Catching Metaphors

Matt Gedigian, John Bryant, Srini Narayanan, and **Branimir Ciric**
International Computer Science Institute
1947 Center Street. Suite 600
Berkeley, CA 94704, USA
{gedigian, jbryant, snarayan}@icsi.berkeley.edu

Abstract

Metaphors are ubiquitous in language and developing methods to identify and deal with metaphors is an open problem in Natural Language Processing (NLP). In this paper we describe results from using a maximum entropy (ME) classifier to identify metaphors. Using the Wall Street Journal (WSJ) corpus, we annotated all the verbal targets associated with a set of frames which includes frames of spatial motion, manipulation, and health. One surprising finding was that over **90%** of annotated targets from these frames are used metaphorically, underscoring the importance of processing figurative language. We then used this labeled data and each verbal target's PropBank annotation to train a maximum entropy classifier to make this literal vs. metaphoric distinction. Using the classifier, we reduce the final error in the test set by 5% over the verb-specific majority class baseline and 31% over the corpus-wide majority class baseline.

1 Introduction

To move beyond "factoid" style questions, question answering systems must rely on inferential mechanisms. To answer such commonplace questions as *Which train should I take to get to the airport?* requires justifications, predictions and recommendations that can only be produced through inference.

One such question answering system (Narayanan and Harabagiu, 2004) takes PropBank/FrameNet annotations as input, uses the PropBank targets to indicate which actions are being described with which arguments and produces an answer using probabilistic models of actions as the tools of inference. Initiating these action models is called *simulation*.

Such action models provide deep inferential capabilities for embodied domains. They can also, when provided with appropriate metaphoric mappings, be extended to cover metaphoric language (Narayanan, 1997). Exploiting the inferential capabilities of such action models over the broadest domain requires a system to determine whether a verb is being used literally or metaphorically. Such a system could then activate the necessary metaphoric mappings and initiate the appropriate simulation.

2 Metaphor

Work in Cognitive Semantics (Lakoff and Johnson, 1980; Johnson, 1987; Langacker, 1987; Lakoff, 1994) suggests that the structure of abstract actions (such as *states*, *causes*, *purposes*, and *means*) are characterized cognitively in terms of *image schemas* which are *schematized* recurring patterns from the embodied domains of force, motion, and space.

Consider our conceptualization of events as exemplified in the mapping called the Event Structure Metaphor.

- States are locations (bounded regions in space).
- Changes are movements (into or out of bounded regions).

Proceedings of the 3rd Workshop on Scalable Natural Language Understanding, pages 41–48,
New York City, June 2006.

- Causes are forces.
- Actions are self-propelled movements.
- Purposes are destinations.
- Difficulties are impediments to motion.

This mapping generalizes over an extremely wide range of expressions for one or more aspects of event structure. For example, take states and changes. We speak of *being in* or *out of* a state, of *entering* or *leaving* it, of *getting to a state* or *emerging from* it. This is a rich and complex metaphor whose parts interact in complex ways. To get an idea of how it works, consider the submapping Difficulties are impediments to motion. In the metaphor, purposeful action is self-propelled motion toward a destination. A difficulty is something that impedes such motion. Metaphorical difficulties of this sort come in five types: blockages; features of the terrain; burdens; counterforces; lack of an energy source. Here are examples of each: Blockages: *He's trying to get around the regulations*. *We've got him boxed into a corner*. Features of the terrain: *It's been uphill all the way*. *We've been hacking our way through a jungle of regulations*. Burdens: *He's carrying quite a load*. *Get off my back!* Counterforces: *Quit pushing me around*. *She's leading him around by the nose*. Lack of an energy source: *I'm out of gas*. *We're running out of steam*.

In summary, these metaphors are ontological mappings across conceptual domains, from the source domain of motion and forces to the target domain of abstract actions. The mapping is conventional, that is, it is a fixed part of our conceptual system, one of our conventional ways of conceptualizing actions. Conventional metaphors capture generalizations governing polysemy, over inference patterns, and governing novel metaphorical language (Lakoff and Turner, 1989).

