Toward Improving the Automated Classification of Metonymy in Text Corpora*

Francis M. O. Ferraro
Department of Computer Science
University of Rochester

May 5, 2011

Abstract

In this paper, we explore methods for improving the automatic classification of schematic metonymies in corpus-based text. Using a pre-existing dataset of thousands of samples, we formulate the hypothesis that a better modeling of the underlying syntactic, semantic and conceptual meanings within a document (sample) can aid automated metonymy classification. To test this hypothesis, we build upon previous researchers' work on metonymy resolution systems but also introduce novel features, including, but not limited to, the extraction and analysis of conceptual paths between syntactically and semantically connected words. We initially explore three models for this classification task, but settle for final evaluation on two of them. Purely quantitatively, the results indicate that more work needs to be done, but on a more detailed analysis, we explain the strengths of our ideas behind our methods, and the weaknesses of the databases used for feature extraction. Finally, we present ways that this work can be continued and expanded.

*This work was completed in partial fulfillment of the requirements for an Honors Bachelor of Science Degree in Computer Science from the Department of Computer Science at the University of Rochester, in Rochester, NY, USA.

Contents

List of Tables

List of Figures

1 Introduction

Various problems still plague automatic processing of text and automatic knowledge extraction from text. Many commonly studied problems include anaphora and pronoun co-reference, and word sense disambiguation (WSD), though those two are by no means the only problems yet to be tackled and solved: in general, figurative language poses a problem as it requires abstract thought and requires one to represent one object or idea in terms of another object or idea. Common examples of figurative language include metaphor and metonymy. We restrict this work to a consideration of metonymy, though it has been argued that a truly representative model must incorporate explanations of multiple types of figurative language (Fass 1997); as a result, we try to use methods that allow themselves to be generalized. In doing so though, we examine common types of metonymy, which will be defined shortly. This decision was influenced by the number of NLP tasks that can benefit from metonymy resolution. For instance, metonymy has been an issue for machine translation for quite some time (Kamei and Wakao 1992, Onyshkevych 1998). Question and answering systems could also benefit from better metonymy resolution systems (Stallad 1993), as could anaphora and pronoun co-reference (Markert and Hahn 2002). Thus improving metonymy recognition and classification could have a large benefit.

1.1 Defining Metonymy

Though in general it is notoriously difficult to arrive at an agreed upon definition of figurative language, it is widely accepted that **metonymy** is a figure of speech in which one references an object by a closely related name rather than the actual name Lakoff and Johnson (1980). A word that indicates a metonymy is called a **metonym**. The following sentences show examples of metonymy, where the metonyms are italicized (throughout this work, potentially metonymic phrases are italicized):

(1) *London* and *Paris* discussed important policy matters at the convention.

(2) As a classicist, Bill enjoys *Shakespeare*.

(3) The *pen* is mightier than the *sword*.

The above sentences indicate how varied metonymy can be, and as a result, how recognizing metonymy, even for humans, can vary in difficulty. For instance, in sentence 1, both *London* and *Paris* are metonyms for (representatives of) the British and French governments. In 2, *Shakespeare* refers to the works of Shakespeare (the phrase "As a classicist" rules out the interpretation that Bill actually knew Shakespeare). The old adage given in 3 metonymically compares the *pen* – literary and diplomatic force – and the *sword* – military force.

Given the wide scope of metonymy, it is not surprising that metonymy has received a thorough discussion in linguistic literature. Over the years, this has involved distinguishing metonymy from other tropes, such as synecdoche and metaphor (Bisang et al. 2006, Fass 1988, Lakoff and Johnson 1980, Seto 1999), and developing typologies and hierarchies for metonymy classification (Fass 1997, Hilpert 2006, Stern 1965). Indeed, one common and theoretically useful result has been to develop metonymy categories, which attempt to get at the underlying cognitive patterns used for metonymy. Such categories include THE-PART-FOR-THE-WHOLE, such as referring to the "nice *wheels*" (the car) of a friend; PLACE-FOR-PRODUCT, such as referring to "the *Bordeaux*" (wine) that stands up well to the meal; and OBJECT-FOR-REPRESENTATION, such as pointing to a region on a map and saying "This is *Russia*." Providing a comprehensive list of these categories would be prohibitive and, as will be discussed, distracting; for an initial listing, the reader is directed to Lakoff and Johnson (1980).

Researchers have tried to develop high-level "trigger" clues, reliable heuristics and rules to map metonymic phrases and sentences into the above groupings and patterns (Bisang et al. 2006, Deignan 1999; 2006). For humans, the conceptual utility of these groupings and patterns is high; unfortunately, computers have not been able to leverage them to the greatest extent, in part due to the high-level and abstract nature of the heuristics used to perform groupings. Thus although select heuristics, such as analyzing the differences in connotations of "flame" and "flames" in different contexts (Deignan 2006), can provide insight into the detection and resolution of figurative language for humans, these individualized heuristics require expert insight and long study. Although large-corpora collocation analysis has, to an approximation, captured some distributional data helpful to this connotation-heuristic creation (Nastase and Strube 2009), scaling these

heuristics to a wider range of figurative language, including metonymy, does not seem complete. If computers were better able to conceptualize ideas and realize connotations between ideas, it may be possible to better learn these latent connotation clues.

Approaching metonymy identification and classification with a more computationally-conscious view, Pustejovsky (1991) described a *qualia* structure of nouns through different roles. Briefly, these roles are constitutive (describing how an object relates to its constituents); formal (how a given noun distinguishes itself within some larger domain); telic (describing a noun's purpose and function); and agentive (an explanation of a noun's existence). Although these qualia and roles are oft-cited in research, not much recent work has attempted to model or incorporate them (an exception is Shutova (2009)); this something we address in this work. These roles may have significant utility, but as with the previously-mentioned high-level trigger clues, efficiently and correctly extrating the qualia structure can present computational issues (e.g., scalability).

Although this scalability problem is not new, Markert and Nissim (Markert and Nissim (2005; 2002)) were some of the first to address it and propose a solution, by defining a computational distinction between *unconventional* and *schematic* metonymy . As briefly discussed in Markert and Nissim (2005), the main difference between unconventional and schematic metonymy is in the ability for a schematic metonymy to be "productive" in different contexts and with different metonyms (assuming two metonyms are of the same category — locations, organizations, etc.). In the following examples, sentence 4 (from Lakoff and Johnson (1980), Markert and Nissim (2005)) demonstrates an unconventional metonymy, whereas sentence 5 (from Markert and Nissim (2005)) demonstrates a schematic metonymy:

(4) The *ham sandwich* is waiting for his check.

(5) *Paris* sleeps.

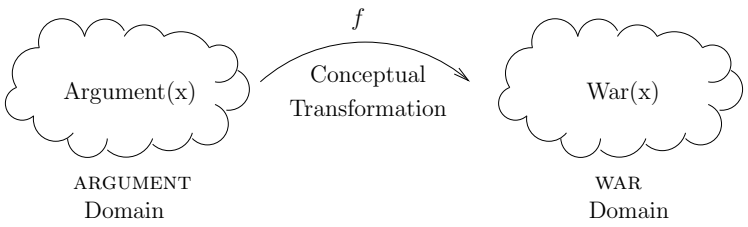
A general hallmark of schematic metonymy is that it is less creative than unconventional metonymy so unconventional metonymy tends to be more tailored to specific situations. This generalization is reflected in the above two examples, as 4 is more creative and can only be applied in select situations. This contrasts with 5, which can be generalized to discuss any geopolitical location. Another potential discriminating characteristic of schematic metonymy is in its ability to easily be placed into standard cognitive categories; in the above schematic metonymy example, there is a clear X-FOR-Y grouping choice that can be made (PLACE-FOR-PEOPLE). This paper will focus on schematic metonymy.

1.2 Metonymy, Metaphor and Figurative Language

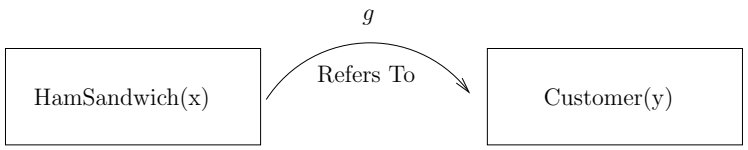
Particularly after having encountered metonymy formally for the first time, it is a natural next step to question how metonymy compares to other types of figurative language. While we cannot hope to do such a discussion justice in this work, we provide a brief overview and direct the reader to Fass (1997), Lakoff and Johnson (1980) for more complete discussions. Although in this section we examine how metonymy differs from **metaphor**, a very common type of figurative language, we stress that the differences between metonymy and metaphor we discuss here are in no way definitive or without some controversy. As they are both types of figurative language, they continue to inspire.

As described in Lakoff and Johnson (1980), metaphor is the process of *understanding* and *conceptualizing* one object in terms of another. At its core, it is a conceptually transformative function between domains. For instance, there is the canonical ARGUMENT-IS-WAR metaphor: claims may be *indefensible*, or a point could be *on target* Lakoff and Johnson (1980). Metaphor maps ideas and ways of thinking about a topic in one domain onto a different domain. Thus if we adopt a mathematical notation, metaphor can be described as a function f between sets (domains). Please see Figure 1a for a visualization of this. As metaphor describes an overall conceptual mapping, rarely are individual words targeted as a metaphor; rather, entire phrases or sentences are described as being metaphoric.

Metonymy, meanwhile, is at its core a reference task. If one object is used to refer to another, then while there is a connection (function g) between the two referents, it is not required that the function be a conceptually transformative one. For instance, in example 4 does the speaker actually think of, or conceptualize, the customer as a ham sandwich? All that can be discerned is that the customer is being refered to by "the ham sandwich." Figure 1b provides a visualization. Unlike metaphor, metonymy is word- or phrase-based; that is, specific words can easily be classified metonymically, but it does not make sense for a sentence to be classified metonymically.



(a) Illustration of a metaphorical transformation f; f is a conceptually transformative mapping, bringing one domain to another.



(b) The metonymic transformation is realized by g, a referential mapping.

Figure 1: A visualization of the differences between metaphor (a) and metonymy (b). Metaphor can be thought of as a conceptually transformative mapping, bringing one domain (ARGUMENT) into another (WAR). Metonymy can be thought of as a referential function.

While the above discussion and definitions show that metaphor and metonymy are distinct types of figurative language, it is important to know that they can co-occur in the same sentence. An easy example is

(6) During negotiations, $\{America\}_1$ $\{deflected\ the\ criticism\}_2$ of its treatment of prisoners,

in which the first group ("America") is used metonymically, while the second ("deflected the criticism") indicates metaphor.

Given that we have these two types of figurative language, it is also natural to wonder how frequently each of them is used, and how frequently they co-occur. Unfortunately, as Fass (1997) describes, there are not many empirical distributional studies of metonymy. Of those studies that do include metonymy, it is rarely separated from other figurative language, such as metaphor (Martin (1994)). Therefore, one can expect the variance in distributional studies and results to be high. Despite this, it is still worth considering the fact that Martin (1994) found that for every one hundred words a combination of metaphor and metonymy occurs five times. However, while concrete frequencies — to support the claims of how frequent metonymy is — would be ideal, we believe that we can currently progress without them and still be correct in giving metonymic usage a high occurrence count. For consider how natural metonymy is, and how easily it can be used.

1.3 Computationally Viewing Metonymy

While some initial motivation for automating metonymy resolution was given at the beginning of Section 1, there are still some issues that must be addressed before proceeding.

Although metonymy resolution can be treated as a special case of word sense disambiguation Markert and Nissim (2002), one problem with figurative language is that after a while, the terms and phrases can become so common that they are lexicalized; at this point, the conceptual reasoning behind the metonymy is, to an extent, irrelevant. Consider a slight variant of an oft-cited metonymy: "the hired *gun* [killed Bill]" Lakoff and Johnson (1980). In the strictest sense, it is not the gun itself that is "hired" (guns are bought, not hired) but the person who used the gun was hired. Despite this lack of literal meaning, this sense of "gun" has become lexicalized: in WordNet 3.0 there is the synset "a professional killer who uses a gun." This indicates that a specific type of person has become associated with a particular object *and its function* (or a result of using it). We note that schematic metonymy should theoretically be relatively unaffected by

lexicalization. Even though the metonyms fit a pattern, the particular usage and referents generally need to be determined from context.

This problem of lexicalization is further reason for focusing on schematic metonymy. Metonymies involving lexicalization reduce the resolution task to one of selecting the best sense out of some set of predefined senses, where we know that the "correct" answer is actually listed. By focusing on schematic metonymy and non-lexicalized figurative language as much as possible, we are instead trying to select the best sense that helps provide the strongest evidence for a particular conceptual mapping.[1] Even though we have a predefined set of description/classification categories (X-FOR-Y), the senses are not necessarily fully conceptualized.

1.3.1 Viewing Metonymy Resolution as Entailment Production

The above discussion of lexicalization brought up the issue of treating metonymy resolution as WSD. However, we may also view the process of automating metonymy detection and classification as being akin to automating entailment production. Considering sentence 1, a knowledge base may contain only one constant term for *London* even though *London* can take on a number of meanings. Recognizing and classifying the demonstrated metonymy can be thought of as creating rules and appropriate premises such as "*London* may sometimes refer to the British government" or "*London* may sometimes refer to the geopolitical entity *London*."[2] Although the classifiers and rules researchers construct may not be in the form of logic formulae, we can still view these resolution systems as a "wrapper" around formulae: we may think of it as trying to create sets of rules that govern correct interpretation. A grand goal of this work is that a metonymy resolution system can serve as a stand-alone metonymy detector, or it can be streamlined into a larger NLP task as an initial processing phase. However, NLP tasks at various stages, such as machine translation as an end-stage and pronoun co-reference as an intermediate-stage, can still benefit from resolving this ambiguity.

1.4 Outline of this Work

As with any experiment, we must decide what data and evaluation metrics to use, which are described in Section 2. In doing so, we see examples of the specific types of metonymies we will be analyzing, which is dependent on the particular dataset. After that, in Section 3 we explore related and previous work. While some of the most useful information is from previous researchers using the dataset we use, it is also useful to examine work on other types of metonymy (e.g., logical metonymy) or on other datasets. The information presented in Sections 2 and 3 allows us to form our main hypothesis: that a better modeling of the underlying syntactic, semantic and conceptual meanings within a document can aid automated metonymy classification. Thus, in Section 4 we describe in more detail our hypotheses, how our methods continue the narrative of metonymy resolution and how we will test the hypotheses. In order to apply these models, we must extract various features; this definition, extraction and engineering process is described in Section 6. We present results and evaluate our methods in Section 7. We also analyze these methods in both large (7.1) and small (7.2) scale evaluations.

2 Datasets and Evaluation

In many tasks where results may be easily quantified, there are two common and well-known issues that generally arise. Both of the issues are due to the fact that it is incredibly difficult to construct a dataset that accurately represents the "real" distribution of whatever is being modeled. Although the reader may be familiar with these issues, we will briefly introduce them; this will allow us to refer to these issues as we progress through this work.

The first issue is what we term *scale invariance* of a method, while the second case is what we term *dataset invariance* of a method. Scale invariance is rather broad, but as one may be able to surmise, it describes how the efficacy of a method is affected by changing a non-trivial characteristic of the dataset, be

[1] As will be discussed in Section 2.1, there is a degree of subjectivity in assigning the labels LITERAL, X-FOR-Y, OTHER-MET and MIXED.

[2] The governing premises are not shown in this prose.

it the number of entries, the number of labels, the complexity of the entries, etc. Dataset invariance, on the other hand, measures how using different datasets to evaluate a method affects the results of that method.

Markert and Nissim note that the lack of a standardized dataset plagues the advancement of metonymy resolution Markert and Nissim (2007). As a result, researchers either tend to hand-construct a very small number of example sentences or use different datasets. In either case, some combination of scale and dataset invariance is introduced, and results from some researchers are not necessarily comparable to those of other researchers. This lack of comparability has made it difficult to "objectively" measure how the state-of-the-art develops. However, it is important to consider comparability of methods, particularly as large-scale text corpora become more important in natural language tasks. Methods can be tailored to very specific types and examples of metonymy and as a result they may be able to perform well on a small number of sentences and a somewhat limited domain, but they may not be widely applicable and scalable.

