Advances in Cognitive Systems 2 (2013) 297–315 Submitted 12/2012; published 12/2013

Analogical Word Sense Disambiguation

David Barbella BARBELLA@U.NORTHWESTERN.EDU
Kenneth D. Forbus FORBUS@NORTHWESTERN.EDU
Qualitative Reasoning Group, EECS Department, 2133 Sheridan Rd, Evanston IL 60208 USA

Abstract

Word sense disambiguation is an important problem in learning by reading. This paper introduces *analogical word-sense disambiguation*, which uses human-like analogical processing over structured, relational representations to perform word sense disambiguation. Cases are automatically constructed using representations produced via natural language analysis of sentences, and include both conceptual and linguistic information. Learning occurs via processing cases with the SAGE model of analogical generalization, which constructs probabilistic relational representations from cases that are sufficiently similar, but also stores outliers. Disambiguation is performed by using analogical retrieval over generalizations and stored examples to provide evidence for new word occurrences based on prior experience. We present experiments demonstrating that analogical word sense disambiguation, using representations that are suitable for learning by reading, yields accuracies comparable to traditional algorithms operating over feature-based representations.

1. Introduction

Scaling up cognitive systems to reason in a variety of domains requires extensive knowledge, but knowledge capture is a well-known bottleneck for AI. Learning by reading (Barker et al., 2007; Carlson et al., 2010; Forbus et al., 2007) represents a promising approach to solving this problem. Natural language understanding, necessary for this process, involves several well-known sub-problems. One of these is *word sense disambiguation*, i.e., choosing which meaning of a word is intended in a particular sentence. The set of meanings for a word are called its *sense inventory*. Word sense disambiguation is a difficult task; Navigli (2009) reports that, even for instances of "coarse-grained, possibly binary, sense inventories", interannotator agreement only reaches about 90 percent for human annotators, and on more fine-grained sense inventories it is between 67 percent and 80 percent.

While word sense disambiguation has been heavily explored (Navigli, 2009), current techniques have two serious limitations for learning by reading. The first is that the sense inventories used are mostly derived from lexical information (e.g., WordNet synsets) or terms from Wikipedia. They do not naturally produce or use the kind of rich conceptual information needed to support reasoning. Structured, relational representations are important to extract from language for reasoning. For example, understanding a verb in a sentence involves inferring how it connects the constituents to the event or relationship expressed by the verb. The second limitation is that most existing techniques rely on large hand-annotated corpora for training. For

example, the Brown Corpus (Francis & Kučera, 1964) contains one million words, and the Wall Street Journal Corpus (Paul & Baker, 1992) contains 30 million words. In contrast, people manage to learn to read without being handed massive amounts of input labeled in terms of their internal representations. Understanding how to learn to read without such large annotated corpora is an important problem, both scientifically and to make the construction of large-scale cognitive systems truly practical.

Understanding texts well enough to produce representations that are accurate enough for reasoning is a different problem than typically addressed by computational linguistics, which has mostly fallen back to feature-based representations. While our version of the word sense disambiguation problem is very different, we take inspiration from instance-based techniques (Daelemans, Van Den Bosch, & Zavrel, 1999), which use k-nearest neighbor over feature-based representations as an approximation to similarity. Our system produces relational representations during language understanding, which suggests that using human-like analogical processing techniques could be a useful approach to disambiguation in learning by reading.

This paper describes a technique for word sense disambiguation based on human-like analogical processing. The approach supports incremental learning. It assumes a natural language system that produces as output structured, relational representations that combine linguistic and conceptual information. Cases for each word sense used in an understood sentence are constructed automatically from the relevant aspects of the understanding of the sentence. These cases are accumulated and either stored or generalized for future use. Given a new sentence, the system uses analogical retrieval of cases or generalizations of cases to provide evidence about how to disambiguate word senses, based on a partial analysis of the new sentence by the language system. To prime the pump, we assume that there are other sources of evidence for disambiguation available, such as top-down abduction (Tomai & Forbus, 2009) or a small set of user-labeled examples. This research uses the Companion cognitive architecture (Forbus et al., 2009), but we believe our approach will be useful for any learning by reading system that produces structured representations to support reasoning.

We begin by summarizing relevant background, including the context motivating this work and the representations and natural language system used. Then we describe our technique in detail. Our hypothesis is that analogical processing can provide accuracy on word sense disambiguation tasks involving relational representations that is comparable to existing techniques operating over simpler, feature-based representations, and that it can learn incrementally and rapidly. After this, we describe two experiments that provide evidence for these hypotheses. Finally, we close with related and future work.

2. Background

The knowledge base contents we use come from ResearchCyc (www.research.cyc.com), with our own extensions for analogical processing. ResearchCyc is a publicly available large-scale knowledge base of over two million formally represented facts. Its ontology incorporates over 58,000 concepts, such as `FamousPerson` and `Star`, organized hierarchically (e.g., `Person` is a superordinate concept of `FamousPerson`). There are over 8,000 relations and logical functions, with their own hierarchical organization. The knowledge is contextualized into *microtheories*, collections of facts organized hierarchically. This supports representing and reasoning about multiple worlds (e.g. counterfactuals, fiction, alternate interpretations). In our

reasoning system, cases are implemented as microtheories. This rich stock of conceptual knowledge has been demonstrated to be useful in prior learning by reading experiments (Curtis et al., 2009; Forbus et al., 2007). For example, Lockwood and Forbus (2009) demonstrated that a learning by reading system, albeit with manual disambiguation, could learn enough from a simplified English chapter (including sketched diagrams) in a Navy manual to do well answering questions from the back of the chapter. The disambiguation system described here is being developed as part of a next-generation learning by reading system in which disambiguation is totally automatic.