2.1 Metaphors vs. Different Word Senses

Presumably, one could treat the metaphoric usage of *run* as a different sense, much in the same way that *move forward on a business plan* is treated as a different sense from literal *move forward*. From a parsing/information extraction point of view, these two approaches are equivalent in terms of their representational requirements.

The benefit of employing the metaphor-based approach, as suggested in the introduction, comes when performing inference. As shown by (Narayanan, 1997), a metaphorical usage and a literal usage share inferential structure. For example, the aspectual structure of *run* is the same in either domain whether it is literal or metaphorical usage. Further, this sharing of inferential structure between the source and target domains simplifies the representational mechanisms used for inference making it easier to build the world models necessary for knowledge-intensive tasks like question answering (Sinha and Narayanan, 2005).

3 Objective

While this work in Cognitive Semantics is suggestive, without a corpus-based analysis, it is hard to accurately estimate the importance of metaphoric information for Natural Language Processing (NLP) tasks such as Question Answering or Information Distillation. Our work is a first step to remedy this situation. We start with our computational definition of metaphor as a mapping from concrete to abstract domains. We then investigate the Wall Street Journal (WSJ) corpus, selecting a subset of its verbal targets and labeling them as either metaphoric or literal. While we had anticipated the pervasiveness of metaphor, we could not anticipate just how pervasive with over 90% of the labeled data being metaphoric.

Provided with labeled training data, our task is to automatically classify the verbal targets of unseen utterances as either metaphoric or literal. Motivated by the intuition that the types of a target's arguments are important for making this determination, we extracted information about the arguments from the PropBank (Kingsbury et al., 2002) annotation for each sentence, using WordNet (Fellbaum, 1998) as the type hierarchy.

3.1 Using Verbal Arguments

A metaphor is a structured mapping between the roles of two frames that makes it possible to describe a (usually) more abstract concept in terms of a more concrete one (Lakoff and Johnson, 1980). The more abstract concept is referred to as the *target* domain while the more concrete concept is referred to as the

1. MET : Texas Air has {run} into difficulty...
2. LIT : "I was doing the laundry and nearly broke my neck {running} upstairs to see ...

Figure 1: Examples taken from the WSJ Corpus. MET indicates a metaphoric use of the target verb and LIT indicates a literal use.

source domain. More precisely, the metaphor maps roles of the target frame onto the source frame.

Figure 1 shows some example sentences with a particular verbal target *run* in curly braces. Example 1 is a *metaphoric* usage (marked by MET) of *run* where the *destination* role is filled by the state of *difficulty*. Example 2 is a *literal* usage (marked by LIT) of *run*.

The arguments of a verb are an important factor for determining whether that verb is being used metaphorically. If they come from the source domain frame, then the likelihood is high that the verb is being used literally. In the example literal sentence from Figure 1, the theme is a person, which is a physical object and thus part of the source domain.

If, on the other hand, the arguments come from the target domain, then it is likely that the verb is being used metaphorically. Consider the metaphorical *run* from Figure 1. In that case, both the theme and the goal of the action are from the target domain. Thus any approach that tries to classify sentences as literal or metaphoric must somehow incorporate information about verbal arguments.

4 Data

Because no available corpus is labeled for the metaphoric/literal distinction, we labeled a subset of the WSJ corpus for our experiments. To focus the task, we concentrated on motion-related frames that act as the source domain for the Event Structure Metaphor and some additional non-motion based frames including *Cure* and *Placing*. Figure 2 shows the selected frames along with example lexical units from each frame.

To identify relevant sentences we first obtained from FrameNet a list of lexical units that evoke the selected source frames. Since WSJ is labeled with PropBank word senses, we then had to determine which PropBank senses correspond to these

Frame	Example LUs
Motion	*float, glide, go, soar*
Motion-directional	*drop, fall, plummet*
Self-motion	*amble, crawl, hobble*
Cause-motion	*catapult, haul, throw, yank*
Cotheme	*accompany, escort, pursue*
Placing	*cram, heap, pocket, tuck*
Cure	*cure, ease, heal, treat*

Figure 2: The frames selected for annotation and some of the lexical units that evoke them.