Thus, as in many other NLP tasks, we are faced with the problem of developing systems that have either broad coverage, which may not be as accurate or precise as desired, or developing systems that provide very accurate analyses for a much smaller domain. Attempting to bridge those views, Markert and Nissim released data from a metonymy resolution competition from the 2007 SemEval conference Markert and Nissim (2007). These data are from a subset of the BNC Version 1.0 and focus on schematic metonymies involving locations or organizations. As many NLP systems are trained on newswire data (such as from the BNC) it is reasonable to assume that many of these systems encounter location- and organization-based metonymy. Thus part of the goal can be seen as increasing the correctness (approaching human standards) on a widely used subset of common data.

The SemEval 2007 dataset is not the only one in use. Other datasets provide excellent opportunities for determining the dataset invariance of a method or algorithm; however, studying all of them was beyond the scope of this work and so some of these datasets are discussed as opportunities for future work (Section 8). Therefore, for reasons just discussed, as well as in Section 1, we have chosen to use the data from the SemEval 2007 competition. The data are described in the following section.

2.1 SemEval 2007 Metonymy Competition

The organizers of the SemEval 2007 task on metonymy recognition and classification (Markert and Nissim 2007) were experienced with metonymy recognition systems (Markert and Nissim 2002, Nissim and Markert 2003; 2005). Further, as they were familiar with computational metonymy resolution they knew many of the problems troubling this automated task and wanted to address some of these issues by making the task more standardized and more applicable to corpus-based natural language systems. In the terminology of this work, they mention that most of the promising work in figurative language does not adequately address either scale invariance or dataset invariance and so results are difficult to compare. The five different systems (Brun et al. 2007, Farkas et al. 2007, Leveling 2007, Nicolae et al. 2007, Poibeau 2007) to participate in this contest are reviewed in greater depth in Section 3.2.1. Other researchers have used subsets of this dataset as well (Nastase and Strube 2009, Peirsman 2006); their work is also reviewed.

The task only considers schematic metonymy (see Markert and Nissim (2005), and/or page 7 of this work), in either the LOCATION (countries) or ORGANIZATION (companies) domains (Sections 2.1.1 and 2.1.2, respectively). To gather the data, countries or companies were identified in the BNC, Version 1.0; this country or company is what we call a *potentially metonymic phrase* (**PMP**). To provide a sufficient context for the PMP, up to three sentences surrounding the one containing the PMP were extracted: two sentences preceding, and one sentence following. Using this context and a self-made annotation guide (Markert and Nissim 2005), Markert and Nissim gave each PMP a predefined label: either literal or metonymic, and if metonymic, what category of metonymy? These categories are inspired by those proposed in Lakoff and Johnson (1980). While many of the metonymy-type labels are domain dependent, there are two domain-independent metonymy categories (Section 2.1.3). The κ score across all annotations for inter-annotator agreement is reported as being near 0.90 Markert and Nissim (2007).

Normally the high κ score may assuage fears about the annotations, but given the highly subjective nature of metonymy, there is still a potential issue. This concern has not escaped other researchers and reviewers: due to the fact that metonymy (and metaphor) are very prevalent in our language and culture (Fass 1997, Lakoff and Johnson 1980), there has been expressed concern that some metonymic readings were missed, and coercions took place (Poibeau 2007). Table 1 presents the distribution of labels throughout the

Label	Training	Testing
literal	737 (79.68%)	721 (79.41%)
place-for-people	161 (17.41%)	141 (15.53%)
place-for-event	3 (0.32%)	10 (1.10%)
place-for-product	0 (0%)	1 (0.11%)
othermet	9 (0.97%)	11 (1.21%)
mixed	15 (1.62%)	20 (2.20%)
obj-for-name	0 (0%)	4 (0.44%)
obj-for-representation	0 (0%)	0 (0%)
Total	925	908

(a) The training/testing distribution for LOCATION-based PMPs.

Label	Training	Testing
literal	690 (63.30%)	520 (61.76%)
org-for-members	220 (20.18%)	161 (19.12%)
org-for-event	2 (0.18%)	1 (0.12%)
org-for-product	74 (6.79%)	67 (7.96%)
org-for-facility	15 (1.38%)	16 (1.90%)
org-for-index	7 (0.64%)	3 (0.36%)
othermet	14 (1.28%)	8 (0.95%)
mixed	59 (5.41%)	60 (7.12%)
obj-for-name	8 (0.73%)	6 (0.71%)
obj-for-representation	1 (0.09%)	0 (0%)
Total	1090	842

(b) The training/testing distribution for ORGANIZATION-based PMPs.

Table 1: The distribution of all labels for both LOCATION- and ORGANIZATION-based samples in the SemEval 2007 dataset.

dataset, which illustrates a potential cause for concern. Examining the data, we see that approximately 80% of the location candidate metonymys were judged to have a literal meaning, while only 63% of the organization candidate metonymys were judged to be literal[3]. Given the distribution concerns raised in Section 1.2, these distributions do warrant review. Other researchers, in part to make the computational handling of metonymy more rigorous and sound, have expressed a desire for further examination and analysis of these data (Bisang et al. 2006, Hilpert 2006, Poibeau 2007). Therefore, given the lack of a formal, well-specified as agreed-upon definition of metonymy (Fass 1997), the data are sufficient and usable for our purposes; any mistakes or misannotated samples will be noted and dealt with on a case-by-case basis.

One of the primary assumptions made while annotating and labeling the data is that there is only one PMP per sample. Although Fass (1997) does not examine metonymy by itself, he notes that figurative language (notably a combination of metaphor and metonymy) tends to appear in groups. According to these findings then, the assumption of a single PMP per sample is not always a good assumption. Indeed, as will be seen shortly, words near a PMP generally assume the label of the PMP — thus locations near a literally-read location-based PMP tend to be literally read as well, while locations near a metonymically-read location-based PMP tend to be metonymically read; see sentence 8 for an example.

What follows in the next three subsections are categories used, along with possible high-level diagnostic tests and heuristics for determining category membership. The diagnostics are a result of our own thinking, but also information made available with the data Markert and Nissim (2005). Appropriate examples are given, though due to space considerations, only the illuminating portions of each example are given.

Recall that throughout the rest of this paper, a candidate metonymy is called a *possibly metonymic*

[3]These numbers were compiled only from the sentences/candidates for which there was agreement between the annotators.

phrase (PMP), and at times refered to symbolically by `X`.

2.1.1 Location Category

For the location category there are the following categories:

- LITERAL These generally can be characterized as geographical descriptions or intra-location relationships. In the former, we have physical, ecological and geographical descriptions (including cardinal modifiers), as in 7:

  (7) ... if they weren't already at the bottom of the North Sea – just off the southern coast of *Norway*. ...

  The latter can generally be described as "`X`'s `Y`," or "`Y` of/in `X`." Sentence 7 provides an example of this as well. However, this diagnostic is also more nuanced, as personnel descriptions, political membership, or as Markert and Nissim (2005) describes, the recipient of state-targeted actions (such as boycotts, sanctions, etc.) may be appropriate clues. Information regarding possession and governing prepositions of the PMP is important to extract for this category.

- PLACE-FOR-PEOPLE

  This metonymy occurs when a place represents either a person(s) or organization(s) associated with it. Common heuristics can include replacing the PMP `X` by some combination of "(people/group/organization) (associated with/inhabiting/frequenting) `X`." In sentence 8, we may use the paraphrase "with hardline stalwarts such as *the people associated with* Cuba" to obtain a (more-or-less) equivalent reading:

  (8) ... East Germany was again glad to oblige, along with such hardline stalwarts such as *Cuba*, North Korea,...

  (9) But it's encouraging that so many scientists, politicians and (of late) influential business people have been making it unmistakably clear to White House officials that the *US* position on global warming and the whole Earth Summit process absolutely stinks.

  (10) ... [The resolution] demands that Iraq comply fully... and decides, while maintaining all its decisions, to allow *Iraq* one final opportunity...

  Notice how 8 shows that locations near PMP-locations can assume the same reading label: both *Germany* and *North Korea* could be given a PLACE-FOR-PEOPLE reading.

  A PLACE-FOR-PEOPLE reading can also include descriptions of physical actions; mental or emotional states; or some attribute of the people inhabiting the country (but not the country itself). Sentences 9 and 10 provide further examples; note in 9 the human actions of "advocating" and taking a "position" are key indicators of this metonymy. Contrast this to 10, where "allow" also anthropomorphizes *Iraq* (as does "comply").

- PLACE-FOR-EVENT

  This label is used when a PMP refers to some event that took place in that location. These events include athletic/sporting events, war and major political events (e.g., Watergate). In 11, *Italy* is used to represent the soccer tournament being held there:

  (11) After their European Championship victory and Milan's orange-tinted European Cup triumph, Holland will be expected to do well in *Italy*.

  Just as for the LITERAL category, governing preposition information is important.

- PLACE-FOR-PRODUCT

  This label is applied when, for a PMP `X`, we can find some product `Y` such that `Y` was produced in `X`. Unfortunately, this is a very rare label (in the entire dataset there is only one instance, which is in the testing portion); thus rather than provide the example from the data, we provide 12, which is of our own creation (Bolesławiec, Poland is famous for its pottery).

(12) I wasn't sure what to get Lydia, so I got her some *Bolesławiec*.

In this example, *Bolesławiec* refers to some pottery from Bolesławiec, Poland. Although similar examples may be constructed, we have found that at the core they tend to be very similar (the locations tend to be used with a determiner).

- OTHERMET

  This can be described as the miscellaneous, catch-all category, to be used when there are no clues to indicate a literal reading (indeed, perhaps there are some clues against such a reading) but the above categories are not appropriate. Unfortunately, this can diverge from the idea of schematic metonymy, and there few good heuristics (other than none of the other heuristics work well for the sentence). An example of this is in 13:

  (13) ... Asked about the role of Japan, he replied that Japan's economy was lagging behind; only now was *Japan* going into recession...

  In this case, it is the entire economy of Japan – the people, companies, organizations, etc. – that are entering a recession. There is no good predefined class for this, though a literal reading, one that solely represents a geopolitical entity, clearly is not appropriate.

2.1.2 Organization Category

In keeping with the participants and annotators, we denote this class as "organization," even though it is only for business and company organizations (proper nouns); it does not include other types of organizations, such as non-profit, economic or political organizations. For the organization category, there are the following categories:

- LITERAL

  A literal reading is to be used when the PMP `X` refers to the organization as a whole, rather than one particular aspect or attribute (such as the people involved, the products it produces, etc.). This is demonstrated in 14:

  (14) The main competition to *IBM* came from a group of other US mainframe producers, often referred to as the 'seven dwarfs' or the Bunch.

  (15) They are nonetheless a big comedown from the 1960s, when federal trustbusters took on giants the size of AT&T and *IBM* and broke up a merger of Procter & Gamble and Clorox.

  Note that this diagnostic is slightly less nuanced than that of the literal reading for locations. That being said, if artifacts or attributes of an organization are mentioned (as in 15, size is considered an attribute), a literal reading tends to be assigned.

- ORG-FOR-MEMBERS

  This is analogous to PLACE-FOR-PEOPLE in the location class: ORG-FOR-MEMBERS contains those metonymies where a business is used to stand for the people either comprising it (e.g., employees, spokespeople, administrative divisions) or associated with it (e.g., shareholders). Typically, these types involve communication (example 16), action sentience and volition (example 16) and emotional experiences (example 17). Semantic θ-roles are useful for this type of metonymy.

  (16) The UK phone company reports that it has also successfully tested wireless access to its messaging service via RAM Mobile Data's national network. It says it expects to offer a wireless electronic messaging service by the end of the year, and that it is aiming for beta testing during the summer. Pricing has yet to be decided, but *BT* is aiming for the look and feel of the service to be the same as for its existing messaging services.

  (17) The price also reflects the ominous presence of BellSouth which has tabled a rival offer for LIN. While *BT* can be pleased that it bought into McCaw...

We do note that in 17, there can be some ambiguity: the initial predicate (*pleased*) indicates a metonymy, while the latter predicate (*bought into*, which connotes the company's assets as a whole) can be seen as indicating a literal reading. This demonstrates the subjectivity of the labeling and is a consequence of using data annotated by others but not subject to rigorous linguistic analysis Hilpert (2006).

- ORG-FOR-EVENT

  This category is somewhat rare, as it requires an organization to be associated with some event in history. This can include economic events (such as the scandals of Enron and WorldCom), political events/outcomes (see example 18) and (potential) disasters (e.g., referencing "Three Mile Island"), among others. However, regarding the data here, most of these diagnostics are hypothetical as there are so few examples of this category in the data.

  (18) In its *Philip Morris* decision in November 1987, the Court held that...

- ORG-FOR-PRODUCT

  This category is analogous to the location-based PLACE-FOR-PRODUCT, although there are many more examples in the data for ORG-FOR-PRODUCT than for PLACE-FOR-PRODUCT. This reading is appropriate if a PMP refers to a product that said PMP produces. See sentence 19 for an example.

  (19) Los Angeles-based IDOC says it has a platform-independent translation management package called XL8 that facilitates simultaneous software releases in multiple languages. The company says the stuff automates moving from Macs to PCs to *Suns* despite the fact they all use different character sets to create their on-screen texts.

  As will be seen in sentence 31 (pg 16), the product of the company cannot be explicitly mentioned. If a product is explicitly mentioned, there is a defined assocation between the company and its product and so the company receives a literal tag.

- ORG-FOR-FACILITY

  This label is applied when one references a building that is related to an organization. Examples are given in 20 and 21.

  (20) Friends clamoured to escape the trauma of walking endlessly up and down Glasgow's Byres Road with a bottle of Hirondelle looking for a party to gatecrash, and so these Hogmanay house parties swelled in numbers yearly until the queue for the bathroom in the morning rivalled *McDonald's* in Red Square. We found a cottage in Torridon that was idyllic.

  (21) Steven was always difficult, he bought every kitchen gadget and electrical device known to man, and he already had enough leather gloves, ties and scarves to restock *Marks and Spencer*.

  Although some may initially view 21 as literal, upon closer examination one see that *restock* indicates metonymy: restocking an object requires a physical location in which the restocking event can take place.

- ORG-FOR-INDEX

  This category is used to refer to the stock price or value of a company. Please see sentence 22 for a more nuanced version of this label.

  (22) Friday's uptick, leaving the index still close to its lowest level since January 1987 and more than 20 per cent down since the start of this year, was prompted by some old-style 'suasion between the Ministry of Finance and the Big Four, Nomura, Nikko, Daiwa and Yamaichi. Investment trust money flowed back into blue chips such as Sony, *Honda* and Pioneer, but banks, financials and government bonds continued to slide.

- OTHERMET

  Similarly to the location-based OTHERMET, sometimes the above categories are not sufficient for classification. In that case, a catch-all, organization-based OTHERMET label may be applied. Please see sentences 23 and 24 for examples.

  (23) But while the Suzuki may be a nimble machine, Schwantz's 1990 *Suzuki* team-mate Niall Mackenzie is a good judge of what makes the Schwantz/RGV combination so rapid.

  (24) Damon Hill is still unsure of his *Williams'* future even though he will be chasing a hat-trick of Formula One wins in tomorrow's Italian Grand Prix at Monza.

  In 23, note that while the first instance of *Suzuki* represents a ORG-FOR-PRODUCT metonymy, it is the second instance of *Suzuki* which is tagged as the PMP. (This also indicates potential quirks of the data, along with the fact that there is only one annotated PMP per sample, even if multiple metonymies exist.[4]) Meanwhile, in 24 *Williams'* does not have a literal label because it refers to the gestalt of a particular person's experience with the company, rather than deal with the company in general.

2.1.3 Class-Independent Categories

The following categories can be applied in either the location domain or the organization domain; this is in contrast to PLACE-FOR-EVENT applying only with the location domain and ORG-FOR-EVENT applying only with the organization domain. A discussion of these three has been delayed until now to prevent repetition. At least one example from each domain is given per category.

- OBJECT-FOR-NAME

  The most common diagnostic is whether a PMP `X` can be replaced by the phrase "name of `X`" (or if such a paraphrase exists or is warranted). In this case, `X` refers to the name of a particular noun, rather than the noun itself. Sentence 25 demonstrates a location-based metonymy, while sentences 26 and 27 demonstrate organization-based metonymy.

  (25) Some phonologists maintain that a syllabic consonant is really a case of a vowel and a consonant that have become combined. Let us suppose that the vowel is. We could then say that, for example, '*Hungary*' is phonemically...