ResearchCyc also incorporates substantial lexical knowledge, including formalizations of words (e.g., `Star-TheWord`), denotation information (i.e., senses of `Star-TheWord` include the KB concepts `Star` and `FamousPerson`), and semantic frame translations for verbs into the underlying conceptual representations. This lexical knowledge provides the representations for word senses. Consider the sentence "He is known as the star of the ABC sitcom and as a host of *It's Showtime at the Apollo*." During language processing, the language system represents the particular occurrence of `Star-TheWord` as an automatically generated discourse variable (e.g., `star-2358`). It also generates a semantic choice set to represent its possible meanings. For example `(isa star-2358 FamousHuman)` and `(isa star-2358 Star)` are choices in this semantic choice set. The representations used for verbs are generally more complex because they connect multiple constituents referred to in a sentence. ResearchCyc mostly uses Davidsonian representations for verbs. In these, the event denoted by the verb is reified and other facts connect that event to the actors and entities that it relates. For example, the word "entertained" in the sentence "He emerged as a league star and his scoring entertained crowds" produces the interpretation `(and (isa entertain-2496 EntertainmentEvent) (spectators entertain-2496 crowd-2587) (doneBy entertain-2496 score-2435))`, which not only includes the intended sense of the word, but also role relations indicating the actors involved.

Our natural language system uses these frame representations to encode the links between the different entities in the sentence. A few nouns also have more complex representations; for example, one possible choice for "temperature" in "The heat flows to the object with the lower temperature" is `(and (temperatureOfObject object-22573 temperature-22651) (isa temperature-22651 Temperature))`, which not only encodes that `temperature-22651` is a temperature, but also that it has a specific relationship with the object – it is the temperature of that object. For prepositions, the sense is largely bound up in the predicate of the statement of the interpretation. For example, in the sentence "It is the second brightest star in the northern celestial hemisphere", the intended interpretation for "in" is `(in-UnderspecifiedContainer star-5486 hemisphere-5560)`. As the system is also deciding among choices with the same word sense but different semantics at the sentence level, our task is somewhat harder than traditionally-defined word sense disambiguation.

We use the Explanation Agent Natural Language Understanding system (Tomai & Forbus 2009) for natural language understanding, which is built into the Companion cognitive architecture. EA NLU uses Allen's (1994) bottom-up chart parser for handling syntax. The system reifies its syntactic analysis in the reasoning system, using predicates that extend the Cyc ontology. This supports a knowledge-rich interpretation process, which uses abduction (Hobbs, 2004) and discourse representation theory (Kamp & Reyle, 1993), implemented via microtheory

inheritance. The system uses this linguistic and conceptual information in cases for analogical word sense disambiguation. The semantics generated by the system are derived from ResearchCyc's frames, as described above.

EA NLU is designed to create representations useful for reasoning from text (e.g., Barbella & Forbus, 2011; Lockwood & Forbus, 2009), not for broad syntactic coverage. Instead, we use simplified English (Clark et al., 2005; Kuehne & Forbus, 2004) in which complex sentences are broken up into multiple shorter, simpler ones, and sentences with certain uncommon sentence constructions are reconfigured into other forms. We believe that this is a reasonable and practical solution for teaching cognitive systems, because people commonly use similar simplifications themselves when communicating with non-native speakers (e.g., the Simple English Wikipedia).

3. How Analogical Word Sense Disambiguation Works

The fundamental idea of analogical word sense disambiguation is that choices made in similar prior circumstances are good sources of evidence for what to do in understanding new sentences. The system uses analogy to construct generalizations of training examples and analogical retrieval to choose the most similar from among those generalizations and examples in order to disambiguate new choices later. Our work differs from prior research in that it adopts conceptual representations for its sense inventories, and uses relational representations instead of feature-based representations. We assume there are other methods of disambiguation to fall back on when analogies are not available, and that successful instances of disambiguation are used to incrementally provide cases that can be used in analogical learning. We will not discuss those methods here. For evaluation, we use hand-made disambiguation choices to prime the pump, to focus on what analogy can contribute.

The analogical processing techniques we employ are based on Gentner's (1983) structure-mapping theory. Analogical matching is performed by the Structure-Mapping Engine, SME (Falkenhainer, Forbus, & Gentner, 1989), which takes two structured, relational inputs, a *base* and *target*, which are structured, relational representations. It uses a greedy merge algorithm (Forbus, Ferguson, & Gentner, 1994) to produce, in polynomial time, one or more *mappings*. A mapping contains a set of *correspondences* that indicate what entities and statements in the base and target align, a *score* indicating the structural quality of the match, and *candidate inferences* that describe what information from the base can be projected into the target. SME is a component in the models of analogical retrieval and generalization, described below, which are used in our algorithm. These models have all been used to simulate a variety of psychological phenomena (e.g., Dehghani et al. 2008, Friedman & Forbus, 2008), and they have also been used in performance systems (e.g., Hinrichs & Forbus 2012; Lockwood & Forbus 2009), making them reasonable starting points for disambiguation.