Cure Frame LU	PropBank Sense
alleviate	alleviate.01
cure	cure.01
ease	ease.02
heal	heal.01
rehabilitate	rehabilitate.01
resuscitate	resuscitate.01
treat	treat.03

Figure 3: The lexical units that evoke the *Cure* frame and each unit's associated PropBank sense[2].

FrameNet lexical items. The lexical items that evoke the *Cure* frame and the corresponding PropBank senses are shown in Figure 3.

As anyone who has inspected both PropBank and FrameNet can attest, these two important lexical resources have chosen different ways to describe verbal senses and thus in many cases, determining which PropBank sense corresponds to a particular FrameNet sense is not a straightforward process. Verbs like *slide* have a single PropBank sense used to describe both the *slid* in *The book slid off the table* and the *slid* in *I slid the book off the table*. While FrameNet puts *slide* both in the *Motion* frame and in the *Cause-motion* frame, PropBank uses the argument labeling to distinguish these two senses.

Periodically, PropBank has two senses, one for the literal interpretation and one for the metaphoric interpretation, where FrameNet uses a single sense. Consider the word *hobble* and its two senses in PropBank:

- hobble.01 "walk as if feet tied together"
- hobble.02 "tie the feet of, metaphorically 'hinder'"

Frame	#MET	#LIT	Total	%MET
Cause-motion	461	44	505	91
Cotheme	926	8	934	99
Motion-directional	1087	21	1108	98
Placing	888	110	998	89
Self-motion	424	86	510	83
Cure	105	26	131	80
All_Frames	3891	295	4186	93

Figure 4: The number of targets annotated metaphoric or literal, broken down by frame.

Lexical Unit	#MET	#LIT
alleviate	8	0
cure	7	3
ease	81	0
heal	3	0
rehabilitate	1	0
resuscitate	2	0
treat	3	23

Figure 5: The lexical units that evoke the *Cure* frame and each unit's counts for metaphoric (#MET) and literal (#LIT) usage.

Because we intended to classify both literal and metaphoric language, both PropBank senses of *hobble* were included. However most verbs do not have distinct literal and metaphoric senses in PropBank.

The final step in obtaining the relevant portion of the WSJ corpus is to use the lists of PropBank senses that corresponding to the FrameNet frames and extract sentences with these targets. Because the PropBank annotations label which PropBank sense is being annotated, this process is straightforward.

Having obtained the WSJ sentences with items that evoke the selected source frames, we labeled the data using a three-way split:

- MET: indicating metaphoric use of the target
- LIT: indicating literal use of the target
- ? : indicating a target that the annotator was unsure of

For our experiments, we concentrated only on those cases where the label was MET or LIT and ignored the unclear cases.

As is shown in Figure 4, the WSJ data is heavily weighted towards metaphor over all the frames that we annotated. This tremendous bias towards metaphoric usage of motion/cause-motion lexical items shows just how prevalent the Event Structure Metaphor is, especially in the domain of economics where it is used to describe market fluctuations and policy decisions.

Figure 5 shows the breakdown for each lexical item in the *Cure* frame. Note that most of the frequently occurring verbs are strongly biased towards either a literal or metaphoric usage. *Ease*, for example, in all 81 of its uses describes the easing of an economic condition and not the easing of pain. *Treat* on the other hand, is overwhelmingly biased towards the treating of physical and psychological disorders and is only rarely used for an abstract disorder.

5 The Approach

As has been discussed in this paper, there are at least two factors that are useful in determining whether the verbal target of an utterance is being used metaphorically:

1. The bias of the verb
2. The arguments of the verbal target in that utterance

To determine whether the arguments suggest a metaphoric or a literal interpretation, the system needs access to information about which constituents of the utterance correspond to the arguments of the verbal target. The PropBank annotations fill this role in our system. For each utterance that is used for training or needs to be classified, the gold standard PropBank annotation is used to determine the verbal target's arguments.

For every verbal target in question, we used the following method to extract the types of its arguments:

1. Used PropBank to extract the target's arguments.
2. For each argument, we extracted its head using rules closely based on (Collins, 1999).