  (26) In the computer industry, the power of trade marks can readily be seen as, in a relatively short space of time, names such as Apple computer, *IBM*, WordStar, Lotus 1-2-3 and BBC computer have become household names. Trade marks are especially important in a fast-moving industry

  (27) He hands Holmes the *Safeway* bag, waits while he frowns over the contents.

  In 25, a key indicator of this metonymy type is the fact that the PMP is quoted, thus focusing attention on it; its subject position in the clause also indicates this; examining the lower-level clauses, as well as the high-level sentences in the PMP is located can therefore be useful.

  Interestingly, this metonymy can be indicated by the use of organization names as nominal modifiers as demonstrated in the latter two examples. This can result in some confusion, which even the annotators note Markert and Nissim (2005), and so using this tag in larger systems (i.e., aside from trying to compare to other systems that have used these data) could be optional: all that would be required is a retraining of classifiers.

- OBJECT-FOR-REPRESENTATION

  This category, like OBJECT-FOR-NAME, has the potential to cause confusion; for thoroughness, however, it will still be discussed. Generally, one can try to paraphrase a PMP `X` with "a representation of `X`." There are no examples of location-based OBJECT-FOR-REPRESENTATION metonymy provided in the

[4] Although adding these readings to the annotations would result in a more complete and thorough dataset, it is beyond the scope of this work. It also introduces difficulties, such as different annotator views; to be more rigorous, multiple annotators would have to be used as well.

data.[5] Sentence 28 is given as an organization example (and is from the data) while 29 is given as a location example (though this is from Markert and Nissim (2005)).

(28) Here Tim King, of consultants Siegel & Gale (3M, Citicorp), had a special fear. BT's pipes-of-Pan motif was, for him, somehow too British. Graphically, it lacked what King calls the 'world class' of *IBM*, Apple Computer, Ford, Sony and Shell.

(29) *Malta* is here.

In 28, *IBM* metonymically refers to the artistic representation IBM presents to the world (presumably through its logo and other associated graphics). Appropriate context for 29 may be given by someone looking at a map and uttering the sentence.

Just as with OBJECT-FOR-NAME, we consider OBJECT-FOR-REPRESENTATION for completeness (and so that results can be compared more easily). To be used with other systems that do not need such discrimination, our methods are simple to change.

- MIXED

Other times, there is sufficient support for both metonymic and literal readings. This occurs if there are at least two, explicitly mentioned predications that indicate conflicting readings. For instance, in 30, "reached agreement" indicates a metonymy while political membership indicates a literal location-based reading:

(30) The three EC members of the Security Council — Britain, *France* and Belgium — have reached agreement on what we want the Security Council to do.

Compare this to 31, in which "ordering" indicates ORG-FOR-MEMBERS, though the *explicit* association of the company with the products it produces indicates a literal reading:

(31) Boeing orders checks on 700 jumbo jets The plane maker, *Boeing* is ordering new inspections on seven hundred jumbo jets following last month's Amsterdam air tragedy.

2.2 Distributed Files

Although the dataset, made freely available for research purposes[6], contains many useful representations of the data, we present examples here of what we found useful or what a significant number of other researchers using this dataset found useful. An example of input data for sentence 10 is presented in Figure 2. For ease of reference, we repeat the metonymic (place-for-people) sentence 10 below:

(10) ... [The resolution] demands that Iraq comply fully... and decides, while maintaining all its decisions, to allow *Iraq* one final opportunity...

Although formatted prose (Figure 2a) was not included, we present it here to clearly indicate the entire passage. While this type of representation may be sufficient for people, to proceed computationally we needed, at the very least, for the PMP to be clearly marked as the target word to focus on; we also needed to represent the PMP within its context (its own sentence, and surrounded by the external context sentences). Although multiple distributed formats provided this information, we found that the minimally-annotated XML (Figure 2b) provided relevant, useful data that were also easy to use. As can be seen, the XML data is simply text with a special "annot" tag indicating a PMP annotation.

The annotators also provided, for each sample, a manually constructed set of important, related words (Figure 2c). A word and the relation connecting it to the PMP were said to be important if they played a significant role in determining the reading label for that PMP. Therefore, for every sample the PMP (column three) and the related, important word (column two) are given. The connecting relation (the entire collection of which form a subset of standard relations in dependency parses) is also given, this time in column 4. For ease of reference, the sample ID (column one) and reading type (column five) are given; that is, the ID and

[5] For rare cases of metonymy, the annotators supplied additional examples that were neither to be used for training nor for testing. Even with these additional data, there are no instances of this metonymy type.

[6] http://www.comp.leeds.ac.uk/markert/MetoSemeval2007.html

Keesings Contemporary Archives. November 1990
Acting under Chapter VII of the UN Charter;

1. Demands that Iraq comply fully... and decides, while maintaining all its decisions, to allow *Iraq* one final opportunity, as a pause of goodwill, to do so;

2.

(a) The sample as may be seen in prose.

```
<sample id="samp1550">
<bnc:title>Keesings Contemporary Archives. November 1990</bnc:title>
<par>
Acting under Chapter VII of the UN Charter;
1. Demands that Iraq comply fully... and decides, while maintaining all its decisions, to allow <annot><location reading="metonymic" metotype="place-for-people" notes="OFF">Iraq </location></annot> one final opportunity, as a pause of goodwill, to do so;
2.
</par>
</sample>
```

(b) The given sample in XML, indicating a PLACE-FOR-PEOPLE metonymy.

samp1550 | allow | Iraq | iobj | place-for-people
ID: samp1550
Related Word Lemma: allow
PMP: Iraq
Relation: iobj
Label: place-for-people

(c) The distributed manual annotation. The sample ID and label were provided only for easy cross-reference, and were not used themselves as features to classifiers.

Figure 2: Example data for sentence 10, a place-for-people metonymy.

label were provided solely to reduce the cross-referencing that would have to be done in training and testing. The labels were not used as features to classifiers.

Regarding the manual grammar annotations, it should be noted that for nearly every sample there is one and only one related dependency word. Thus in most cases, the manual relation annotation is a proper subset of what is returned by a probabilistic dependency parser, as described in Section 5.1.

2.3 Evaluation Measures

Each of the SemEval participants was evaluated against a baseline, which was just the assignment of the most frequent category; for both the location and organization metonymy schemas, the most frequent category was literal. The systems were also evaluated with three different granularities: coarse (literal or not); medium (literal vs metonymic vs mixed); and fine (literal, mixed, class-dependent and class-independent readings).

As the underlying task at hand is an n-ary classification problem, we evaluate the various approaches using standard information retrieval evaluation measures, such as accuracy, recall, precision and F_1 scores. Note that accuracy is influenced solely by the granularity of the specific classification instance, while recall, precision and F_1 are determined by both the granularity and the class in question.

Using the following table,

		GOLD STANDARD	
		True	**False**
PREDICTION	**True**	A	B
	False	C	D

we may easily define the aforementioned metrics:

$$\text{accuracy} = \frac{\#\text{ correct predictions}}{\#\text{ predictions}} = \frac{A+D}{A+B+C+D}$$

while for a given category c,

$$\begin{aligned}
\text{recall}_c &= \frac{\#\text{ correct assignments of } c}{\#\text{ instances of } c} = \frac{A}{A+C} \\
\text{precision}_c &= \frac{\#\text{ correct assignments of } c}{\#\text{ assignments of } c} = \frac{A}{A+B} \\
F_{1,c} &= \frac{2\text{recall}_c\text{precision}_c}{\text{recall}_c + \text{precision}_c}.
\end{aligned}$$

3 Related Work

As discussed in Section 2, scale invariance and dataset invariance pose two significant challenges for researchers working on classification tasks. Particularly for natural language classification tasks, these challenges can commonly create a tension between trying to provide the best computational explanation of a given phenomenon and the best coverage for that phenomenon. Discussion in Section 2 previewed the fact that figurative language resolution is no stranger to these tensions, though the following survey of past work done on metonymy and general figurative language resolution provides concrete realizations of the tensions.

Throughout this section, unless we say otherwise, "abstracting a word via WordNet" should be taken to mean examining the abstract-type classification (artifact, product, etc.) of that word. Please see Table 2 and Section 5.2 for a more formal definition.

3.1 Small-Scale Metonymy Systems

Most of the early work on metonymy resolution was focused on selectional preferences (Fass 1991, Hobbs et al. 1993, Pustejovsky 1991). However, due to the fact that focusing solely on violation of selectional preferences was shown to miss a significant number of metonymies (Markert and Hahn 2002), this section will focus on modern work that combines statistical approaches with some semantic knowledge. However, as most modern work has been corpus-based, this section will be devoted to work on logical metonymy – metonymy that indicates an eventive reading, as in sentence 2 (Bill enjoys *reading* Shakepeare). Although logical metonymy does not always conform to the schematic metonymy we are interested in, examining and understanding previous methods can still provide valuable insight.

Recent successful work in logical metonymy has been done by Lapata and Lascarides (2003), which has since been expanded by Shutova (2009), Shutova and Teufel (2009). In the former, Lapata and Lascarides use the BNC to extract collocation and co-occurence data to model the likelihood of an eventive interpretation conditioned on the verb itself and its object complement; recall though that our PMPs are not always this syntactically nice and easy to handle. The system pursued by Shutova (2009), Shutova and Teufel (2009) obtains collections of synsets. Noting that examining the effect of the noun complement and the effect of the verb could aid logical metonymy detection, they clusters verb senses based on semantic similarity. Similarity is determined by examining BNC collocation data for collections of synsets. Since the basis features are representations of sets, various set operations may be applied to generate actual features (for classifiers). Using k-means clustering, Shutova and Teufel determined that the most expressive feature set represented the union of pairwise intersections of synsets, compensating for data sparsity while attempting to find common-ground among different synsets (Shutova and Teufel 2009). Specifically, this suggests that a fine-grained, more holistic analysis of the interaction between syntactic and semantic features can yield improved results. Unfortunately, one of the potential weaknesses of this method is that its scale and dataset invariance are neither intuitive nor discussed: the model was trained on five sentences and tested on another five sentences; each set targeted logical metonymy and was of the authors' own creation.

3.2 Large-Scale Metonymy Systems

In natural language processing, as a larger number of text corpora has become available, so too has the interest in corpus-based datasets grown. While this is true for metonymy resolution systems Krishnakumaran and Zhu (2007), Martin (1994), Nastase and Strube (2009), Nissim and Markert (2003), Peirsman (2006), until Nissim and Markert (2005) there was no widely standardized large dataset for metonymy recognition. However, the key insight of Markert and Nissim to treat metonymy resolution as a special case of word sense disambiguation facilitated the creation of a metonymy corpus-based dataset: WSD allowed a small number of concrete (schematic) "senses" or readings. The works of Markert and Nissim (2002), Nissim and Markert (2003; 2005) were later extended to form a SemEval 2007 competition Markert and Nissim (2007), aimed at increasing interest in metonymy resolution and advancing the state-of-the-art. As this current work is focused on the SemEval dataset, most of this section will be spent discussing the work done there. Although the depth of the following survey is rather unorthodox, we believe that it is useful, and possibly necessary, in order to better understand the data. This goal is especially important given concerns that have been raised regarding the data.

3.2.1 Users of the SemEval 2007 Data

As to be expected, as the granularity became smaller (from coarse to fine), the performance of all systems degraded. Looking at the accuracy, there was minimal performance dropoff, but there is a big drop off in $F_{1,c}$ score for nearly every category and every system. As a good general rule though, the F scores are very poor, across the board. This means that at least one of either recall or precision needs to be significantly improved. In many cases, both do.

Markert and Nissim (2007) mention that "shallower" methods (such as Leveling (2007), Poibeau (2007)) did not perform well enough to beat the baseline. Further, those two systems only participated in the location task. With few exceptions, the performance of the remaining three systems (as measured by accuracy) were minimally different.

As a brief overview of the results, the baseline accuracy score for the location class was 79.4%, while the baseline accuracy for the organization class 61.8%. The low scores for location were 75.4%, 75.0% and 74.1%, while the high scres were 85.2%, 84.8% and 84.4% (the scores are for the coarse, medium and fine granularities, respectively). The low scores for organization were 73.2%, 71.1% and 70.0%, while the high scores were 76.7%, 73.3% and 72.8% (again, coarse, medium, and fine granularities). The differences in the accuracy results underscore the fact that, as mentioned in Section 2.1, the organization class can be more ambiguous – cf. sentences 17, 26 and 27 – and thus harder to detect and classify.

Dealing only with the location class, Leveling (2007) operated under a variant of the distribution hypothesis: they assumed (somewhat vaguely) that "metonymic location names can be identified from the context." They extract three different contexts using variable-sized windows (a certain number of words to the left and right of the PMP, including the PMP itself). The first context was hypernymy information; the sentence context was basic syntactic and lexical information; the third was word information (such as capitalization, symbol information, POS tags, etc.). They use WordNet solely to extract synset IDs (the first context), and not for word sense disambiguation.

Using leave-one-out cross-validation with a memory-based learner, they experiment with the parameters (the window sizes for the three contexts). They find that generally context on either side of the PMP is necessary, though for WordNet only hyponymy information for the PMP itself is useful. This method generated results that were not significantly different from the baseline. Various factors could be influencing these results: first and foremost, the memory-based approach could be a significant limitation. They were also plagued by poor tokenization, lemmatization and sentence chunking.

For both the location and organization classes, GYDER Farkas et al. (2007) had the best accuracy results across granularities (though to be fair, sometimes the differences – within 0.001 – were extremely small). They used a publicly-available maximum entropy toolkit, with Gaussian prior equal to 1, and 5-fold cross-validation to optimize feature usage. The main features used include the provided grammatical annotations, determiner classification, and the plurality of the PMP. From the grammatical annotations, the most useful feature they were able to extract was WordNet hypernym information for the first sense of the PMP; other features such as Levin verb classes and a manually built "trigger table" of metonymy-inducing words and phrases provided insignificant benefit. Determiner classification (whether a determiner

was definite, indefinite, demonstrative, possessive, etc.), along with a boolean indication of whether the PMP was sentence initial, proved discriminative.

They provide a coherent exposition of other features with which they experimented but found to be unhelpful. A more thorough description is left to their report Farkas et al. (2007); but briefly, most of the less discriminative features could be considered to have high variance (or be very specific to that PMP), such as named-entity labels, orthographic features and the inflectional category of the closest verb to the PMP.

Aside from being a good resource as to what features have been useful, they conclude that detecting metonymy, rather than classifying it into the various categories, is the main challenge that must be overcome, partially because the different categories can signal very different contexts: cf. sentence 32 (a ORG-FOR-PRODUCT of our own creation) with 20.

(32) Bill will be here soon; he wanted to grab *McDonalds* first.

To help with detection, they suggest that more semantic knowledge should be used. They also discuss their desire for different annotations and datasets, as data sparsity was a significant issue for some of the categories.

Poibeau went for the minimalist approach, trying to establish a lower-bound. For instance, Poibeau does not make use of a POS tagger, syntactic parser or semantic parser. Instead, the approach is mainly based on a distributional analysis of surface word forms and a filtering process to restrict the metonymic readings to country and capital names, even though such an over-simplification has no linguistic basis Poibeau (2007).

Operating only on the coarse-grained setting, he takes a window around the PMP and associates the words in that window with two classes, either literal or non-literal. This collocation analysis results in lists of discriminative words — that is, words that are either very frequent or very rare in one corpus compared to another. The list yielded high precision but very low recall; he concludes that syntactic, let alone semantic, analyses would be very useful. Further, he shows that surface forms can be a fairly reliable way to reduce the size of the search space.

A high-performing system, `UTD-HLT-CG` Nicolae et al. (2007) extracts some surprisingly simple features, which may be categorized as either syntactic or semantic. The syntactic features include POS tags and lemmas for words adjacent to the PMP (including the PMP), determiner analysis, and governing preposition analysis. Noting that possession can indicate metonymy, they examine that as well as whether a PMP is inside quotes. The semantic features include semantic roles of the PMP and adjacent tokens (for both the words and the corresponding lemmas), Levin roles and Lexical Conceptual Structure (LCS) roles. Interestingly, WordNet is not used.

To determine which feature combination yields the best results, they put everything into a suite of machine learning algorithms, and experiment with decision trees, decision rules, logistic regression, and "lazy" classifiers such as k-nearest neighbor. They did not find that the number of features resulted in data overfitting.