We begin by describing how cases are constructed from the partial analysis of a sentence. Then we describe how the system uses analogical learning to accumulate experience. Finally, we describe how it generates evidence for disambiguation based on that accumulated experience.

3.1 Case Construction and Representation

By analyzing a disambiguation selection made by some other means, the system generates a case for analogical learning. (For our current work, the training step is done using gold standard

disambiguations. In the future, our hope is that the system will be able to use its own disambiguations – at least ones in which it has high confidence – as training material, becoming semi-supervised.) There are many facts that might be relevant, including detailed parse information, information about choices made in the sentence and in nearby sentences, and characteristics of the document from which the text was drawn. Based on preliminary experiments, we have focused on a small set of relationships that describe the local context for the interpretation of an occurrence of a word, i.e., those which concern the sentence in which it occurs. (We discuss possible extensions in the section on future work.)

The system constructs a separate case for each occurrence of each word in every sentence, to provide a focused context for making a word sense choice. This process is used to generate both the cases used in training and the cases used as retrieval probes for later disambiguation. When the system attempts to disambiguate a new occurrence of a word, it makes a case for it that is used as a probe for analogical retrieval. The case (or generalization from several cases) returned from retrieval is used as evidence for disambiguation.

Consider the sentence, "The heat flows to the object with the lower temperature." Each content word in the sentence will have a case generated for it. Consider the construction of the case for the word "flows". This case includes a set of statements that describe the context in which this decision appears, which incorporates semantic, surface, and parse information. In our example, these statements include:

- A statement indicating the word sense used. For this example, the word sense includes both the sense itself, `FluidFlow-Translation`, and a role relation statement that connects the action to one of its actors, `(primaryObjectMoving flow-22523 heat-22505)`. In a case for a noun, this will often be simpler, and only indicate that the word belongs to a particular collection (e.g., `FamousHuman`);
- Statements describing the choices selected for the other words in the sentence (e.g., `ThermalEnergy`). These statements encode the semantics of the sentence. Most nouns generate statements describing what something is, such as `(isa heat-22505 ThermalEnergy)`, while other words generate more complex statements, such as `(possessiveRelation object-22573 temperature-22651)`.

This provides a very simple form of semantic information about the context. Other facts record surface information:

- A statement indicating that the word covered by the choice set is `flows;`
- Statements indicating the other words in the sentence (`heat, lower,` etc.).

Information from the syntactic analysis of the sentence is also included, such as:

- The subject of the sentence (`heat`);
- The part of speech of the word (`verb`);
- If the word appears in a compound noun phrase, the word(s) with which it appears are also recorded. In the sentence "He emerged as a league star and his scoring entertained crowds," for the word "star" this would be (`league`);
- If the word has a prepositional phrase attached to it, the preposition; in this example, "to the object…" is attached to "flows", so (`to`) is recorded.
- If the word appears in a prepositional phrase, the preposition.

Table 1 illustrates a portion of the case generated for the decision about "flows" made in our example sentence, "The heat flows to the object with the lower temperature." Logical functions are used to reify information about constants in order to improve sensitivity in analogical retrieval. Neither SME nor MAC/FAC (described below), by design, is sensitive to specific constants that are the arguments to relations, such as `verb`. By redundantly encoding relational information as attributes (e.g., `(partOfSpeechOfChoiceSetFn verb)`), analogical retrieval picks up on information it would otherwise miss.

Table 1. A portion of a disambiguation case. Only ten out of a total of 32 facts are shown. The first three facts shown concern conceptual information. Nine facts out of 32 concern semantics, with the rest encoding syntactic or lexical information.

```
;;; Subset of the case generated for "The heat flows to the object with the lower
;;; temperature."
;;; The word sense used, FluidFlow-Translation, and a role relation connecting the
;;; action to one of the actors.
(selectedChoice ChoiceSet-22757 Choice-22758
  (and (isa flow-22523 FluidFlow-Translation)
       (primaryObjectMoving flow-22523 heat-22505)))

;;; Other choices made in the sentence.
(otherSentenceSelectedChoice (isa heat-22505 ThermalEnergy))
(otherSentenceSelectedChoice
  (possessiveRelation object-22573 temperature-22651))

;;; The word covered by the choice set
(isa ChoiceSet-22757 (ChoiceSetForPhraseFn (TheList flows)))

;;; Other words that are present in the sentence – "heat" and "lower".
(isa ChoiceSet-22757 (wordInSentenceOfChoiceFn heat))
(isa ChoiceSet-22757 (wordInSentenceOfChoiceFn lower))

;;; The subject of the sentence
(isa ChoiceSet-22757 (subjectOfSentenceOfChoiceSetFn (TheList heat)))

;;; The part of speech of the word
(isa ChoiceSet-22757 (partOfSpeechOfChoiceSetFn verb))

;;; Preposition attached to the word
(isa ChoiceSet-22757 (hasPPWithLeadingPrepositionFn to))
```

