance of our system as compared to baseline is quite high, across all four languages of interest when measured against human assessed gold standard.

In this article, we discuss examples of metaphors belonging to a specific Target domain – "Governance". However, we can run our system through data in any domain perform the same kind of metaphor identification. In cases where the Target domain is unknown, we plan to use our Topic tracking module to recognize content words that may form part of a metaphorical phrase. This is essentially a process that is the reverse of that described in Section 3.3. We will find the salient Target concepts where there are directly dependent relations with the imageable verbs or adjectives.

In a separate larger publication, we plan to discuss in detail revisions to our Mapping module as well as the discovery and analyses of more complex conceptual metaphors. Such complex metaphors are based upon evidence from one or more instance of linguistic metaphors. Additional modules would recognize the manifold mappings, affect and force associated with the complex conceptual metaphors.

Acknowledgments

This research is supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of

Defense US Army Research Laboratory contract number W911NF-12-C-0024. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoD/ARL, or the U.S. Government.

References

Allbritton, David W., Gail McKoon, and Richard J. Gerrig. 1995. Metaphor-Based Schemas and Text Representations: Making Connections Through Conceptual Metaphors, *Journal of Experimental Psychology*: Learning, Memory, and Cognition, Vol. 21, No. 3, pp. 612-625.

Baumer, Erik. P.S., White, James., Tomlinson, Bill. 2010. Comparing Semantic Role Labeling with Typed Dependency *Parsing in Computational Metaphor Identification. Proceedings of the NAACL HLT 2010 Second Workshop on Computational Approaches to Linguistic Creativity, pages 14–22,* Los Angeles, California, June 2010.

Bradley, M.M. & Lang, P.J. 2010. Affective Norms for English Words (ANEW): Instruction manual and affective ratings. Technical Report C-2. University of Florida, Gainesville, FL.

Broadwell George A., Jennifer Stromer-Galley, Tomek Strzalkowski, Samira Shaikh, Sarah Taylor, Umit Boz, Alana Elia, Laura Jiao, Ting Liu and Nick Webb. 2012. Modeling Socio-Cultural Phenomena in Discourse. *Journal of Natural Language Engineering, Cambridge Press.*

Broadwell, George A., Umit Boz, Ignacio Cases, Tomek Strzalkowski, Laurie Feldman, Sarah Taylor, Samira Shaikh, Ting Liu, Kit Cho, aand Nick Webb. 2013. Using imageability and topic chaining to locate metaphors in linguistic corpora. in Ariel M. Greenberg, William G. Kennedy, Nathan D. Bos and Stephen Marcus, eds. *Proceedings of the 6th International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction SBP 2013.*

Carbonell, Jaime. 1980. Metaphor: a key to extensible semantic analysis. *Proceedings of the 18th Annual Meeting on Association for Computational Linguistics*.

Charteris-Black, Jonathan 2002 Second Language Figurative Proficiency: A Comparative Study of Malay and English. *Applied Linguistics* 23/1: 104-133.

Coltheart, M. 1981. The MRC Psycholinguistic Database. *Quarterly Journal of Experimental Psychology*, 33A, 497-505.

Fass, Dan. 1991. met*: A Method for Discriminating Metonymy and Metaphor by Computer. *Computational Linguistics, Vol 17:49-90*

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

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

Gedigian, M., Bryant, J., Narayanan, S., & Ciric, B. (2006). Catching Metaphors. *Proceedings of the Third Workshop on Scalable Natural Language Understanding ScaNaLU 06* (pp. 41-48). Association for Computational Linguistics.

Klein, Dan and Manning, Christoper D. 2003. Accurate Unlexicalized Parsing. *Proceedings of the 41st Meeting of the Association for Computational Linguistics*, pp. 423-430.

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

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

Lakoff, George. 2001. *Moral Politics: what Conservatives Know that Liberals Don't.* University of Chicago Press.

Malkki, Liisa. 1992. National Geographic: The Rooting of People and the Territorialization of National Identity Among Scholars and Refugees. *Society for Cultural Anthropology* 7(1):24-44

Martin, James. 1988. A Computational Theory of Metaphor. *PH.D. Dissertation*

McGraw, K. O., & Wong, S. P. 1996. Forming inferences about some intraclass correlation coefficients. *Psychological Methods, 1*(1), 30-46.

Musolff, Andreas. 2008. What can Critical Metaphor Analysis Add to the Understanding of Racist Ideology? Recent Studies of Hitler's Anti-Semitic Metaphors, Critical Approaches to Discourse Analysis across Disciplines, http://cadaad.org/ejournal, Vol. 2(2): 1-10.

O'Halloran, Kieran. 2007. Critical Discourse Analysis and the Corpus-informed Interpretation of Metaphor at the Register Level. *Oxford University Press*

Shrout, P. E., & Fleiss, J. L. 1979. Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86* (2), 420-428.

Shutova, E. 2010. Models of Metaphors in NLP. *In Proceedings of ACL 2010, Uppsala, Sweden.*

Shutova, E. and S. Teufel. 2010a. Metaphor corpus annotated for source - target domain mappings. *In Proceedings of LREC 2010, Malta.*

Shutova, E., T. Van de Cruys and A. Korhonen. 2012. *Unsupervised Metaphor Paraphrasing Using a Vector Space Model*, In Proceedings of COLING 2012, Mumbai, India

Turney, Peter., Yair Neuman, Dan Assaf, and Yohai Cohen. 2011. Literal and metaphorical sense identification through concrete and abstract context. In Proceedings of EMNLP, pages 680–690, Edinburgh, UK

Wilks, Yorick. 1975. Preference semantics. *Formal Semantics of Natural Language*, E. L. Keenan, Ed. Cambridge University Press, Cambridge, U.K., 329--348.

Wilson, M.D. (1988) The MRC Psycholinguistic Database: Machine Readable Dictionary, Version 2. *Behavioural Research Methods, Instruments and Computers, 20(1), 6-11*.