Brun et al. (2007) created `XRCE-M`, an unsupervised metonymy resolution system, which built upon a commercial parser. Although they construct rules (such as, "If a location name is the subject of a verb referring to an economic action, then it is a PLACE-FOR-PEOPLE"), they do not provide a deep discussion of the efficacy of these rules, how many there were, and how they determined predicate values (such as "a verb referring to an economic action").

To supplement the syntactic information and symbolic rules, they adapt Harris's distributional hypothesis Harris (1954). However, the approach is somewhat complicated and describing it in adequate detail would be distracting; the reader is therefore directed to their paper for an overview and a complete example Brun et al. (2007).

Although the system performs very well (one could consider that, based solely on accuracy, it got second place), Brun et al. (2007) note that the data seems to be somewhat skewed, as the results from the test portion were lower than the results from the training portion. These remarks echo concerns voiced by Leveling (2007), Nicolae et al. (2007). Parsing errors, MIXED classifications and uncovered contexts (from their distributional approach) were the primary sources for missed detections and classifications.

One of the goals in creating and releasing the data for the competition was to inspire future work in metonymy resolution. In addition to facilitating this work, the SemEval dataset provided the basis for Nastase and Strube (2009). Noting that local grammatical context can significantly determine a reading label, they rely on the distributed, manually-constructed grammatical annotations in order to build off of base features

from Nissim and Markert (2005), such as the grammatical role of the PMP, an analysis of the determiner (if any) of the PMP and grammatical role distributions. To expand upon these base features, they start by estimating selectional preference features, which involves extracting WordNet-abstracted dependency collocations from the BNC. To refine the experimental probability estimates they use a combination of WordNet and Wikipedia to extract *IsA* and *Supersense* relations.

To expand these statistical data, they mine Wikipedia pages to define the binary predicates *has-product* and *has-event*; these look for references to and uses of manufacturer- or event-based nouns in pages. While these are good ideas one may expect their utility to be diminished since some metonymys require inference chaining, or be more anecdotal in nature. Using an SVM implementation Hall et al. (2009), they verified this concern. Despite this, they still achieved, nearly across the board, the best recorded results on the SemEval dataset.

Technically Peirsman (2006) did not evaluate against the SemEval data, but rather against the dataset from Nissim and Markert (2005); however, the method and results are still interesting and applicable to our work, since the SemEval data are based on those used in Nissim and Markert (2005). Reacting to the complexity of the method in Nissim and Markert (2005), which included iterative smoothing over a thesaurus, Peirsman wanted to show that a simpler learning method could perform competitively. Operating under the hypothesis that we should be able to use the fact that schematic metonymy is possibly pattern-based, he used a memory-based approach. Arguing that people learn and adapt to new situations simply by remembering what they have already seen, he extracted overt syntactic and semantic features (much like what is described above and in the beginning of Section 6). Though he does not explicitly say it, it is very probable that he used the manually constructed grammatical annotations, just as others have done. Using these features he achieved an accuracy of 86.6% (74.6%) and an F_1 of 0.612 (0.651) on locations (organizations). There is no mention of granularity, so it is again most likely safe to assume that this was evaluated under what we would call a coarse granularity. However, it is unclear the exact ways in which his dataset differed from ours.

3.3 Computationally Approaching Other Types of Figurative Language

Recently, there has been work in removing underspecification in knowledge-base queries (Fan et al. 2009), and analyzing sentence cohesion to classify whether a phrase is being used literally or not (Li 2008, Li and Sporleder 2010). In the former case, removing underspecification, Fan et al. work with what they call "loose speak," of which metonymy is a type. At its core, their algorithm represents sentences and phrases as triples between two classes and a relation, and attempts to determine if there was a domain mismatch between the classes and the types expected by the domain and range of the relation. They perform this domain analysis via a graph search, refining the representation and moving up the hierarchy until it passes certain constraints.

Meanwhile, (Li 2008, Li and Sporleder 2010) have detected non-literal language usage within passages by analyzing the cohesion; basically, they determine if some words do not "fit" with the rest of the sentence. Specifically, for a given query phrase, they determine how related words are (generally pairwise) in a sentence and then apply those features to a cohesion-graph approach (Li 2008) or Gaussian mixture models. Although they say that their approach handles metonymic language, the examples offered in their papers are primarily idiomatic phrases, where the wording of a specific phrase may be brought into question.

4 Computationally Modeling Metonymy

After examining metonymy on a more theoretical basis, and analyzing it with respect to other tropes and types of figurative language, we saw how metonymy is realized in our dataset of choice. In addition to providing concrete examples, this process allowed us to get a sense of potentially useful heuristics that people may actually use when trying to classify and disambiguate schematic PMPs. This whirlwind introduction to schematic metonymy also enabled us to start formulating computationally feasible features that could be extracted and help in our overall task. We were then able to analyze the intersection between past methods and our own intuitions and heuristics, as developed in Section 2. By studying past work on schematic metonymy, in addition to logical metonymy, other tropes and idiomatic phrases, we have thus been able to get a sense of what features should be discriminative, or at the very least helpful, in metonymy classification.

Most notably, systems that are able to consider the intersection of syntactic and semantic features have tended to perform well.

The main hypothesis of this work is that a better modeling of the underlying syntactic, semantic and conceptual meanings within a document can aid automated metonymy classification. Note that as lexical and syntactic features are very easy to extract, but semantic and pragmatic features are more difficult to extract, we can think of this as trying to better map the lexical/syntactic features into semantic/pragmatic ones (what we call "mapped semantic" features). However, there are many possible ways to test this hypothesis: this section describes three primary methods strongly considered to test the hypothesis, including one method for which investigation and development was halted in order to focus on the other two methods. The first implementation, which is no longer being tested in this work but still described in Section 4.2, experiments with the idea that a rule-based approach, typically discounted among modern researchers and in large-scale NLP projects, can perform competitively on this type of schematic metonymy. The second, described in Section 4.3, assumes that the features of a PMP subsume those of its associated context words. It experiments with various machine learning algorithms on a suite of novel syntactic, semantic and "mapped semantic" features. Finally, the third, given in Section 4.4, attempts to provide a more explicit model of how the lexical and syntactic interact, while inducing semantic correlations and relationships.

4.1 Standardizing Notation

Prior to discussing models, it is useful to standardize a notation that will be used throughout the remainder of this work. As discussed in Section 2.1, the potentially metonymic phrases (PMPs) are contained in sentences that have a context window of up to two previous sentences and one following sentence. The PMP-based sentence, along with its context window, is interchangeably called a **document** or a **sample**. It is helpful to be able to describe a document in at least two ways, based on what we consider the foundational building blocks of the sentence. First, since the sample is just an ordered collection of words, we may view a sample as a (potentially flattened) sequence $\mathcal{W}$. Specifically, we may give the sentence containing the PMP an index of 0, so that the initial two have indices -2 and -1, respectively, and the following sentence has index 1. We reference a word at position j (0-indexed) in sentence $i \in \{-2, -1, 0, 1\}$ by $w_{i,j}$. The sequence $\mathcal{W} = \left\{ \{w_{i,j}\}_{j=0}^{l_i - 1} \right\}_{i=-2}^{1}$ represents the entire passage (l_i is the number of words in sentence i). Note that by thinking of a sample this way we preserve the ordering of words and so easily allows for index-based subsequences to be extracted.

However, if we want to capture some of the more complex linguistic properties of a document, then it helps to first have foundational building-blocks. Dependency relations are extracted from syntactic characteristics, but can also provide some semantic information; as much work has been done on creating reliable dependency parsers, it is natural to consider this information the "foundation" of a document. We easily represent a dependency relation δ between two words $w_{i,j}$ and $w_{i,k}$ by the relational pair $\delta(w_{i,j}, w_{i,k})$. If we assume that the important information about a sentence is captured by these dependency relationships, we can consider a sentence $\mathcal{S}$ to be a set of these relationship pairs. A document $\mathcal{D}$ is then simply a finite (ordered) set of sentences. Thus the following holds:

$$\mathcal{D} \ni \mathcal{S} \ni \delta(w_{i,j}, w_{i,k}). \tag{1}$$

Note that by using these two representations simulatenously, neither linear orderings nor tree-structure constraints impede our ability to easily examine subsets (subsequences) of words. For instance, for a PMP `X` (in sentence i, position j), if we want to analyze all intra-sentential "context" words $\mathcal{C}_k$ connected to `X` via a dependency parse and within a k-word window, that set $\mathcal{C}_k$ is just

$$\mathcal{C}_k = \{w \mid \delta(w, \texttt{X}) \text{ or } \delta(\texttt{X}, w)\} \cup \{w_{i,m} \mid |m - j| < k\}. \tag{2}$$

4.2 Rule-Based Approach

In crafting the dataset, Markert and Nissim (2007) specifically decided to focus on location- and organization-based metonymy, in part due to the ease with which those categories lend themselves to schematic metonymy. Further, as seen above many of the metonymies can be justified using certain rules and heuristics (see

Markert and Nissim (2005), and Section 2.1 of this work). Despite these facts, most researchers are sticking with statistical techniques without entertaining a rule-based approach: out of all the reseachers to use this dataset (including the five participants in the SemEval task), only one group has incorporated a rule-based approach into a detection system Brun et al. (2007). While it is potentially promising that Brun et al. (2007) was consistently in the top two in the competition, attempting to learn from their rules — from the more successful rules to the less successful rules — is not straight-forward. They offer an example of a rule, such as "A location-based PMP, which is the subject of a verb relating to an economic action (such as *import, repay*, etc.) is a PLACE-FOR-PEOPLE metonymy," though unfortunately, there is no discussion of the effectiveness of the rules in their work. Similarly, the entire set of rules is not made publicly available. This leaves the narrative incomplete, as we are unable to know basic information about the rules, such as how many there were and how precise they were. It also makes improving upon the rules and advancing a high-performing system difficult.

Despite the above difficulties, it is especially for a rule-based method that the "if-then" heuristics presented in Section 2.1 seemed promising. At a high-level, these rules could be represented fairly easily in first-order logic formulae. This would have the potential benefit of allowing for easy inference, as well as providing for a smooth incorporation into larger knowledge, planning and question-and-answer systems. Rules also have the benefit of being more friendly and accessible to human intuition and reasoning (as compared to, say, the weights in a support vector machine or a multilayer perceptron). Since in trying to construct first-order logic rules it is only natural to create a correspondence between extracted features and predicates, a significant portion of the problem reduces to reliably extracting the necessary features.

However, in this feature extraction task lies the problem. The rules and heuristics provided in Section 2.1 are reasonable to humans, but it is non-trivial to develop an algorithm that accurately measures whether a paraphrase can be added to a PMP within the context of a sentence in order to achieve a proper level of specification. We appeal to sentence 8, where an algorithm must be able to reliably decide that mapping *Cuba* to *the people associated with Cuba* arrives at a suitable level of specificity in order to recognize and classify the metonym.

Brun et al. attempted to get around this problem by adapting a commercial parser to generate these rules; however, little information is presented or made readily available regarding the engineering required to generate the features. For instance, it is not clear how the "dedicated lexicons," which are used to determine whether or not a verb relates to an economic action, are populated. However, even with these dedicated lexicons there is still the above mentioned paraphrasing problem, among others.

These issues were not what paused work on the rule-based approach. Ideas similar to those described in Sections 6 and 8 were considered for predicate extractors. Unfortunately, as discussed later, those methods were eventually determined to be insufficient, due to a combination of poor coverage, low precision and high noise. Especially when we considered that many of the other features and predicates, as outlined in Section 2.1, were to be used in our second and third methods (Sections 4.3 and 4.4, respectively), it was determined that hand-crafting "hard-logic" rules would more-or-less amount to emulating a standard machine learning algorithm.

We therefore determined that, considering the above constraints — in addition to time constraints, it was the most logical and potentially productive decision to pursue the other two modeling methods. However, this is not to be taken as advocating against a rule-based approach. On the contrary, we believe that a rule-based approach could be very competitive, and needs to be studied. It is unfortunate that pursuing the approach with the focus and energy it would need and deserve, would have diverged and diverted attention from the main questions being asked. Studying a rule-based metonymy classifier for corpus text is an area that has the potential to be very fruitful, and is discussed in Section 8.

4.3 PMP Targeted Classification

Recall from Section 3 that most modern work on metonymy resolution reduces first to the extraction of a single feature vector $\Phi(\cdot)$, and then to the application of various machine learning algorithms to try and learn from these features. Even though metonymy can be seen as potentially requiring the synthesis of multiple words or phrases, as the goal of our classification task is to classify a single specific (noun) phrase, it is reasonable to represent a passage by a single feature vector. While the auxiliary and surrounding context words for a given PMP (e.g., those words connected to the PMP in a dependency tree) provide most of the

discriminatory features, our task is not one of sequence-tagging: we are only concerned with the label of the PMP (note though that in the next section, we examine ways to more explicitly model those discriminatory context words). We may thus consider characteristics and features of the context words to be inherent to that specific PMP.

However, even though we may group the extracted features from the PMP and its related words together, our feature engineering still needs to be able to model the underlying meanings in the samples; the process required to extract these features is discussed later in Section 6. Although the two highest ranked groups operating on the SemEval data used a support vector machine Nastase and Strube (2009) and a maximum entropy model Farkas et al. (2007), there is no clear indication how the novel features we extract will interact with other, more deeply studied features. Therefore, once the feature vector Φ is populated, a variety of classifiers will be tested. These classifiers include support vector machines, neural networks, k-nearest neighbor, Naive Bayes, and decision tree-based learners. Although some of the more powerful tree-based learners, such as a random forest Breiman (2001), may have an experimental tendency to overfit[7], they may still be able to provide an approximation for how the extracted features would work in a rule-based approach. Note that this would very likely be a very rough approximation and non-promising results may not necessarily be conclusive.

4.4 Hidden Conditional Random Fields (hCRFs)

Ever since having been proposed, Harris's distributional hypothesis — that words occurring in the same contexts are synonymous Harris (1954) — has become a staple in NLP tasks where semantic correlations may be extraordinarily predictive, but at the same time extremely difficult to obtain. Specifically, consider that by using explicit collocation data on our dataset, Poibeau (2007) was able to achieve high precision — despite a linguistically-ignorant approach — and Brun et al. (2007) were able to achieve high accuracy across granularities. Further, the work of Shutova and Teufel (2009) indicates that such collocation data can be useful for semantic clustering and potentially for disambiguation. However, in each of the above cases, the collocation data are modeled (computed) explicitly, and then used as a feature. As a step toward arriving at a better mapping into semantic space, one of the questions we wanted to ask is if these collocation data can be modeled more implicitly and learned from the samples themselves, by *explicitly* modeling the interactions ("transitions") between words. Learning these most naturally implies the use of a chain structure, such as a conditional random field.

Conditional random fields (CRFs) have been quite successful in NLP, particularly for chain-based tasks Finkel et al. (2005), McCallum et al. (2004), Peng and McCallum (2004). Unlike hidden Markov models, they are not limited by overly-simplistic independence assumptions, and unlike maximum entropy models, they can avoid the label bias problem Lafferty et al. (2001). However, as Klein and Manning (2002) have shown, the observation bias problem can be just as significant as the label bias problem.[8] It therefore is important to choose the probabilistic method that best models the data in question; simply choosing the most "powerful" does not necessarily lead to better performance. However, given the success of a maximum entropy approach Farkas et al. (2007), we believe that observation bias is not a significant problem.

Additionally, we believe that a discriminative approach is, in the end, more appropriate since there are complex interactions among the linguistic information of the words in a passage, and our task is not concerned with modeling data so as to support generation. Further, the best approach employed by Nastase and Strube (2009) provides support for a discriminative method.

Although conditional random fields (CRFs) can theoretically be defined using cliques of any size in a graph, typically for practical matters (training time, size of training set) a linear chain CRF is used. In a linear-chain CRF, the sequence of state variables $\mathbf{y} = \{y_t\}$ is conditioned on a sequence of observation data $\mathbf{x} = \{x_t\}$, for $t = 1, 2, \ldots, T$ and a set of parameters θ; the cliques are just the edges in the graph. Adopting

[7] Breiman (2001) claims that random forests do not overfit, although as we discuss in Section 7, significant disparities between training and testing error may suggest otherwise. Other work has briefly echoed these concerns as well Statnikov et al. (2008).