3.2 Analogical Learning

Generalizing from experience is an important form of information compression. Analogical generalization is performed via SAGE, an extension of SEQL (Kuehne et al., 2000). SAGE organizes knowledge in *generalization contexts* (Friedman & Forbus, 2008), which accumulate knowledge about a concept. Each generalization context has a *trigger*, a pattern that, when satisfied by an incoming example, indicates that that example should be included in that generalization context, i.e., a class label. (For example, in word sense disambiguation, the triggers are statements about the choice of sense made for an occurrence of a word.) Each

generalization context maintains a set of generalizations and unassimilated examples. As new examples arrive, the system uses MAC/FAC for retrieval, treating each generalization context as a case library. If the new example is sufficiently close to a prior generalization, as determined by SME's score exceeding an *assimilation threshold*, SAGE assimilates that example into the generalization. This involves updating the frequency information of statements in the generalization, based on its overlap with the new example (Halstead & Forbus, 2005). For example, if a generalization contains one hundred swans and only one of them is black, then in that generalization

```
(probability (hasColorBlack <swan>) 0.99)
(probability (hasColorWhite <swan>) 0.01)
```

will hold. If the new example is sufficiently close to a previously unassimilated example, a new generalization will be constructed by replacing non-identical entities with arbitrary individuals and starting to accumulate frequency information for each matching statement. Statements whose probability becomes sufficiently low over time are filtered out. The ability to maintain multiple generalizations is useful for handling disjunctive concepts, and the ability to maintain unassimilated examples is useful for dealing with exceptions and edge cases. A novel feature of SAGE is that it typically achieves robust performance after only a dozen cases for each concept, making it faster (in terms of number of examples required) than most statistical learners.

During training, the system produces cases using examples disambiguated by other means, and then uses SAGE to put those examples into generalization contexts. Examples that are similar to each other are merged into generalizations, while examples that are not sufficiently similar to any of the previously encountered examples remain in memory as ungeneralized examples. For example, suppose the system is training on the word "star" and one of the examples it is given is "A red star's temperature is lower than the Sun's temperature." A case representing information about the sentence and indicating that the word "star" is the word being disambiguated is stored in the generalization context. Later, the system is given the example "A blue star's temperature exceeds the Sun's temperature." The case constructed for this sentence is very similar to the one constructed for the first sentence, so it is added to a generalization with that sentence, even when the assimilation threshold is quite high. Later, the system is given the sentence "A star changes hydrogen into helium through nuclear fusion" as an example. As the case generated for this sentence is not sufficiently similar to the existing generalizations or cases, it remains in memory as an ungeneralized example.

The system uses analogical retrieval to find one or more cases that are close to an incoming example. The MAC/FAC model of analogical retrieval (Forbus, Gentner, & Law, 1995) assumes a case library of structured, relational representations and a probe case that is also structured. For each case, it automatically computes a special kind of feature vector, a *content vector*, from the structured representations. Each dimension of a content vector corresponds to a predicate, with its weight indicating the number of occurrences of statements involving that predicate in the case. The dot product of two content vectors provides an estimate of the similarity score that SME would compute on the corresponding structured representations, but is extremely cheap. The first stage calculates, conceptually in parallel, the dot product of the content vector for the probe with the content vectors for every case in the library. The best, and up to two others if they are

sufficiently close, are returned as the output of the first stage. The structured representations of these cases are then compared to the structured representation of the probe, again conceptually in parallel, using SME. The best, and up to two others if they are sufficiently close, are returned as the result. The cheap and parallelizable first stage is important because it provides potential scalability. Since each new sentence the system reads will lead to the accumulation of several new cases, there is the potential for accumulating massive case libraries as the same words are encountered repeatedly.

When the system generates a disambiguation case, it is added to the appropriate generalization context so that SAGE can process it. Each combination of word and word sense has its own generalization context – for example, there is a generalization context for the `FamousHuman` sense of "star", and another for the astronomical sense. This is done so that each case provides evidence for only one sense. Recall the example of the color of swans from Section 2; if only one generalization context were used for each word, then generalizations could be constructed that were ambiguous in their decisions. Moreover, SAGE eventually drops low-frequency relationships, when their probability goes below a threshold, and thus information about infrequent but still valid uses of a word sense could be lost.

Recall that SAGE uses a threshold to decide when a new example is close enough to assimilate what is retrieved (either a prior example or a generalization) with the example or just store the new example. The setting of the assimilation threshold thus determines a tradeoff between how much SAGE generalizes versus simply stores cases. We use this ability in Section 4 to examine how much benefit generalization provides over learning by only accumulating cases.

3.3 Retrieval and Evidence Production

Given a new occurrence of a word to disambiguate, our technique uses MAC/FAC to retrieve relevant prior cases and/or generalizations, if any. Recall that each generalization context covers a word plus a particular sense for that word. The union of all generalization contexts involving that word, for all of the senses seen so far, constitutes the case library. In the example above, the system uses a separate generalization context not only for different senses of the word "flows", including `<(TheList flows), (isa ?entity FluidFlow-Translation)>`, but also for different combinations of role relations, such as `<(TheList flows), (and (isa ?entity FluidFlow-Translation) (primaryObjectMoving ?entity ?object)>`. In our example, the second of these is the correct sense; the system will be successful if it retrieves a case or generalization that had been part of that generalization context.