Feature Schema	Example Instantiation	Comment
verb	*verb=treat*	The verbal target
ARG0_TYPE	uninstantiated	ARG0 (Doctor role) not present
ARG1_TYPE	uninstantiated	ARG1 (Patient role) not present
ARG2_TYPE	*ARG2_TYPE=anemia*	The WordNet type is *anemia*.
ARG3_TYPE	*ARG3_TYPE=drug*	The WordNet type is *drug*.

Figure 6: The feature schemas used for classification. The instantiated features are drawn from the sentence *The drug is being used primarily to* {*treat*} *anemias.*

3. If the head is a pronoun, use the pronoun type (without coreference resolution) as the type of the argument.

4. If the head is a named entity, use the Identifinder tag as the type of the argument (BBN Identifinder, 2004).

5. If neither, use the name of the head's WordNet synset as the type of the argument.

Consider the sentence *The drug is being used primarily to* {*treat*} *anemias.* The PropBank annotation of this sentence marks *the drug* as ARG3 and *anemias* as ARG2. We turned this information into features for the classifier as shown in Figure 6.

The *verb* feature is intended to capture the bias of the verb. The *ARGX_TYPE* feature captures the type of the arguments directly. To measure the tradeoffs between various combinations of features, we randomly partitioned the data set into a training set (65% of the data), a validation set (15% of the data), and a test set (20% of the data).

6 Results

6.1 Classifier Choice

Because of its ease of use and Java compatibility, we used an updated version of the Stanford conditional log linear (aka maxent) classifier written by Dan Klein (Stanford Classifier, 2003). Maxent classifiers are designed to maximize the conditional log likelihood of the training data where the conditional likelihood of a particular class c on training example i is computed as:

$$\frac{1}{Z} exp(f_i \cdot \omega_c)$$

Here Z is a normalizing factor, f_i is the vector of features associated with example i and ω_c is the vector of weights associated with class c. Additionally, the Stanford classifier uses by default a Gaussian prior of 1 on the features, thus smoothing the feature weights and helping prevent overfitting.

6.2 Baselines

We use two different baselines to assess performance. They correspond to selecting the majority class of the training set overall or the majority class of verb specifically. The strong bias toward metaphor is reflected in the overall baseline of 93.80% for the validation set. The verb baseline is higher, 95.50% for the validation set, due to the presence of words such as *treat* which are predominantly literal.

6.3 Validation Set Results

Figure 7 shows the performance of the classifier on the feature sets described in the previous section. The overall and verb baselines are 605 and 616 out of 645 total examples in the validation set.

The first feature set we experimented with was just the verb. We then added each argument in turn; trying ARG0 (Feature Set 2), ARG1 (Feature Set 3), ARG2 (Feature Set 4) and ARG3 (Feature Set 5). Adding ARG1 gave the best performance gain.

ARG1 corresponds to the semantic role of *mover* in most of PropBank annotations for motion-related verbs. For example, *stocks* is labeled as ARG1 in both *Stocks fell 10 points* and *Stocks were being thrown out of windows*[3]. Intuitively, the mover role is highly informative in determining whether a motion verb is being used metaphorically, thus it makes sense that adding ARG1 added the single biggest

[3]This is an actual sentence from the training set.

FSet	Feature Schemas	M	L	Total	%Tot
1	*verb*	599/605	20/40	619/645	95.97
2	*verb, ARG0_TYPE*	601/605	17/40	618/645	95.81
3	*verb, ARG1_TYPE*	602/605	19/40	621/645	96.28
4	*verb, ARG2_TYPE*	600/605	19/40	619/645	95.97
5	*verb, ARG3_TYPE*	599/605	20/40	619/645	95.97
6	*verb, ARG1_TYPE, ARG3_TYPE*	602/605	19/40	621/645	96.28
7	*verb, ARG1_TYPE, ARG2_TYPE, ARG3_TYPE*	601/605	18/40	619/645	95.97
8	*verb, ARG0_TYPE, ARG1_TYPE, ARG2_TYPE*	602/605	18/40	620/645	96.12
9	*verb, ARG0_TYPE, ARG1_TYPE, ARG2_TYPE, ARG3_TYPE*	602/605	17/40	619/645	95.97

Figure 7: For each Feature Set, the feature schemas that define it, along with the ratio of correct to total examples on the validation set for metaphor (M), literal (L) and total (Total) is shown.

jump in performance compared to the other arguments.