[8] Briefly, the label bias problem corresponds to low-entropy in next state transitions, while the observation bias problem corresponds to the observations being disproportionately more predictive than the states. Label bias can result in a model being so confident in its transitions that it disregards the input, while observation bias can result in a model being so confident in its observations that it does not adequately consider its state. A thorough exposition of the label bias and observation bias problems are left to Lafferty et al. (2001) and Klein and Manning (2002), respectively.

common notation, the conditional distribution is given by

$$P(\mathbf{y}|\mathbf{x},\theta) = \frac{1}{Z(\mathbf{x})}\prod_{t=2}^{T}\Phi(y_{t-1},y_t,\mathbf{x};\theta_t),$$

where

$$\Phi(y_{t-1},y_t,\mathbf{x};\theta_t) = \exp\left(\sum_k \theta_{t,k} f_k(y_{t-1},y_t,\mathbf{x})\right)$$

is a potential function, $\{f_k\}$ are predetermined feature functions and $Z(\mathbf{x})$ is a normalization function. For linear-chain CRFs, training is a convex optimization problem, which generally means that the training is both efficient and reaches non-local (global) optima Lafferty et al. (2001).

The examples shown throughout this work indicate that a variety of features – lexical, syntactic, semantic and pragmatic – help influence metonymy resolution. Therefore, we believe that modeling these features, *and how they interact*, is necessary. Since we have the potential to consider arbitrary subsequences of $\mathcal{W}$, modeling the various features could require an account of long distance (possibly inter-sentential) relations and dependencies. What we would like to do is be able to use a CRF to create a chain across some (possibly improper) subsequence of $\mathcal{W}$.

However, there is at least one potentially crippling issue that must be addressed: CRFs label individual time occurrences in data sequences, but we are only concerned with one "time occurrence:" the label of the PMP. To achieve our sequence-based model, we can begin by treating each word in the given subsequence as a single time occurrence, but we quickly run into trouble since we have only one labeled time occurrence (the PMP) per sample. This poses a problem because all of the other words in the subsequence provide clues for the intended reading, but there is no predefined, obvious label for them. One initial solution would be to label all the words in a subsequence with the PMP's label: if the PMP is to be read literally (metonymically), then all words in the sequence are to be read literally (metonymically). Even though this could work, as has been done in computer vision tasks without harm Kjellström et al. (2008), and indeed appeals to intuition (surrounding auxiliary sequence words should be read the same), it 1) places an unnecessary emphasis on particular word/time step labeling, and 2) could make the evaluation unnecessarily complex. Instead, we use hidden conditional random fields (hCRFs) Quattoni et al. (2007) — a variant of CRFs — which have been shown to yield greatly improved results Wang et al. (2006).

In hCRFs, the observed data variables are augmented by hidden data variables; these hidden variables allow the model to capture both near and remote dependencies Wang et al. (2006). Rather than attempt, in a possibly suboptimal manner, to explicitly endow the observed data with additional structure, we let the model learn the latent structure. A depiction of an hCRF for this work is given in Figure 3.

The details of the training (and derivation of the distribution model) are left to Quattoni et al. (2007), Wang et al. (2006), though we present the final distribution model here. The probability of an observation y given observation data $\mathbf{h}$ (hidden variables) and $\mathbf{x}$ is

$$P(y|\mathbf{x};\theta) = \frac{1}{Z(\mathbf{x})}\sum_{\mathbf{h}}\prod_{t=1}^{T}\Phi(y,h_t,\mathbf{x};\theta_t^h)\prod_{t=2}^{T}\Phi(y,h_{t-1},h_t,\mathbf{x};\theta_t^{hh}),$$

where the potential functions are given by

$$\begin{aligned}\Phi(y,h_t,\mathbf{x};\theta_t^h) &= \exp\left(\sum_k \theta_{t,k}^h f_k(y,h_t,\mathbf{x})\right)\\ \Phi(y,h_{t-1},h_t,\mathbf{x};\theta_t^{hh}) &= \exp\left(\sum_k \theta_{t,k}^{hh} f_k(y,h_{t-1},h_t,\mathbf{x})\right).\end{aligned}$$

5 External Resources

In this section, we list the external resources, tools and libraries used to create the system. As many of the resources we use are well-known and well-studied (e.g., WordNet), we do not provide an extensive exposition.

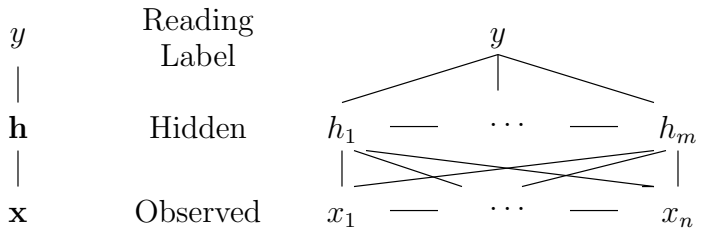


Figure 3: A hidden conditional random field (hCRF), the primary graphical model proposed for use in this work. See Wang et al. (2006) for full details.

However, we discuss and describe those resources that a reader is less likely to be familiar with, or those for which a special notation or assumption is adopted throughout this work.

5.1 Parser

Even if we were not trying to make the classification of metonymy as automated as possible, the manually-constructed grammatical annotations (see Section 2.2) would not have been sufficient for our purposes. Recall that those annotations only provide a dependency relation if it was related to the PMP; without this human "oracle" an automated system will need a way to obtain (a superset of) these relations. Therefore, despite the admonition that the use of automated parsers significantly decreases performance Nastase and Strube (2009), we used the Stanford dependency parser Klein and Manning (2003a;b) for parsing each of the documents and extracting the dependency relations. We also obtained the part-of-speech tags and parse trees, though we did not find the parse trees to be of use. We noticed that we did not have significant problems with the automated parses, despite what was indicated in Leveling (2007).

5.2 Knowledge Bases

As briefly discussed in Section 3, early metonymy resolution work focused on the violation of selectional preferences. Even though limitations of using selectional preferences have been identified Markert and Hahn (2002), examples presented in this work clearly indicate that an analysis of thematic roles can be helpful (e.g., sentence 5). To help extract the expected θ-roles, we use VerbNet 3.1 Kipper et al. (2000; 2006).

However, not all PMPs can be decided by verb-argument analysis (e.g., as a nominal modifier, as in sentence 9). For those, we must analyze words related to the PMP; this analysis must not be too sparse so as to prevent machine learning algorithms from actually learning relationships, but at the same time not overly general. Wordnet 3.0 Fellbaum (1998) allows us to obtain this desired degree of specification. We use the MIT Java Wordnet Interface[9] to access WordNet.

In what follows, we describe a notation we adopt, which will help describe the algorithms in Section 6. Although sets are technically unordered, when we mention sets we assume that some ordering has been placed on them, and that this ordering is applied throughout. For instance, for the WordNet-defined part-of-speech tags $\mathcal{P} = \{\text{NOUN}, \text{VERB}, \text{ADJ}, \text{ADV}\}$, we assume all iterations over $\mathcal{P}$ are in the order first NOUN, then VERB, etc. This is to simplify the presentation. Note that the precise orderings for any of the sets discussed does not matter; all that matters is that an order be given and used consistently throughout. In addition we make use of the lexical file/abstract entity-type information for each word. Since these may be categorized by part-of-speech tag, let $\mathcal{L}_{\mathbb{P}}$ be the set of abstract entity-types for a given part-of-speech tag $\mathbb{P} \in \mathcal{P}$; for example, NOUN.ARTIFACT, NOUN.COMMUNICATION $\in \mathcal{L}_{\text{NOUN}}$ and VERB.CHANGE $\in \mathcal{L}_{\text{VERB}}$. Note that for ease of reference we may define $\mathcal{L} = \bigcup_{\mathcal{P}} \mathcal{L}_{\mathcal{P}}$. The order of enumeration across all $\mathcal{L}$ is given by the column-read ordering for each $\mathcal{L}_{\mathbb{P}}$, as in Table 2.

[9] http://projects.csail.mit.edu/jwi/

noun			verb		adj	adv
TOPS	FEELING	POSSESSION	BODY	MOTION	ALL	ALL
ACT	FOOD	PROCESS	CHANGE	PERCEPTION	PERT	
ANIMAL	GROUP	QUANTITY	COGNITION	POSSESSION	PPL	
ARTIFACT	LOCATION	RELATION	COMMUNICATION	SOCIAL		
ATTRIBUTE	MOTIVE	SHAPE	COMPETITION	STATIVE		
BODY	OBJECT	STATE	CONSUMPTION	WEATHER		
COGNITION	PERSON	SUBSTANCE	CONTACT			
COMMUNICATION	PHENOMENON	TIME	CREATION			
EVENT	PLANT		EMOTION			

Table 2: A listing of the abstract WordNet categories we consider. We may reference the entire collection by $\mathcal{L}$, and each column by $\mathcal{L}_{\mathbb{P}}$, where $\mathbb{P}$ is the given part-of-speech tag. As an example, VERB.BODY $\in \mathcal{L}_{\text{VERB}}$. The ordering is given first by POS tag, and second within each $\mathcal{L}_{\mathbb{P}}$ by a column enumeration (top to bottom, left to right).

5.2.1 ConceptNet

Although WordNet provides a lot of useful information, it is rather limited by its primary organization as an *IsA* hierarchy. For instance, it is by no means easy, let alone elegant, to represent desire or motivation. Indeed, as Nastase and Strube (2009) implicitly noted, there is only so much that can be done with WordNet alone. They attempted to use encyclopedic information, as provided by Wikipedia, to help metonymy classification. Unfortunately, this was non-discriminative: while Wikipedia tries to provide general knowledge[10], it is still written by, and more importantly for, humans. Humans do not need, or want, the type of more axiomatic knowledge required to solidify conceptual connections between syntactically related words. We make use of ConceptNet Liu and Singh (2004) in our attempts to extract this more axiomatic, encyclopedic knowledge.

ConceptNet Liu and Singh (2004) is a semantic network that attempts to expand upon the idea of WordNet by not only describing an *IsA* relation between two words, but also more abstract and descriptive relations such as *ConceptuallyRelatedTo* and *CapableOf*. These relationships are automatically extracted from input data (text, other natural language) that are added to the database by any registered user of the Common Sense Initiative website[11]. Formally, ConceptNet is defined as a directed multigraph (pseudograph) $G = (\mathcal{V}, \mathcal{E})$ where $V \in \mathcal{V}$ is called a **concept** and $E \in \mathcal{E}$ defines some sort of relationship between two concepts; it is because we wish to encode more semantic information than a simple existence (*IsA*) arc allows that G is a multigraph. For applications, it suffices to consider a concept simply as a lemmatized word/phrase. To represent a given relation, each edge relationship E can be defined as a triple of some predefined concept-relation R between concepts V_1 and V_2. As stated, the relationships $R \in \mathcal{R}$ are predefined to include *IsA*, *HasA*, *ConceptuallyRelatedTo*, etc. A complete listing of the relations is given in Table 3.

Letting $E = (R, V_1, V_2)$, $E \in \mathcal{E}$ if and only if some editor or moderator determined that $R(V_1, V_2)$ makes sense. The functional relation $R(V_1, V_2)$ is generally read as "a V_1 R V_2." For example, the following are defined:

- CapableOf(dog, bark)
  A dog is capable of barking.
- UsedFor(bone, dog)
  A bone is used (in some manner) for dogs.
- ConceptuallyRelatedTo(pet, dog)
  A pet is conceptually related to (in an unspecified way) a dog.

[10] Not though, necessarily common sense knowledge alone.
[11] http://csc.media.mit.edu/conceptnet/

Relationship ($R \in \mathcal{R}$)	**Guiding Questions**	**Ranking** $\mathfrak{R}(R)$
IsA	What kind of thing is it?	23
SymbolOf	What does it represent?	23
CapableOf	What can it do?	22
ConceptuallyRelatedTo	What is related to it in an unknown way?	22
MadeOf	What is it made of?	21
HasProperty	What properties does it have?	20
LocatedNear	What is it typically near?	19
AtLocation	Where would you find it?	18
HasA	What does it possess?	18
PartOf	What is it part of?	17
UsedFor	What do you use it for?	17
Causes	What does it make happen?	16
Desires	What does it want?	15
ReceivesAction	What can you do to it?	14
CreatedBy	How do you bring it into existence?	13
DefinedAs	How do you define it?	12
MotivatedByGoal	Why would you do it?	11
CausesDesire	What does it make you want to do?	10
ObstructedBy	What would prevent it from happening?	9
HasSubevent	What do you do to accomplish it?	8
HasPrerequisite	What do you need to do first?	3
HasLastSubevent	What do you do last to accomplish it?	2
HasFirstSubevent	What do you do first to accomplish it?	1

Table 3: Our own defined example ranking $\mathfrak{R} : \mathcal{R} \to [0, 23]$ (23 is the number of recognized relationships for ConceptNet) of relationships R. Higher is better.

- ConceptuallyRelatedTo(god, dog)
  Sometimes ConceptNet has interesting entries. It is most likely because "god" is "dog" reversed.
- Desires(dog, drink water)
  A dog wants to drink water.
- ConceptuallyRelatedTo(dog, pet)

Note that we may define real-valued functions over $\mathcal{E}$. Some functions, including frequency (how many times E was added) and a reliability score (how many people agreed with E), are distributed with the database. While these frequency and reliability metrics are useful, they are not much help when trying to rank relations semantically. Unfortunately, there has been little work in defining general relation rankings and so we are left to define our own internal scoring function $\mathfrak{R} : \mathcal{R} \to \mathbb{R}$; an example of such a function is given in Table 3. We may change $\mathfrak{R}$ to try to better rank the connections (or, from a word sense disambiguation point-of-view, help choose the better sense). For example, we could say that if two concepts are connected by more than one relationship (i.e., $R_1(V_1, V_2)$ and $R_2(V_1, V_2)$), and if $\mathfrak{R}(R_1) > \mathfrak{R}(R_2)$, then we may consider the R_1 relationship to be more reliable *for our task*. However, there are issues with ConceptNet that must be acknowledged; as will be discussed in Section 6.3, a significant number of these issues cannot be avoided in the current release of ConceptNet. Therefore, although we have the ability to define our own scoring function $\mathfrak{R}$, we experimentally determined that using $\mathfrak{R}$ would present more issues than could be appropriately handled at this time.

5.3 Machine Learning Libraries

All of the above mentioned resources allow us to extract features from the samples, but to actually make use of them, we must apply some learning algorithm to them. For many of the standard algorithms, such as those described in and used for Section 4.3, we employ Weka, a widely used and publicly available machine learning library for Java Hall et al. (2009). Unfortunately, Weka does not have native support for (generalized) CRFs or factor graphs, something that would be needed to construct an hCRF. Instead, we use the HCRF library associated with Quattoni et al. (2007), and available online[12].

6 Feature Definition, Extraction and Engineering

The previous section contained descriptions of the models that are used; the features that will power the models are described here. Unless otherwise stated, it is safe to assume that any features being described apply to both the PMP-targeted learner and the hCRF sample tagger. As has been common with many metonymy researchers, we extract syntactically-based (Section 6.1) and semantically-based features (Section 6.2). However, rather than rely on them we attempt to complement them with additional features that derive a deeper conceptual understanding of the sample (as it relates to metonymy).

We also clearly indicate what features are used in what model. Since we present the features in a general fashion, we often define them based on a given input query word. Recall that, unless otherwise stated, the query word for Model One is the PMP, while for Model Two it is any word in the provided subsequence. That is, in Model Two, we apply the feature extractors to every word in the subsequence.

6.1 Syntactically-Based Features

Based on the heuristics from Section 2.1, syntactically-based features can provide big hints for determining the proper reading. That being said, the syntactic features are notably limited and they should not be expected to decide and justify a significant number of samples. The syntactic features of a word $w_{i,j}$ include determiner and preposition analysis, as well as the binary indication of whether or not $w_{i,j}$ is possessive. There is support for these features from Brun et al. (2007), Farkas et al. (2007), Nicolae et al. (2007), in addition to some of the observations in Section 3.2.1. Quotation and parenthetical analysis (if the word is contained in quotes, respectively parentheses) examine whether or not the reader is meant to focus on the PMP, which could indicate a metonymy (e.g., sentence 25). It should be noted that these syntactically-based features are not expected to be discriminative in their own right; rather, they are intended to act as "conditioning" features.

These features are used in both Models One and Two.