The candidate inferences generated by SME during retrieval are examined to determine which word sense choice is best supported by experience. Given the way that cases are encoded, these inferences will include the decision associated with that case or generalization. The system uses that prior decision to generate evidence for the current decision.[1] Returning to our example, when the occurrence of "flows" in "The heat flows to the object with the lower temperature" is used as a probe, the system retrieves a generalization of four previously seen cases, including

[1] In the experiments described here, when analogy is used its choice is always made, and analogical results are never combined with other information, so we do not describe the details of evidential reasoning used in the learning by reading system.

"Despite this, volume flows toward the can with the low depth" and "Volume flows to the lower depth." This generalization is retrieved because it is the most similar generalization or exemplar in the corpus (described below). The analogy between these generates a candidate inference of the form

```
(selectedChoice
  ChoiceSet-16374
  (:skolem Choice-35146)
  (and (isa (:skolem flow34459) FluidFlow-Translation)
    (primaryObjectMoving ?entity (:skolem heat34459)))
```

suggesting the `FluidFlow-Translation` sense that also includes `primaryObjectMoving`, which is correct in this case. The reasons that the generalization is chosen include not only a shared subject and several shared words, but a similar structure – the sentence and the generalization both have a prepositional phrase starting with "to" attached to the word "flows". Table 2 shows a portion of the match between the two cases.

Table 2. A portion of the match from the retrieval made with the probe case for "The heat flows to the object with the lower temperature" in Experiment 2. `(GenEntFn 10 flows)` is a generalized entity that was created when cases were merged during training. Some representations have been simplified for space and readability.

Base Case	Retrieved Target
`((hasPPWithLeadingPrepositionFn to) ChoiceSet-3583789884-50713)`	`((hasPPWithLeadingPrepositionFn t0) (GenEntFn 10 flows))`
`(possessiveRelation object6515 temperature6596)`	`(possessiveRelation can6515 temperature6596)`
`((ChoiceSetForPhraseFn (TheList flows)) ChoiceSet-3583789884-50713)`	`((ChoiceSetForPhraseFn (TheList flows)) (GenEntFn 10 flows))`
`((wordInSentenceOfChoiceFn lower) ChoiceSet-3583789884-50713)`	`((wordInSentenceOfChoiceFn lower) (GenEntFn 10 flows))`

In some situations, the system retrieves an incorrect example; for example, while attempting to disambiguate the first "depth" in the sentence "The depth of the can differs from the depth of the other can," the system retrieves the sentence "Depth differs from volume." While those sentences are similar structurally and have other features in common, the first refers to a specific depth of a particular object, while the second refers to the general concept of depth. The distinction is relatively subtle, but as the representations are different internally, the system does not select the best choice based on this retrieval.

Because a sense that appears more frequently will have more examples in its generalization context, this process automatically incorporates sense frequency as a factor. That is, if a sense is very common in the training corpus, there will be more generalizations and exemplars available for it to match.

Table 3. Accuracy for binary disambiguation of *star*, mean of five-fold cross-validation. Results significant vs. chance; $p < .05$. n = 113 in all conditions. No significant difference between generalization and retrieval.

Sense	Retrieval Only	Generalization: Threshold 0.2	Generalization: Threshold 0.4
FamousHuman	71%	60%	75%
Star	85%	95%	85%
Total	79%	79%	81%

4. Experiments

Our main hypothesis is that analogical processing over relational representations can provide accuracies comparable to other word sense disambiguation algorithms. Since prior research has focused on very different sense inventories, representations, and texts, performance comparisons must be made with care. Nevertheless, they provide a useful component-level benchmark, since accuracy is a desirable property in all cases. To provide evidence for this hypothesis, we conducted two experiments. The first tests the ability of our technique to distinguish between senses of a specific word. As noted above, our disambiguation task uses a very different kind of sense inventory than traditionally used, but distinguishing noun senses provides the closest comparison. The second experiment tests the ability of the technique to learn incrementally, in a setting that is a closer approximation to what learning by reading systems need.

4.1 Experiment 1: Noun Sense Disambiguation

To test performance on a traditional word sense disambiguation task, we created a test corpus of 104 sentences that used the noun "star" in either the sense of a famous person or an astronomical object, as per our running example. The sentences were drawn from the English and Simple English Wikipedias, and simplified when necessary to fit the syntax that EA NLU handles. The sentences were roughly evenly divided between the two word senses. We used five-fold cross validation, with approximately ten examples of each word sense in the test set, with the remainder of each corpus providing the set of cases used as input for training.[2] This experiment is not intended to model a normal reading experience; the senses of "star" do not occur with the same frequencies in this corpus as they do in more natural corpora, and are effectively drawn from many different documents. Indeed, in a normal reading situation, the fact that polysemous words that recur within a document typically carry the same sense (Gale, Church, & Yarowsky, 1992) can be exploited. This experiment is intended to test performance in situations where a comparable number of examples of two different senses exist in the training set.