Once we determined that ARG1 was the best argument to add, we also experimented with combining ARG1 with the other arguments. Validation results are shown for these other feature combinations (Feature Sets 6,7, 8 and 9)

Using the best feature sets (Feature Sets 3,6), 621 targets are correctly labeled by the classifier. The accuracy is 96.98%, reducing error on the validation set by 40% and 17% over the baselines.

6.4 Test Set Results

We retrained the classifier using Feature Set 3 over the training and validation sets, then tested it on the test set. The overall and verb baselines are 800 and 817 out of 861 total examples, respectively. The classifier correctly labeled 819 targets in the test set. The results, broken down by frame, are shown in Figure 8. The final accuracy of 95.12%, represents a reduction of error by 31% and 5% over the baselines.

6.5 Discussion

A comprehensive assessment of the classifier's performance requires a measure of interannotator agreement. Interannotator agreement represents a ceiling on the performance that can be expected on the classification task. Due to the very high baseline, even rare disagreements by human annotators affects the interpretation of the classifier's performance. Unfortunately, we did not have the resources available to redundantly annotate the corpus.

We examined the 42 remaining errors and categorized them into four types:

- 13 fixable errors
- 27 errors caused by verbal biases
- 2 errors caused by bias in the training set

The fixable errors are those that could be fixed given more experimentation with the feature sets and more data. Many of these errors are probably caused by the verbal bias, but a verbal bias that should not be insurmountable (for example, 2 or 3 metaphor to each 1 literal).

The 27 errors caused by verbal biases are ones where the verb is so strongly biased to a particular metaphoric class that it is unsurprising that a test example of the opposite class was missed. Verbs like *treat* (0 metaphoric to 20 literal) and *lead* (345 metaphoric to 0 literal) are in this category.

The two remaining errors are cases where the verb was not present in the training data.

7 Related Work

Previous work on automated metaphor detection includes Fass (1991), Martin (1990), and Mason (2004). Whereas our aim is to classify unseen sentences as literal or metaphorical, these projects address the related but distinct task of identifying metaphorical mappings. All three use the selectional preferences of verbs to identify metaphors. In literal usage, the arguments that fill particular roles of a verb are frequently of a common type. For instance, in the MEDICAL domain, the object of the

Frame	M	L	Total	%Tot	%OBL	%VBL
Cause_motion	78/78	1/10	79/88	**89.77**	88.64	88.64
Cotheme	179/179	0/2	179/181	98.90	98.90	98.90
Cure	26/30	3/3	29/33	87.88	**90.91**	90.91
Motion_directional	242/242	0/2	242/244	99.18	99.18	99.18
Placing	176/181	13/25	189/206	**91.75**	87.86	91.26
Self_motion	87/90	14/19	101/109	**92.66**	82.57	91.74
All_Frames	788/800	31/61	819/861	**95.12**	92.92	94.89

Figure 8: The results of the classifier on the test set, using Feature Set 6. For each frame, the ratio of correct to total examples for metaphor (M), literal (L) and total (Total) is shown. The total percent correct for the frame (%Tot), the overall baseline percentage (%OBL), and the verb baseline percentage (%VBL) are also shown. The cumulative performance over all frames is located in the bottom row of the table.

verb *treat* is usually a pathological state. In the FINANCE domain, the object of *treat* is usually an economic problem. This difference in selectional preference suggests metaphorical usage. Furthermore, it suggests a metaphorical mapping between health problems and economic problems.

The systems described by Fass and Martin exhibit impressive reasoning capabilities such as identifying novel metaphors, distinguishing metaphor from metonymy, and interpreting some metaphorical sentences. But they require hand-coded knowledge bases and thus have limited coverage and are difficult to extend. More similar to our efforts, Mason's CorMet uses a corpus-based approach. In CorMet, domains are characterized by certain keywords which are used to compile domain-specific corpora from the internet. Based on differences in selectional preferences between domains, CorMet seeks to identify metaphorical mappings between concepts in those domains.