6.2 Semantically-Based Features

However, as Leveling (2007), Poibeau (2007) showed, syntactic information is not enough to determine metonymy (let alone the syntactic features we extract). To get around this limitation, we would like to incorporate semantic information into what the model learns. In the past, WordNet has provided crucial ontological information. The hierarchical structure alows us to abstract words to their high-level entity type. We also record synset information. Specifically, if given a part-of-speech tag for a word, we obtain this abstraction for the top k-senses. In the past, researchers have successfully been able to use only the first sense, despite an acknowledgement that this frequently is not the proper sense to examine Nastase and Strube (2009), Peirsman (2006). The semantic features of a word $w_{i,j}$ include the grammatical and semantic role of that word. When applicable, VerbNet is used as well to get the expected grammatical roles induced by the governing verb. As discussed previously in Section 2.1.1, this is not always natural.

In the next section, we define a word-scoring function s, which at its core is based on WordNet. While we use its results as features in some cases, we describe it in the next section, as it is central to the method described there.

All of these features are used in both Models One and Two.

[12] http://sourceforge.net/projects/hcrf/

6.3 Extracting Underlying Conceptual Meaning

The syntactically-based features explain how connected words are used, while the semantically-based features describe what those words mean. By combining the two we can describe why those words are being used in the ways that they are. Although explicitly-computed collocation data can model some generalizations between meaning and usage, such modeling is generally implicit: from these data there are not always easily obtainable explanations for connections. Therefore, we wish to better explain the connections[13].

We are not the first to attempt to describe and extract this interaction Nastase and Strube (2009). However, recall that they simply used Wikipedia to perform a rather limited, existential search for certain keywords, where the search was not "conditioned" on the context. Thus, no matter the context, these specific featuers for a PMP would always be the same. We believe that the context should be provided and used in order to obtain a better explanation. This section describes the novel use of ConceptNet to extract underlying conceptual meaning and connections between two words connected by a dependency relation.

6.3.1 Path Generation

If a word $w_{i,j}$ is connected to some other word y via a dependency relation, we want to determine the most plausible explanation for a connection existing between them. Given a semantic graph, the most natural thing to do is find and analyze path(s) between them. However, as described above, ConceptNet is a semantic multigraph and so relations more complex than *IsA* can be represented, which means that generally path length is no longer sufficient to determine semantic relatedness. To approach this, we determine if there is some "convenient" path — i.e., short, using a small number of relations (predicates) R and having semantically appropriate nodes — between $w_{i,j}$ and y. That is, are there a small number of conceptual relations R_l such that $R_1(w_{i,j}, x_1)$, $R_2(x_1, x_2), \ldots, R_k(x_k, y)$, where all $x_l, 1 \leq l \leq k$ are distinct (note that R_k need not be distinct)?

Although the lack of a found path may itself be indicative of a special connection, determining how easily the path lends itself to particular readings is more important[14]. Further, there is no clearly identifiable "most important" item in determining the explanatory power of a path. For instance, consider the very simple example of trying to finding out how "writing" and "novel" are connected. Three of the found paths are:

- ⟨write⟩ ConceptuallyRelatedTo ⟨novel⟩
- ⟨write⟩ HasFirstSubevent ⟨think⟩ PartOf ⟨human⟩ AtLocation ⟨book⟩ ConceptuallyRelatedTo ⟨novel⟩
- ⟨write⟩ HasFirstSubevent ⟨think⟩ Causes ⟨knowledge⟩ AtLocation ⟨book⟩ ConceptuallyRelatedTo ⟨novel⟩

The first path listed shows that sometimes short paths are weak: a "conceptually related to" relation does not really explain the connection. Therefore, we consider the length of the path, as well as the arcs/relations and the WordNet abstractions of the intermediate nodes that create the path. The WordNet abstractions are necessary for two reasons. The first is that the individual nodes can be very specific, leading to sparse data and making learning difficult. Second, we would like to get an overall sense of the meaning and interplay of the two dependency words, specifically as it relates to our predefined metonymy categories, such as people, artifact, event, product, etc.

The algorithm employed is a bi-directional, breadth-first search. However, there are numerous difficulties, such as an unpredictable branching factor and lack of available, thoroughly tested metrics or heuristics (particularly admissible heuristics). While it may seem that if a reliable heuristic or metric could be developed, then standard shortest path algorithms could be used, we note that unfortunately the intermediate nodes are not known a priori. We note that our search is concerned with both the journey (the length of the path, and the intermediate concepts and relations) as well as the destination.

To get a better idea of what is involved in this problem, we provide the following analogy. Imagine being on the quintessential post-baccalaureate back-packing trip through Europe, and you are in Barcelona and want to get to Paris. You never paid attention in geography and so know none of the intermediate cities/destinations (the intermediate nodes being unknown a priori). While you care about actually making

[13] This section can be thought of as providing a way to extract the qualia structure, mentioned earlier.

[14] In reality, yielding to computational concerns, the search must be truncated. Sometimes this affects the results (returned paths). Thus that no connectivity is found may not be as indicative as it could be, at least theoretically.

Paris → sleep: Paris sleeps. (sentence 5)
⟨paris⟩ IsA ⟨city⟩ HasA ⟨person⟩ CapableOf ⟨tire⟩ CausesDesire ⟨sleep⟩

Albania → provide: . . . to provide Albania with food aid. . . (sentence 33)
⟨albania⟩ IsA ⟨country⟩ HasA ⟨person⟩ CapableOf ⟨bring⟩ IsA ⟨provide⟩

Figure 4: Examples of extracted ConceptNet paths for schematic metonymy.

it to Paris (path existence), you are willing to explore as long as you get there in reasonable amount of time; in fact, you want to explore since that exploration is what defines the entire journey (the need to analyze the elements — the nodes and relations — of the path). You have guide books, but they are woefully out-of-date. Further you do not know either French or Spanish well. Thus asking locally for help is not a reliable solution (heuristics are lacking).

For two very small examples of schematic metonymy, consider the "toy" example 5 (page 7) and example 33:

(33) The G-24 group expressed readiness to provide *Albania* with food aid until the 1993 harvest, and beyond if necessary.

Since right now we do not consider what type of grammatical relation connects two words when performing the graph search, we simply represent the connections as δ(Paris, sleep) and δ(Albania, provide). Figure 4 shows some of the results obtained for searching for paths going from *Paris* to *sleep(s)*, and from *Albania* to *provide*. Note that for both examples, the paths are able to derive the anthropomorphic readings necessary to understand the connections.

However, there are limitations and problems with ConceptNet and as discussed in Section 7, these problems can be quite severe. Most notably, though, ConceptNet stems and lemmatizes all input (at times incorrectly), so *use* $\mapsto$ ⟨us⟩. ConceptNet also does not differentiate among senses for polysemous words, even when the difference can be drastic. For instance, it will map both *aids* (as in, "The US aids foreign countries") and AIDS to ⟨aid⟩.

6.3.2 Path Analysis via WordNet Abstractions

Prior to continuing we wish to say that we will introduce two functions, s and $\mathfrak{D}$ in this section. So as not to confuse the reader, we say now that while s is used for both Model One and Model Two, the function $\mathfrak{D}$ is used only on Model One. Similarly, the next two sections (6.3.3 and 6.4) only apply to Model One.

Note that two paths may differ by only a single relation or a single word, which suggests we may be able to analyze separately the concepts (vertices) or the relations — i.e., that there is a partial independence between words and the relations. Separately analyzing the various components of a path (i.e., the nodes versus the relations) can clarify the data analysis. Given the compositionality of documents (equation (1)), here we define a word-based, path-ranking function $\mathfrak{W}$.

We define $\mathfrak{W}$ to be an $|\mathcal{L}|$-dimensional real-valued function on paths p_i. In what follows, we define a word-scoring function s, which may be thought of as a collection of part-of-speech $\mathbb{P}$-specific word-scoring functions $s_{\mathbb{P}}(w_{ij})$. Each $s_{\mathbb{P}}$ is a vector of integral values, where the k-th component of $s_{\mathbb{P}}(w_{ij})$ represents the number of senses that are categorized under the k-th abstract type in $\mathcal{L}_{\mathbb{P}}$. To fully capture the range of meanings, we also consider the derivationally-related forms of various word senses.[15] Note that the k-th component of $s_{\mathbb{P}}(w_{ij})$ will be 0 if and only if none of the senses of w_{ij} may be traced up to the k-th abstract

[15] We disagree with the argument that considering the derivationally-related forms improperly weights abstract entities. This removes the potentially troublesome aspect of lexicalizing morphology, as well as ensuring the most complete distribution for that word form, regardless of part-of-speech. Given that ConceptNet does not distinguish POS tags, this is needed.

entity in $\mathcal{L}_{\mathbb{P}}$ (we present an example shortly). Thus, we may define

$$s(w_{ij}) = \prod_{\mathbb{P}\in\mathcal{P}} s_{\mathbb{P}}(w_{ij}), \tag{3}$$

where

$$s_{\mathbb{P}} = \prod_{k=1}^{|\mathcal{L}_{\mathbb{P}}|} \left\{ \mathbb{1}_{[w_{ij}\in\mathcal{L}_{\mathbb{P}}(k)]} \right\},$$

using the indicator function

$$\mathbb{1}_C = \begin{cases} 1, & C \text{ is true,} \\ 0, & C \text{ is false.} \end{cases}$$

Note that we may think of $s_{\mathbb{P}}(w_{ij})$ as providing a count of how many senses map to a given abstract entity; it does not attempt to indicate which abstraction is more likely to occur. Thus the word-scoring function indicates the existence distribution of the abstract entities, rather than the use distribution. We could extend these definitions to obtain proper distributions but as will be discussed, there may be a benefit to disregarding probabilistic information.

As a brief example, consider the word *knife* in some path p_i. In WordNet 3.0, *knife* is listed as both a NOUN and a VERB. The first and second senses of *knife*.NOUN (including derivationally related forms) may be traced to NOUN.ARTIFACT and VERB.CONTACT, and the third to NOUN.SHAPE; the sole sense of *knife*.VERB may be traced to VERB.CONTACT (with a related NOUN.ARTIFACT. Therefore there are three counts for NOUN.ARTIFACT, one for NOUN.SHAPE and three for VERB.CONTACT. Given the predefined ordering over $\mathcal{L}_{\mathbb{P}}$ from Table 2, we have

$$\begin{aligned} s_{\text{NOUN}} &= (0,0,0,3,\ldots,0,1,0,0,\ldots,0) \in \mathbb{R}^{|\mathcal{L}_{\text{NOUN}}|} \\ s_{\text{VERB}} &= (0,\ldots,0,3,0,\ldots,0) \in \mathbb{R}^{|\mathcal{L}_{\text{VERB}}|} \\ s_{\text{ADJ}} &= (0,0,\ldots,0) \in \mathbb{R}^{|\mathcal{L}_{\text{ADJ}}|} \\ s_{\text{ADV}} &= (0,0,\ldots,0) \in \mathbb{R}^{|\mathcal{L}_{\text{ADV}}|}. \end{aligned}$$

Then

$$\begin{aligned} s(\text{knife}) &= (s_{\text{NOUN}}, s_{\text{VERB}}, s_{\text{ADJ}}, s_{\text{ADV}}) \\ &= (\overbrace{0,0,0,3,\ldots,0,1,0,0,\ldots,0}^{\text{NOUN}}, \overbrace{0,\ldots,0,3,0,\ldots,0}^{\text{VERB}}, \overbrace{0,0,\ldots,0}^{\text{ADJ}}, \overbrace{0,0,\ldots,0}^{\text{ADV}}). \end{aligned}$$

Note that unless dimensionality reduction is done, $\mathfrak{W}$ is an $|\mathcal{L}|$-dimensional vector. Attempting to map this vector to a single number manually, as opposed to learning a mapping, is error-prone and runs the high risk of improperly losing information. The defining issue in word-sense disambiguation is the selection of the correct sense for analysis. As has been mentioned, some work has tried to circumvent this issue by choosing the first sense, on the assumption that the first sense is the most common and so it will be the most likely to be conceptualized in metonymic usage. While this may simplify algorithms, it could do so at a price, as even for schematic metonymy this could pose a problem: the underlying conceptual reasoning may not be the most commonly used (recall that schematic metonymy is standard in the categorical realm — e.g., PLACE-FOR-PEOPLE — and in the ability for the sample to be productive). Further, although the PMP (either terminal node of p_i) is a noun-phrase, there is no guarantee that any of the other words in the path will be in the form of a noun (or noun-phrase). Because the number of abstract entity categories is not constant between part-of-speech tags (i.e., $|\mathcal{L}_{\text{NOUN}}| \neq |\mathcal{L}_{\text{VERB}}|$) creating $\mathfrak{W}$ to be well-defined across part-of-speech tags could pose a problem if we do not consider all tags.

Recall from equation (1) the compositional structure of our documents $\mathcal{D}$ (repeated here for convenience):

$$\mathcal{D} \ni \mathcal{S} \ni \delta(w_{i,j}, w_{i,k}).$$

Thus while technically $\mathcal{D}$ is comprised of sentences (which in turn are comprised of dependency relations), for computation issues, we only consider the sentence containing the PMP, and so are able to remove the

mention of sentence in these algorithms. We note that adding multiple sentences is not necessarily trivial, as the further one goes from the PMP, the more likely it will be that any information contained in those sentences could create confusion, and so should be correspondingly weighted less heavily than information obtained from words near the PMP.

Because restrictions such as "consider only those paths where the PMP is a terminal node" are easy to incorporate, we will present the general algorithm that considers all paths. In our experiments we found that the PMP-restricted version performed better than the general version, but due to serious noise issues in ConceptNet, are not convinced that the results are truly indicative of the strength of each algorithm.

We consider a sample document $\mathcal{D}$, which is comprised of dependency relations δ. For every dependency relation δ we extract k paths p, where each path p has $|p|+1$ words w. Unless otherwise stated, we may assume that a total document score can be obtained by simply summing over all path-scores:

$$\mathfrak{D}(\mathcal{D}) = \sum_{\delta \in \mathcal{D}} \sum_{p \in \delta} \mathfrak{W}(p).$$

This path-scoring function $\mathfrak{W}$ is defined to be linear with respect to the composite words w_{ij} in path p_i; that is,

$$\mathfrak{W}(p) = \sum_{w \in p} s(w).$$

In this setting, although a word w may appear multiple times, every word in a path gets equal weight. Thus, naturally, words that occur more frequently in paths will provide stronger evidence for a particular abstract entity as being an underlying "theme" through the document. We may extend the above by attaching a weight function per word:

$$\mathfrak{W}(p) = \sum_{w \in p} \mathcal{N}(w) s(w),$$

where $\mathcal{N}(w) \in \mathbb{R}$. As a first attempt, we let $\mathcal{N}_1(w)$ be the number of times w is a member of any path in a document; that is,

$$\mathcal{N}_1(w) = \sum_{\delta \in \mathcal{D}} \sum_{p \in \delta} \sum_{x \in p} \mathbb{1}_{[x=w]}.$$

Using this definition will clearly make some abstract entities more prominent. Although there is a great possibility that this definition will exacerbate the "hub" problem, it may have the benefit of highlighting the (select few) abstract entities with the greatest number of occurrences. Alternatively, we may define $\mathcal{N}_2$ to be the number of times w appears in some path for different dependencies (though not counting multiple occurrences):

$$\mathcal{N}_2(w) = \sum_{\delta \in \mathcal{D}} \mathbb{1}_{[\exists p \in \delta . w \in p]}.$$

This $\mathcal{N}_2$ measures how important a word is to any underlying semantic relationships in a document. Although this can be considered the analogue of term frequency, we choose not to normalize by the number of dependencies in the document, although that is always an option.

While $\mathcal{N}_2$ widens any score differences (i.e., leading entities increase their lead), we may also wish to level the field by equalizing the score vector. This leads to

$$\mathcal{N}_3(w) = \frac{1}{\mathcal{N}_2(w)}.$$

Unfortunately, none of these word-weighting functions $\mathcal{N}_i$ performs as well as the simplest,

$$\mathcal{N}_0(w) = 1.$$

Using this function, the general form our scoring function is

$$\mathfrak{D}(\mathcal{D}) = \sum_{\delta \in \mathcal{D}} \sum_{p \in \delta} \sum_{w \in p} s(w). \tag{4}$$

We found that using all of the dependencies $\delta(\cdot, \cdot)$ for a given document was too noisy. We thus restricted score extraction to those dependencies where one, and only one, of the words was the PMP.

6.3.3 Subpath Analysis

This section only applies to Model One.