Table 3 summarizes the results of three conditions, varying only in assimilation threshold. By setting the threshold to 1.0, SAGE only stores examples, as otherwise cases must identical to be assimilated at that threshold. The 0.2 condition is the opposite extreme, leading it to form generalizations more readily, while the 0.4 condition is more conservative. Importantly, there is no significant difference in accuracy between these three conditions, suggesting that, for this task,

[2] The "astronomical body" sense appears with slightly greater frequency in the corpus because a greater number of sentences that use that sense also use it more than once.

SAGE is insensitive to the particular threshold used. Accuracy is significantly and substantially different from chance (which would be 50%) in all conditions.

Moreover, the accuracy achieved is in the range reported for state of the art algorithms with more traditional word sense inventories. For example, Hearst (1991) also used a method involving storing syntactic and semantic information from training data for disambiguation, and hence is the most comparable system to ours. Of the six words that Hearst used, accuracies after 40 examples ranged from 75% to 85%,[3] and from 82% to 88% on the two words for which 50 or more examples were used. Thus although our results are not directly comparable due to differences in training data, test data, and representations, our accuracy is comparable to Hearst's.

Table 4. Generalizations created on first fold of Experiment 1.

Trial	Examples	Ungenl'd Examples	Gens Created (Star)	Gens Created (FamousHuman)	Min Gen Size	Max Gen Size
Retrieval Only	92	92	0	0	N/A	N/A
Generalization: Threshold 0.2	92	34	16	12	2	3
Generalization: Threshold 0.4	92	44	14	10	2	2

Our results are also similar to those reported by other systems. GAMBL, which employs example-based learning and a genetic algorithm for selecting which features to use, achieves an accuracy of 74% on binary course-grained word senses (Decadt et al., 2004). The IMS system (Zhang & Ng, 2010), which uses support vector machines and cross-linguistic parallel corpora, achieved a 73% accuracy rating on a lexical sample test corpus. The best-performing systems on Semeval-2007, which featured a lexical sample task,[4] achieved an f-score of about 89% (Pradhan et al., 2007). There are several differences between that task and ours. One hundred words were tested, the number of training examples ranged from 32 to 1009 (mean 223), and the number of senses present ranged from one to 13 (one to eight when senses with fewer than three instances are excluded). The corpus is not sense-balanced, with the most-frequent-sense baseline achieving 78% accuracy.

There are word sense disambiguation methods that provide higher accuracy, but at the cost of vastly more training data. For example, Yarowsky (1995) achieves 94 to 98% accuracy on similar tasks using a semi-supervised method. However, that method used an (unlabeled) corpus of 400 to 11,000 examples per word, whereas ours requires relatively little labeled training data and does not require a large corpus to operate.

Acquisition rate is also an important concern for learning by reading, as word sense disambiguation is just one component of a larger system. Consequently, we also measured how few examples sufficed to achieve reasonable levels of performance. Keeping the assimilation threshold at 0.4, when we varied the number of training examples from 2 to 80, the accuracy ranged from 73% to 81%, a difference that is not statistically significant (n = 113). In other words, by twenty examples the system is already performing reasonably well.

[3] This leaves out "bass", which hit 100% after only 25 examples, and thus was an outlier.
[4] http://nlp.cs.swarthmore.edu/semeval/tasks/task17/description.shtml

We noted earlier the lack of significant difference between the accuracy for the retrieval-only and generalization conditions. Could this be because retrieved cases are doing all the work, in both conditions? This is not the reason. Table 4 illustrates, for one fold, the number of generalizations and examples produced, where size indicates the number of examples assimilated to produce a generalization. As expected, a lower assimilation threshold leads to more and larger generalizations. In the 0.4 condition, generalizations provided evidence 23% of the time, and in the 0.2 condition, generalizations provided evidence 21% of the time. This difference is not statistically significant. As the system accumulates yet more experience, some of what are now cases would become generalizations, as sufficiently similar sentences are encountered. Thus we believe that the full benefit of generalization may only be apparent when used in the intended context of learning by reading, where a system will be reading very large amounts of text. We expect that generalizations will be used most of the time as a system becomes more experienced, and the information compression they provide will help keep memory loads reasonable.

4.2 Experiment 2: Word Sense Disambiguation in Connected Text

The previous experiment served to examine accuracy in the kind of test traditionally used in word sense disambiguation research. In this subsection we introduce a new kind of test, involving connected text, that is designed to be a more natural evaluation for word sense disambiguation in the context of learning by reading. Recall that our hypotheses are that (1) analogical word sense disambiguation will have reasonably high accuracy, compared to other systems, while reading connected text about a topic and that (2) its rate of learning will be rapid. Experiment 2 provides evidence for both hypotheses.

We divided Chapter 2 of a popular science book (Buckley, 1979) into sections approximately ten sentences long each. There were a total of 90 sentences after simplification, leading to nine subsections. For each section, the disambiguation procedure was run using the generalization contexts generated from gold standards of the previous sections as the source of cases for disambiguation. (This means that for the first section there will be no correct answers.) We note that this is an idealization: mistakes in understanding commonly occur when people read, and they will be even more likely in reading systems that have a much weaker grasp on the world. Nevertheless, we believe that this is a useful idealization because it shows in a pure form what the word sense disambiguation model is doing, independent of any particular error model. After recording accuracy, cases for that section were automatically generated, based on hand-selected word sense choices, to extend the generalization contexts accordingly.