One shortcoming of using syntactic arguments is reflected by CorMet's mistaken identification of a mapping between *institutions* and *liquids*. This arises from sentences like *The company dissolved* and *The acid dissolved the compound*. Such sentences suggest a mapping between the subjects in the target domain, *institutions*, and the subjects in source domain, *liquids*. Using semantic roles avoids this source of noise. This is not to suggest that the syntactic features are unimportant, indeed the selectional preferences determined by CorMet could be used to select which arguments to use for features in our classifier.

Our approach considers each sentence in isolation. However the distribution of metaphorical usage is not uniform in the WSJ corpus (Martin, 1994),. It is therefore possible that the information about surrounding sentences would be useful in determining whether a usage is metaphorical. CorMet incorporates context in a limited way, computing a confidence rating, based in part upon whether a metaphoric mapping co-occurs with others in a systematic way.

8 Conclusion

Metaphors are a ubiquitous phenomenon in language, and our corpus analysis clearly bears this out. It is somewhat gratifying that with a judicious combination of the available wide-coverage resources (WordNet, FrameNet, PropBank) we were able to build classifiers that could outperform the baseline even in the most skewed cases. Our results show the utility of our approach and more generally the maturity of the current NLP technology to make progress in attacking the challenging and important problem of interpreting figurative language.

However, this is only the first step. As with all semantic extraction methods and technologies, the proof of utility is not in how good the extractor is but how much it helps in an actual task. As far as we can tell, this problem remains open for the entire semantic parsing/role labeling/extraction field despite the flurry of activity in the last four years. In the case of metaphor interpretation, we have some initial encouragement from the results published by (Narayanan, 1997) and others.

Our classifier relies on PropBank senses, so we can use the high performance classifiers available for PropBank. The price is that we have to construct mappings from FrameNet frames to PropBank senses. However, this is a one-time effort pursued by many groups, so this should not present a problem to extending our approach to cover all frames and metaphors. Additionally, we are in the process of linking the metaphor detector to a metaphor inference system. We hope to have initial results to report on by conference time.

References

BBN Identifinder. 2004. http://www.bbn.com/for_government_customers/data_indexing_and_mining/identifinder.html.

Michael Collins. 1999. *Head-Driven Statistical Models of Natural Language Parsing*. Ph.D. thesis, University of Pennsylvania.

Dan Fass. 1991. Met*: a method for discriminating metonymy and metaphor by computer. *Comput. Linguist.*, 17(1):49–90.

Christine Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

Mark Johnson. 1987. *The Body in the Mind: The Bodily Basis of Meaning, Imagination and Reason*. University of Chicago Press.

Paul Kingsbury, Martha Palmer, and Mitchell Marcus. 2002. Adding semantic annotation to the penn treebank. In *Proceedings of the Human Language Technology Conference*.

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

George Lakoff and Mark Turner. 1989. *More Than Cool Reason: A Field Guide to Poetic Metaphor*. University of Chicago Press.

George Lakoff. 1994. The contemporary theory of metaphor. In Andrew Ortony, editor, *Metaphor and Thought*. Cambridge University Press.

Ronald Langacker. 1987. *Foundations of Cognitive Grammar I: Theoretical Prerequisites*. Stanford University Press.

James Martin. 1990. *Computational Model of Metaphor Interpretation*. Academic Press.

J.H. Martin. 1994. A corpus-based analysis of context effects on metaphor comprehension. Technical report, Boulder: University of Colorado: Computer Science Department.

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

Srini Narayanan and Sanda Harabagiu. 2004. Question answering based on semantic structures. In *Proceedings of the International Conference on Computational Linguistics*.

Srini Narayanan. 1997. *Knowledge-Based Action Representations for Metaphor and Aspect*. Ph.D. thesis, University of California at Berkeley.

Steve Sinha and Srini Narayanan. 2005. Model-based answer selection. In *Proceedings of the AAAI Workshop on Inference for Textual Question Answering*.

Stanford Classifier. 2003. http://nlp.stanford.edu/software/classifier.shtml.