As noted above, the extracted paths contained (partially) independent components: the nodes and the relationships. The abstracting/scoring functions $s(\cdot)$ and $\mathfrak{D}$ represent the types of entities that are considered in connecting, conceptual paths. However, analyzing how various edges (relationships) interact with their nodes should also yield useful information. For instance, consider sentence 33, with the extracted path

$$\langle \text{albania} \rangle \text{ IsA } \langle \text{country} \rangle \text{ HasA } \langle \text{person} \rangle \text{ CapableOf } \langle \text{bring} \rangle \text{ IsA } \langle \text{provide} \rangle \text{ .}$$

Although it is encouraging that the underlying conceptual nodes are heavily skewed towards representing people (through both the "person" and "bring" concepts), what is truly promising is the way that they interact. Thus a subpath analysis shows that a person is capable of doing something that is more-or-less synonymous with a word for which there exists a linguistic relationship with the PMP.

We continue to restrict ourself to dealing only with those relations $\delta(y, \texttt{X})$ or $\delta(\texttt{X}, y)$, where the PMP is X. Thus, we would like to determine some basic things about the paths (that are important to the PMP, though in this case is all of them): does a person desire (in some manner) y? Is a country located in, or more generally is in some spatial relationship with, y? Is a person capable of (doing) y? Do a lot of the relations in the path indicate a physical location status?

Note that as each connecting edge in the path is supposed to use a simple predicate structure, performing this type of subpath analysis can be considered employing a variant of selectional preferences. Unfortunately, given the current state of ConceptNet (most notably, the unpredictable branching factor and general sparsity), creating "literal constraint" rules that have sufficient coverage can be quite involved. This is because there are only a small number of predefined, "supported" relations, and they are not always sufficent for capturing verbal predicates. In order to capture these predicate meanings, an involved analysis of the paths (and relations and nodes) must be done. The above sub-path analysis provides a good first-approximation to solving the problem.

6.4 Measuring Semantic Relatedness

This section only applies to Model One.

We also attempt to measure conceptual similarity via internet searches. In general, we have search queries x and y, and $f(\cdot)$, which is the number of pages returned by searching its argument. Thus $f(x)$ represents the number of results returned from a search of x, $f(x, y)$ represents the number of results returned from a search of xy (searching x and y simultaneously).[16]

For a PMP X, we may define the context words $\mathcal{C}_k$ of X by equation (2), which is repeated below:

$$\mathcal{C}_k = \{w \mid \delta(w, \texttt{X}) \text{ or } \delta(\texttt{X}, w)\} \cup \{w_{i,m} \mid |m - j| < k\};$$

that is, all dependency-related words, with all words within k of X. Similarly to Li (2008), Li and Sporleder (2010), we would like to determine the semantic cohesion of the group of words that most likely influence how the PMP should be read. In a way, this can be seen as appealing to the internet to extract implied selection preferences.

While Li and Sporleder (2010) talk about using web similarity results to help with metonymy classification (within the more general context of figurative language), there is no specific, individualized examination of web similarity on metonymy. Thus, we adopt many of the same features they do, such as the average similarity between the PMP and a context word,

$$x_1 = \frac{1}{|\mathcal{C}_k|} \sum_{c \in \mathcal{C}_k} \text{sim}(\texttt{X}, c),$$

the average similarity between pairwise distinct context words,

$$x_2 = \frac{2}{\binom{|\mathcal{C}_k|}{2}} \sum_{(c_1, c_2) \in \mathcal{C}_k \times \mathcal{C}_k c_1 \neq c_2} \text{sim}(c_1, c_2)$$

[16] While mathematically we are not entirely comfortable with overloading f in such a way, we use that notation to remain consistent with past literature and work.

	Dec. Tree	Random Forest	7-NN	NB	MLP	SVM
LOCATION: coarse						
accuracy	82.3%	81.9%	81.5%	67.0%	82.1%	83.4%
literal	0.894	0.893	0.890	0.780	0.894	0.899
non-literal	0.452	0.424	0.423	0.410	0.425	0.540

Table 4: Accuracy and F_1 Scores of PMP-targeted modeling using different classifiers on the entire dataset.

and relationships between x_1 and x_2,

$$x_3 = x_1 - x_2.$$

Knowing and testing cut-points for x_3 is also done. Notice how x_3 can provide a grounding: if the context words are all pairwise very similar, but when taken with the PMP have a low yield, then adapting past work, metonymy could be taking place.

To actually obtain the above features, we make use of the Normalized Google Distance (NGD) as presented in Cilibrasi and Vitanyi (2007). The NGD is designed to measure the semantic relatedness of two phrases x and y by

$$NGD(x,y) = \frac{\max\{\log f(x), \log f(y)\} - \log f(x,y)}{\log N - \min\{\log f(x), \log f(y)\}},$$

where N represents the number of pages indexed by the search engine. To help clarify its meaning, we stress that *NGD* should be interpreted as a metric rather than as a semantic relatedness measure: terms x and y are more closely related than x and z if and only if $NGD(x,y) < NGD(x,z)$.

Despite warnings from Li (2008), we wanted to apply pointwise-mutual information to schematic metonymy classification:

$$PMI(x,y) = \log \frac{f(x,y)}{f(x)f(y)}.$$

This turned out to be fruitless, as *PMI* was no more discriminative than *NGD*.

7 Results and Evaluation of Features

Here we present the results from applying our methods to the SemEval 2007 metonymy dataset. We will call the PMP-targeted the first model ("Model One") and we will call the hCRF the second model ("Model Two"). Note that other researchers working on this dataset have typically reported results from k-fold cross-validation. This has been primarily due to the fact that the testing set is generally considered much more difficult than the training. Our own internal experiments verified this. Therefore, unless otherwise stated, all of our results are using 10-fold stratified cross-validation.

7.1 Evaluation on the Dataset

In testing our first model, one of the things we wanted to do was determine what, if anything, the success, or failure, of any particular learning algorithm would tell us. Thus, we tested Model One on a number of Weka classifiers. These ranged from tree-based approaches (a standard decision tree as well as a random forest Breiman (2001)); more memory-based approaches (k-nearest neighbor), emulating Peirsman (2006); simple probabilistic classifiers (Naïve Bayes); non-linear discriminative classifiers (multilayer perceptron — MLP); and a kernel-based discriminative classifiers (support vector machine), as in Nastase and Strube (2009). Normally on this dataset performance does not degrade horrendously when changing task granularity; so as to prevent an overload of information, Table 4 contains only the results of these classifiers in the location/country domain with coarse granularity.

As can be seen, the SVM clearly provides the most predictive power. It is disappointing that neither the decision tree nor the random forest performed all that well. However, when we try and glean information from

		SemEval				
	Granularity	**Baseline**	**Avg.**	**Best**	**Nastase et al.**	**Model One**
LOC.	Coarse	79.4%	81.5%	85.2%	86.1%	83.4%
	Medium	79.4%	81.2%	84.8%	85.9%	82.5%
	Fine	79.4%	80.1%	84.4%	85.0%	82.0%
ORG.	Coarse	61.8%	74.6%	76.7%	74.9%	75.5%
	Medium	61.8%	71.8%	73.3%	72.4%	69.9%
	Fine	61.8%	71.3%	72.8%	71.0%	69.0%

Table 5: Accuracy Results of PMP-targeted modeling on the entire dataset.

the actual structure of the decision tree itself, we see that while the abstract entity-type of the dependency-related words were influential, certain coordinates of $\mathfrak{D}(\mathcal{D})$ were also useful. Most notably, the decision tree was able to discriminate samples based in part on how much the underlying conceptual paths related to communication, a procedule description of something, and the embodying of an entity. Unfortunately, given its poor performance, it does not provide a reasonable approximation for how hand-crafted rules would fair.

Further, the k-NN results indicate that the values for our features do not lend themselves to memorization. Unfortunately, as we will see throughout and later on, this is most likely due to unacceptable, distracting noise. However, as the SVM demonstrates, the data do not seem to be inseparable. In fact, as will be seen, the features can be extraordinarily descriptive.

As the SVM performed the best out of all tested classifiers, we use only an SVM as a classifier in our full Model One. We compare accuracy and F_1 scores to the "most-frequent" baseline (choose the most frequent label), the average and best results from the SemEval competition and the results from Nastase and Strube (2009).

Although we present data for both location and organization, there are certain key differences between the two domains. First, while five participants competed in location domain in the SemEval competition, only the top three systems (by self-selection) competed in the organization task. Since they were the high performing systems, we can expect the average to be proportionately higher than for the location domain.

However, our own evaluations on the organization data deserve note: due to a lack of data features for a majority of the organization-based samples, our algorithm purposefully chose not to even consider classifying them. While this leads to an extremely low coverage score, upon closer analysis of the samples for which enough data could be collected, we see that the distribution of literal readings is close to the distribution for all organization-based samples ($\sim 62\%$ of the full distribution are labeled as literals, while $\sim 65\%$ of our self-restricted subset are labeled as literals).

In Table 5, we present accuracy scores for the above mentioned baselines/systems, while in Table 6 we present F_1. When examining the accuracy, we see that all of our results are above the baseline. However, once we consider the distribution of labels, particularly for the location category, the results, both ours and others', indicate that much more work needs to be done. This is conclusively demonstrated by the F_1 scores.

To analyze our results, it helps to distinguish between the location domain and the organization domain. In the location-domain, our results were lower than the previous best systems, but in the organization-domain, our Model One was competitive. First, let us analyze the location-domain data.

Across granularities, when there was a sufficient number of training examples for a given class, Model One was able to classify better than the average SemEval system, though not better than the best systems for this dataset. Of all systems to have used this dataset (to our knowledge), there have been four that outperformed Model One. Of these four however, it is worthwhile to note that one (Brun et al. 2007) adapted and modified a proprietary parser; while they describe their distributional approach, the description of the changes does not easily permit a re-implementation of their methods[17]. Meanwhile, the other three systems that outperformed Model One may have extracted some complex and discriminating features (Farkas et al.

[17]This is not meant to be accusatory in any way; rather just to note that emulating and recreating their results was very difficult.

	SemEval			
	Avg.	**Best**	**Nastase et al.**	**Model One**
LOCATION: coarse				
literal	0.888	0.912	0.916	0.899
non-literal	0.472	0.576	0.591	0.540
LOCATION: medium				
literal	0.889	0.912	0.916	0.897
metonymic	0.476	0.580	0.615	0.530
mixed	0.017	.083	.160	0.000
LOCATION: fine†				
literal	0.887	0.912	0.916	0.896
place-for-people	0.456	0.589	0.617	0.525
ORGANIZATION*: coarse				
literal	0.746	0.767	0.814	0.809
non-literal	0.615	0.652	0.616	0.658
ORGANIZATION*: medium				
literal	0.814	0.825	0.814	0.795
metonymic	0.577	0.604	0.587	0.577
mixed	0.163	0.308	0.268	0.000
ORGANIZATION*: fine				
literal	0.817	0.826	0.814	0.803
org-for-members	0.608	0.630	0.597	0.634
org-for-event	0.000	0.000	0.000	0.000
org-for-product	0.458	0.500	0.444	0.208
org-for-facility	0.141	0.222	0.381	0.000
org-for-index	0.000	0.000	0.000	0.000
obj-for-name	0.592	0.800	0.588	0.000
obj-for-rep	0.000	0.000	0.000	0.000
othermet	0.000	0.000	0.000	0.000
mixed	0.135	0.343	0.293	0.143

Table 6: F_1 Scores of PMP-targeted modeling on the entire dataset. (*) A subset of the ORG. data was used for testing. The label-distribution of the subset was very similar to that of the full dataset (around 65% literal reading in the subset, compared to around 62% for the full dataset. (†) Because the classifier turned this multiclass problem into a binary decision problem, we only list non-zero classifications. That is, if a class is not listed, though it should be, then that means that the classifier did not consider it.

		SemEval			
	Measure	**Avg.**	**Best**	**Nastase et al.**	**Model Two**
LOC. — coarse	literal	0.888	0.912	0.916	0.877
	non-lit	0.472	0.576	0.591	0.414
ORG. — coarse	literal	0.746	0.767	0.814	0.718
	non-lit	0.615	0.652	0.616	0.393

Table 7: F_1 Results of hCRF modeling on the entire dataset. Unlike Model One, we were able to use all samples in both domains.

2007, Nastase and Strube 2009, Nicolae et al. 2007), but they required the use of the manually constructed grammatical annotations. Our method, on the other hand, was automatic.[18]

Given the above, the lower performance of Model One is possibly understandable and even acceptable. However, we also experimented with using the manual annotations to obtain our features, and the results changed minimally. That is, extracting the same feature, but from a different starting point, resulted in rather insignificant changes. For instance, on the coarse setting, the non-literal F_1 score was 0.540 for our automatic system, and 0.535 for our system with the manual seed data. In the medium granularity though, our metonymic F_1 score was 0.521 for our automated system, and 0.525 for the manual system. The results are similar across the location domain and granularities. Since the differences would be minimal, we only report results for the automatic system. The minimal changes indicate that our methods are somewhat robust to automatic, rather than manual, input; this constrasts with other researchers, notably Nastase and Strube (2009), who report that automating their methods (i.e., automatically extracting dependency relations rather than relying on the manually-provided ones) produced a significant degradation in their performance.[19] In the next section, we briely summarize the strengths and weaknesses of our methods.

Shifting to the organization-domain, we must obviously preface and temper any remarks on our results with the caveat that we used a signficantly reduced dataset, even though the distribution of labels was nearly the same as for the entire dataset. However, this reduced dataset is actually useful for analyzing our methods, particularly the graph search. First, we note that using an instance-based classifier (memory-based learner), Peirsman (2006) found that training on significantly reduced datasets did not translate to a proportional performance reduction. Therefore, the results should not be automatically discounted.

Recall that one of the reasons we used the reduced dataset is because it was only for that subset of documents that our graph-search method was able to actually find paths that could then be analyzed. While we could have tested on the entire dataset (and put dummy-values in for the path features), we decided to use this as an opportunity to informally examine how well our method works when there is less noise[20] Thus, for the organization domain, when it actually found concepts and relations, they tended to be more accurate. Rather than turn the organization domain into examining how well the ConceptNet features help the standard syntactic and semantic features, we were actually better able to test informally the converse of that. The competitiveness of Model One indicates that our method is promising; this point is further considered in Section 7.2.

Unfortunately, the hidden conditional random field was not nearly as successful as we hoped it would be. Due to its rather poor performance on the coarse granularities, we determined that analyzing it in the medium and fine situations would not be worthwhile. Please see Table 7 for results on the location and organization domain. The results are unfortunately below the baselines.

Recall that one of the things we wanted to test was whether or not we could implicitly model collocation

[18] Not including the potentially noisy ConceptNet data, we were able to more-or-less recreate previous SemEval competition results; that is, generally achieve performance within 1.5 of previous results. Note that we did not engineer some of the very specific features from other researchers; this could very well account for the slightly lower performance.

[19] However, we do need to remind the reader that this work has been done two years later than the work of Nastase and Strube (2009). So although they used the Stanford paraser, it is possible that in that time the parser has been improved significantly.

[20] As briefly discussed above, the extracted data for the location domain is very noisy; however, that typically was a result of there being too much information improperly stored in ConceptNet. As the number of assertions that a concept has decreases, we found that generally, the precision and correctness of that information increases.

data by representing it in a chain-like fashion. The hidden variables were an attempt to learn any potential hidden structure in the data, both the syntactically- and semantically-based. Moreover, the hidden variables should have allowed us to use a CRF to exploit its tagging powers, but to do so in a way as to only care about an entire document, rather than the labels for specific words. However, one of the issues we encountered is that there was no consistent number of hidden variables that would produce results. Using say, ten hidden variables for coarse location-based metonymy produced classifications, but ten hidden variables were not necessarily appropriate for coarse organization-based metonymy. This made creating a general model extremely difficult.

7.2 Small-Scale ConceptNet Evaluation

We have mentioned above, and throughout this work, that ConceptNet can suffer from a combination of low recall, low precision, and high noise due to users errors and infidelities or errors in the automated process that populates the database. However, as our Model One results from both the location and organization domains show, there is useful information to be extracted from ConceptNet. This point is further emphasized by the preliminary results illustrated in Section 6.3. In this section, we wanted to briefly expand upon those results, and illustrate how there are some signficant issues involved in working with ConceptNet.

First, recall that ConceptNet stems and lemmatizes all input. Thus, while *aids*, and *aided* map to ⟨aid⟩, so does *AIDS*. Therefore, it was not surprising to see instances of one country (for instance, France) aiding another, but have the underlying conceptual mapping be given as

⟨france⟩ IsA ⟨country⟩ HasA ⟨cow⟩ HasA ⟨sex⟩ Causes ⟨aid⟩.