There are several things worth noting about this experimental setup. First, the system gets many choices correct simply because they are monosemous within the text. If only one sense is ever used and the system has encountered it in a prior section, the reminding will be correct. For example, the word “pan”, within the text used in this experiment, always refers to a cooking pan. We refer to such choice sets as *easy*, because any system with access to the set of training data could get them correct by simply choosing the only attested sense. However, there are also situations where a previously seen word is now being used in a novel sense, and hence there are no cases with the correct sense. We refer to such choice sets as *impossible*. The most interesting situations are where multiple senses have previously been encountered and one of them is correct. We refer to these as *interesting* choice sets. Every choice set is either easy, impossible, or interesting. We focus on the interesting ones in measuring the accuracy of analogical dis-

Table 5. Results from Experiment 2: Progressive performance on sections of a text, with generalization at an assimilation threshold of 0.4.

Section	Total	Total Accuracy	Impossible	Easy	Interesting	Interesting Accuracy
1	40	0%	40 (100%)	0 (0%)	0 (0%)	N/A
2	61	30%	43 (70%)	18 (30%)	0 (0%)	N/A
3	51	39%	28 (55%)	18 (35%)	5 (10%)	80%
4	66	59%	22 (33%)	25 (38%)	19 (29%)	95%
5	68	32%	41 (60%)	17 (25%)	10 (13%)	60%
6	57	35%	31 (54%)	10 (18%)	16 (28%)	63%
7	62	61%	20 (32%)	11 (18%)	31(50%)	87%
8	69	49%	23 (33%)	27 (39%)	19 (28%)	37%
9	54	48%	17 (31%)	6 (11%)	31(57%)	65%

ambiguation, but we tabulate the others because they help provide upper and lower bounds on how much can be obtained by any learning system that accumulates information while reading.

Generally most words are used with only one word sense within a given document (Daelemans, Van Den Bosch, & Zavrel, 1999). There are 131 interesting choice sets in our experiment, out of 528 total (24.8%). This is higher than might be expected, but they represent only 18 distinct words. Most of the 130 words that generate choice sets are monosemous within the text. There are also a small number of words (e.g., “an”, “the”) for which the language system does not generate choice sets. The 18 words that generate at least one interesting choice set are those which are fairly common in the text, including “of”, “can” and “temperature”. Only 52 out of 131 interesting choice sets are choices between word senses alone: the 79 others involve the same word sense but differ in semantic frames, i.e., in role relations that interconnect constituents. Thus this problem is substantially different than what is normally addressed in research on word sense disambiguation.

Table 5 summarizes the results. For each section, Total is the total number of choice sets generated by the NLU system. In general there are fewer choice sets than there are words, as some words, such as determiners, do not generate any. Total accuracy is the number that the procedure got correct. Impossible, Easy, and Interesting indicate how many such choice sets that section contained. The first section had no prior examples, and the second section had no interesting choice sets, so Interesting is zero in both sections. The last column gives the system’s accuracy on interesting choice sets. For them, the mean number of senses previously seen is 3.13. The total accuracy on interesting choice sets is 70.2 percent. Note that this number cannot be directly compared to accuracy numbers from most other disambiguation experiments (including from our Experiment 1), as it covers only disambiguation cases where multiple senses of the word have previously been seen, which is not what experiments typically measure. However, it is still encouraging that it is comparable in accuracy to other disambiguation techniques applied to what we view as simpler disambiguation problems.

The total accuracy on all choice sets is 41.1 percent. Note that this includes section 1, where all choice sets were impossible. The total accuracy on words seen before (easy, interesting, and impossible choice sets that previously appeared in the text with a different sense) is 71.3 percent. Again, owing the nature of the task, we cannot directly compare these numbers to most results on traditional word sense disambiguation, but it does suggest that analogical disambiguation can be useful for learning by reading.

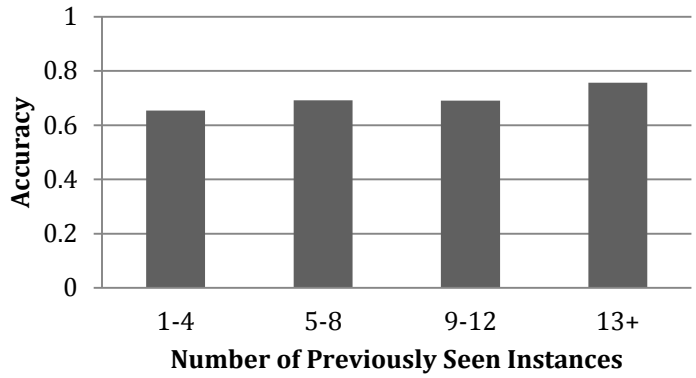

Figure 1. Accuracy on interesting choice sets for experiment 2, broken down by number of previously seen examples, with $n \geq 26$ for all groups. All results are significantly above chance, but there is no significant difference between categories. In 13+ category, the maximum number of previously seen examples is 25, whereas the mean is 16.9.