However, there are also some very idiosyncratic entries. For instance, according to ConceptNet, *war* is an *elephant*. While it is possible someone actually tried to metaphorically say that war is an elephant, it is much more likely that the input was something like, "War was the elephant in the room at the press briefing" (though one could also be talking about a war elephant).

Despite these problems, we have seen that useful information can be extracted. Repeating and expanding upon the information from Figure 4, we have

Paris → sleep: Paris sleeps. (sentence 5)

⟨paris⟩ IsA ⟨city⟩ HasA ⟨person⟩ CapableOf ⟨tire⟩ CausesDesire ⟨sleep⟩

Albania → provide: . . . to provide Albania with food aid. . . (sentence 33)

⟨albania⟩ IsA ⟨country⟩ HasA ⟨person⟩ CapableOf ⟨bring⟩ IsA ⟨provide⟩

Iraq → comply: . . . Demands that Iraq comply fully . . .

⟨iraq⟩ IsA ⟨place⟩ HasProperty ⟨something⟩ CapableOf ⟨choose⟩ ConceptuallyRelatedTo ⟨refuse⟩ Is[not] ⟨comply⟩

Denmark → champion: Charlton believes that European Champions Denmark are the most likely side. . .

⟨denmark⟩ IsA ⟨country⟩ AtLocation ⟨world⟩ HasA ⟨human⟩ AtLocation ⟨war⟩ CausesDesire ⟨conquer opponent⟩ MotivatedByGoal ⟨champion⟩

The above four examples are classified as place-for-people metonymies.

We already noted that with the first two our method is able to successfully derive the anthropomorphic readings necessary to understand the connections. Note that this holds true for the latter two examples as well. It is worth noting that in the fourth example, the path correctly identifies the SPORT-AS-WAR metaphor. Further, we have already seen how a subpath analysis can be beneficial and discriminative. When we examine the results of the scoring function $\mathfrak{D}$, we see that entities relating to communication and state are consistently high. This contrasts with literally-read PMPs, where $\mathfrak{D}$ typically indicates that non-action related nouns are quite common (for instance, NOUN.LOCATION).

8 Future Work and Conclusion

In this thesis, we tested the hypothesis that a better modeling of the underlying syntactic, semantic and conceptual meanings can aid automated metonymy classification. We tested this hypothesis in two main ways: the first (Model One) used a PMP-targeted system, which grouped the features all into one set, while the second attempted to explicitly model interactions among some of the more standard syntactic and semantic features by implicitly modeling collocation data. Moreover, we wanted to make our system as automated as possible, and as invariant to automatically-generated, as opposed to manually-created, data as possible.

Although on the full location-based dataset Model One was outperformed by the systems created both for the competition and as research projects after the competition, it was able to do better than average while remaining fairly robust and invariant to the seed data provided (the manual or automatically created dependencies and annotations). Moreover, we believe issues with ConceptNet beyond our control, such as that it does not distinguish word senses for polysemous words, that it can incorrectly stem/lemmatize words, and that the automatic relation extraction process can be faulty, significantly impacted our work. We believe this in part because when the algorithms worked, the paths it found were able to capture the connections between words. In some cases, it accurately represented the anthropomorphic concepts behind the metonymy; in others,it accurately captured underlying metaphors.

There is additional evidence for the promise of our method from the evaluation on the organization data. Even though having to operate on a significantly reduced dataset (due to the graph search not finding any paths to analyze) is not ideal, it did allow us to isolate and analyze how the standard features were able to interact with the ConceptNet features, without having as much noise in the data. In that case, we found that our automated method was able to perform competitively with manually-reliant systems.

Unfortunately, our Model Two was not nearly as successful as Model One. Many things could be improved, such as the input features, since we restricted it to the standard syntactic and semantic features. Even increasing the number of features could be useful. Doing both of these would be able to more power to the hidden variables and so the model may be able to perform better.

Similarly, we could examine using topic analysis (LDA – Latent Dirichlet Allocation) to perform a "gisting" of various phrases and clauses. This "gisting" could be used to tackle the above-mentioned label problem, or as a feature in its own right. Note that many LDA topic analysis programs are trained on newswire data, so we could expect top performance.

We unfortunately abandoned a rule-based approach, primarily due to a lack of reliable feature extractors. We determined that constructing hard-coded rules with the available feature extractors would have amounted to emulating a standard machine learning algorithm, such as a decision tree. Unfortunately, the decision tree-based algorithms we tested did not find the extracted features useful. In addition to extracting better features, one promising route could be the use of Markov Logic Networks, or MLNs (Domingos and Richardson 2004). As a simple explanation, MLNs attempt to bridge the gap betweeen logic and statistics by adding weights to first-order logical formulae. MLNs could possibly provide the power of heuristics, much as how humans may interpret and disambiguate metonymy (consider the heuristic rules we examined in Section 2).

It seems very likely that we were fairly successful in achieving our goal of automation robustness. There are indications that our methods achieve a relatively high level of dataset invariance, given the examples of our method capturing underlying metaphor and anthropomorphic ideas. However, to more formally test this, we could also apply these methods to other datasets. For instance, there is a metonymy task at the SemEval 2010 workshop[21], which places an emphasis on type coercion. Unfortunately, due to time constraints, we were unable to use that dataset in this thesis. However, we could also apply our methods to datasets for other types of figurative language, such as that used in Fan et al. (2009) or Li (2008).

[21] http://sites.google.com/site/semevalasc/

Acknowledgements

I would first like to graciously and deeply thank my thesis committee, Professor Lenhart K. Schubert, Dr. Mary Swift, and Professor Daniel Gildea. For the many hours of help, whether it be through a lengthy email, a long scheduled meeting or an even longer walk-in, you have helped guide me, brainstorm ideas and provide support. Thank you for all of your input, but most importantly, for helping me and my research mature.

The faculty support was essential and tremendously helpful, but I owe a lot to my all of my friends too. They listened to my ideas and my grumblings, and provided that needed support during late nights in the lab.

I would like to thank the entire URCS staff, especially Marty Guenther, for helping make my time at Rochester so enjoyable and smooth.

I owe a lot to my family as well. They were always there for me, willing to talk no matter the hour. Their constant encouragement, support and love were truly helpful.

And finally, to everyone who has listened to my ideas, I truly appreciate the feedback, comments, concerns and suggestions.

Thank you everyone.

References

Walter Bisang, Hans Henrich Hock, and Werner Winter, editors. *Trends in Linguistics, Studies and Monographs*. Mouton de Gruyter, March 2006.

Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

Caroline Brun, Maud Ehrmann, and Guillaume Jacquet. Xrce-m: a hybrid system for named entity metonymy resolution. In *SemEval '07: Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 488–491, Morristown, NJ, USA, 2007. Association for Computational Linguistics.

Rudi L. Cilibrasi and Paul M.B. Vitanyi. The google similarity distance. *IEEE Trans. Knowledge and Data Engineering*, 19(3), 2007.

Alice Deignan. *Researching and Applying Metaphor*, chapter Corpus-Based Research into Metaphor, pages 177–199. Cambridge University Press, Cambridge, 1999. Lynne Cameron and Graham Low (eds.).

Alice Deignan. *Trends in Linguistics, Studies and Monographs*, chapter The Grammar of Linguistic Metaphors. Mouton de Gruyter, March 2006.

Pedro Domingos and Matthew Richardson. Markov logic: A unifying framework for statistical relational learning. In *PROCEEDINGS OF THE ICML-2004 WORKSHOP ON STATISTICAL RELATIONAL LEARNING AND ITS CONNECTIONS TO OTHER FIELDS*, pages 49–54, 2004.

James Fan, Ken Barker, and Bruce Porter. Automatic interpretation of loosely encoded input. *Artif. Intell.*, 173:197–220, February 2009.

Richárd Farkas, Eszter Simon, György Szarvas, and Dániel Varga. Gyder: maxent metonymy resolution. In *SemEval '07: Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 161–164, Morristown, NJ, USA, 2007. Association for Computational Linguistics.

Dan Fass. Metonymy and metaphor: what's the difference? In *Proceedings of the 12th conference on Computational linguistics*, pages 177–181, Morristown, NJ, USA, 1988. Association for Computational Linguistics. ISBN 963 8431 56 3. doi: http://dx.doi.org/10.3115/991635.991671.

Dan C. Fass. *Processing Metonymy and Metaphor*. Ablex, Greenwich, CT, 1997.

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

Christiane Fellbaum. *WordNet: An Electronic Lexical Database*. Bradford Books, 1998.

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics*, pages 363–370. ACL, 2005.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The WEKA data mining software: An update. *SIGKDD Explorations*, 11(1), 2009.

Z. Harris. Distributional structure. *Word*, 10(23):146–162, 1954.

Martin Hilpert. *Trends in Linguistics, Studies and Monographs*, chapter Keeping an Eye on the Data: Metonymies and Their Patterns. Mouton de Gruyter, March 2006.

Jerry Hobbs, Mark Stickel, Douglas Appelt, and Paul Martin. Interpretation as abduction. *Artificial Intelligence*, 63(1-2):69–142, 1993.

Shin-ichiro Kamei and Takahiro Wakao. Metonymy: Reassessment, survey of acceptability, and its treatment in a machine translation system. In *Proceedings of the 30th Annual Meeting of the Association for Computational Linguists*, pages 309–311, 1992.

Karin Kipper, Hoa Trang Dang, and Martha Palmer. Class-based construction of a verb lexicon. AAAI, 2000.

Karin Kipper, Anna Korhonen, Neville Ryant, and Martha Palmer. Extending verbnet with novel verb classes. In *Fifth International Conference on Language Resources and Evaluation (LREC 2006)*, June 2006.

Hedvig Kjellström, Javier Romero, David Martínez, and Danica Kragić. Simultaneous visual recognition of manipulation actions and manipulated objects. In *Computer Vision ECCV 2008*, volume 5303, pages 336–349, 2008.

Dan Klein and Christopher D. Manning. Conditional structure versus conditional estimation in nlp models. In *2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, pages 9–16, 2002.

Dan Klein and Christopher D. Manning. Accurate unlexicalized parsing. In *Proceedings of the 41st Meeting of the Association for Computational Linguistics*, pages 423–430, 2003a.

Dan Klein and Christopher D. Manning. *Advances in Neural Information Processing Systems 15 (NIPS 2002)*, chapter Fast Exact Inference with a Factored Model for Natural Language Parsing, pages 3–10. MIT Press, 2003b.

Saisuresh Krishnakumaran and Xiaojin Zhu. Hunting elusive metaphors using lexical resources. In *FigLanguages '07: Proceedings of the Workshop on Computational Approaches to Figurative Language*, pages 13–20, Morristown, NJ, USA, 2007. Association for Computational Linguistics.

John Lafferty, Andrew McCallum, and Fernando Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *International Conference on Machine Learning (ICML)*, 2001.

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

M. Lapata and A. Lascarides. A probabilistic account of logical metonymy. *Computational Linguistics*, 29 (2):261–315, 2003.

Johannes Leveling. Fuh (fernuniversität in hagen): metonymy recognition using different kinds of context for a memory-based learner. In *SemEval '07: Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 153–156, Morristown, NJ, USA, 2007. Association for Computational Linguistics.

Linlin Li. A cohesion-based approach for unsupervised recognition of literal and nonliteral use of multiword expressions. Master's thesis, Universität des Saarlandes, 2008.

Linlin Li and Caroline Sporleder. Using gaussian mixture models to detect figurative language in context. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 297–300, 2010.

H. Liu and P. Singh. Conceptnet: A practical commonsense reasoning toolkit. *BT Technology Journal*, 22, 2004.

Katja Markert and Udo Hahn. Metonymies in discourse. *Artificial Intelligence*, 135(1):296–304, 2002.

Katja Markert and Malvina Nissim. Annotation Scheme for Metonymies (AS1). http://nlp.cs.swarthmore.edu/semeval/tasks/task08/summary.shtml, June 7, 2005.

Katja Markert and Malvina Nissim. Metonymy resolution as a classification task. In *EMNLP '02: Proceedings of the ACL-02 conference on Empirical methods in natural language processing*, pages 204–213, Morristown, NJ, USA, 2002. Association for Computational Linguistics. doi: http://dx.doi.org/10.3115/1118693.1118720.

Katja Markert and Malvina Nissim. Semeval-2007 task 08: metonymy resolution at semeval-2007. In *SemEval '07: Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 36–41, Morristown, NJ, USA, 2007. Association for Computational Linguistics.

J Martin. Metabank: A knowledge base of metaphoric language conventions. *Computational Intelligence*, 10(2):134–149, 1994.

Andrew McCallum, Khashayar Rohanimanesh, and Charles Sutton. Dynamic conditional random fields for jointly labeling multiple sequences. In *Workshop on Syntax, Semantics, Statistics; 16th Annual Conference on Neural Information Processing Systems (NIPS 2003)*, 2004.

Vivi Nastase and Michael Strube. Combining collocations, lexical and encyclopedic knowledge for metonymy resolution. In *EMNLP '09: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 910–918, Morristown, NJ, USA, 2009. Association for Computational Linguistics. ISBN 978-1-932432-62-6.

Cristina Nicolae, Gabriel Nicolae, and Sanda Harabagiu. Utd-hlt-cg: semantic architecture for metonymy resolution and classification of nominal relations. In *SemEval '07: Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 454–459, Morristown, NJ, USA, 2007. Association for Computational Linguistics.

Malvina Nissim and Katja Markert. Syntactic features and word similarity for supervised metonymy resolution. In *ACL '03: Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, pages 56–63, Morristown, NJ, USA, 2003. Association for Computational Linguistics. doi: http://dx.doi.org/10.3115/1075096.1075104.

Malvina Nissim and Katja Markert. Learning to buy a Renault and talk to BMW: A supervised approach to conventional metonymy. In *International Workshop on Computational Semantics (IWCS2005)*, 2005.

Boyan Onyshkevych. Nominal metonymy processing. In *ACL '98: The Computational Treatment of Nominals*, 1998.

Yves Peirsman. Example-based metonymy recognition for proper nouns. In *EACL '06: Proceedings of the Eleventh Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 71–78, Morristown, NJ, USA, 2006. Association for Computational Linguistics.

Fuchun Peng and Andrew McCallum. Accurate information extraction from research papers using conditional random fields. In *Proceedings of Human Language Technology Conference and North American Chapter of the Association for Computational Linguistics (HLT/NAACL-04)*, 2004.

Thierry Poibeau. Up13: knowledge-poor methods (sometimes) perform poorly. In *SemEval '07: Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 418–421, Morristown, NJ, USA, 2007. Association for Computational Linguistics.

James Pustejovsky. The generative lexicon. *Computational Linguistics*, 17(4), 1991.

A. Quattoni, S. Wang, L. Morency, M. Collins, and T. Darrell. Hidden conditional random fields. *IEEE PAMI*, 29:1848–1852, 2007.

Ken-Ichi Seto. *Metonymy in Language and Thought*, chapter Distinguishing Metonymy from Synecdoche, pages 255–273. John Benjamins, Amsterdam, 1999. Klaus-Uwe Panther and Günther Radden (eds.).

Ekaterina Shutova. Sense-based interpretation of logical metonymy using a statistical method. In *ACL-IJCNLP '09: Proceedings of the ACL-IJCNLP 2009 Student Research Workshop*, pages 1–9, Morristown, NJ, USA, 2009. Association for Computational Linguistics.

Ekaterina Shutova and Simone Teufel. Logical metonymy: Discovering classes of meanings. In *Proceedings of the CogSci 2009 Workshop on Semantic Space Models*, 2009.

David Stallad. Two kinds of metonymy. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguists*, pages 87–94, 1993.

Alexander Statnikov, Lily Wang, and Constantin Aliferis. A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification. *BMC Bioinformatics*, 9(1):319, 2008.

Gustaf Stern. *Meaning and Change of Meaning.* Indiana University Press, Bloomington, 1965.

Sy Bor Wang, Ariadna Quattoni, Louis-Philippe Morency, and David Demirdjian. Hidden conditional random fields for gesture recognition. In *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2*, CVPR '06, pages 1521–1527, Washington, DC, USA, 2006. IEEE Computer Society. ISBN 0-7695-2597-0. doi: http://dx.doi.org/10.1109/CVPR.2006.132. URL http://dx.doi.org/10.1109/CVPR.2006.132.