Table 5 also highlights how local properties of the text can affect the performance. Sudden shifts in sense use can cause temporary accuracy drops. For example, there is a topic shift at section 5 that is responsible not only for a drop in the accuracy over interesting choice sets, but also for the rise in the number of impossible choice sets. Note also that there is not a monotonic upward trend in accuracy on interesting words from section to section. This might seem surprising, but new words continue to be introduced throughout the text, and words that are not introduced until later have no precedents.

To examine learning rate on interesting choice sets, Figure 1 tabulates accuracy versus the number of available prior cases. Importantly, accuracy is in the 65 percent range even with only a handful of examples. This suggests that analogical word sense disambiguation can provide useful evidence, even in the early stages of becoming familiar with how words are used. The accuracy rises from 65% in the conditions with four or fewer previously seen instances to 76 percent in cases with 13 or more, although this increase is not statistically significant.

5. Discussion

We have described a new approach to word sense disambiguation that uses human-like models of analogical processing operating over relational representations that combine linguistic and conceptual information produced in the course of natural language understanding. The system operates either with analogical retrieval alone or with analogical retrieval and generalization. It provides comparable accuracy to traditional word sense disambiguation algorithms that rely on

feature-based representations, while requiring relatively little training data. It achieves 72% accuracy with just 20 training sentences. While that is not quite as accurate as a supervised system on a binary disambiguation task, it is comparably strong for the quantity of training data. This is important for learning by reading systems, which attempt to learn without the benefit of a large annotated corpus.

We also described an experiment that tested the system's performance in a situation more like those that face learning by reading systems. Our unconventional experimental setup was intended to measure the ability to disambiguate semantics with just an understanding of the previous sections of the text. Our system's performance on words with just a few previously-seen examples was surprisingly high, up to 70% even after only a few examples of a word. This is promising, as it suggests that the technique could be valuable for learning by reading.

5.1 Related Work

Our goal is a system that performs word sense disambiguation in the context of a cognitive agent that learns by reading. Thus the sense inventories we use are drawn from conceptual representations, rather than language-level vocabularies such as WordNet synsets. Nevertheless, the challenges are similar, as noted above. Our system most closely resembles CatchWord (Hearst, 1991) in that it incorporates a mix of semantic and syntactic information from the local context of a word in a set of labeled training data. However, CatchWord does not use analogy for retrieval over structured relational representations constructed from the training data.

We start with the contents of the ResearchCyc KB, plus our own extensions, which are extremely accurate because they are vetted via a careful (and laborious) knowledge engineering process. Machine reading research typically tries to start with less. Some efforts (e.g., Etzioni et al., 2005) only generate word-level descriptions, such as triples of co-occurring words. While relatively straightforward to gather, such representations do not provide the kind of conceptual precision needed for reasoning. Systems like DIRT (Pantel et al., 2007) go further by extracting rules involving triples from dependency trees used in parsing. As Clark and Harrison (2010) point out, DIRT's 12 million inference rules are quite noisy, with longer chains being less reliable. DART (Clark & Harrison, 2009) and KNEXT (Van Durme & Schubert, 2009) have improved on accuracy while still being fully automatically extracted.

Keeping accuracy high when learning by reading is not an easy problem, as work on NELL (Carson et al., 2010) points out: accuracy falls sharply as the system continues to add knowledge. The PRISMATIC knowledge base (Fan et al. 2012) generated as part of the IBM Watson effort is a landmark in learning by reading research, having accumulated close to a billion syntactic sentence frames; these were shown to be useful in factoid question answering that went well beyond the prior state of the art. It sidesteps word sense disambiguation entirely, by treating words themselves as predicates.

Several other prior efforts share our focus on producing highly accurate knowledge that can be directly used for reasoning. Mobius (Barker et al., 2007) and its descendent, Kleo (Kim & Porter, 2009), use CLIB, a compact library of generic concepts as their starting point. These systems use a semantic interpretation process that is not as focused on linking words in a text to senses from a large pre-existing ontology of concepts, and thus are less concerned about word sense disambiguation as part of learning by reading.

5.2 Directions for Future Work

We see two directions for future work as particularly important. First, we want to increase accuracy and learning rates even further. We plan to use the statistical information accumulated in the generalizations to ascertain which relationships and attributes are pulling their weight, and thereby adapt the case construction process accordingly, to increase the discrimination of analogical retrieval. We also plan to explore whether additional conceptual information from the discourse context can be used to speed learning. For example, in a text on solar energy, the word "heat" rarely means spicy or sexy. Additionally, inspired by the use of confidence estimation in IBM's Watson (Gondek et al., 2012), we plan to explore techniques for enabling the system to provide confidence values for the choices it makes, based on numerical similarity estimates computed by SME. These confidence values can be combined with evidence from other disambiguation techniques (e.g., abduction), both to make initial choices and to help guide backtracking during semantic interpretation. Additionally, we plan to explore letting the system use its own disambiguations as precedents, at least where its confidence is sufficiently high.

The second direction is to use this technique in a next-generation learning by reading system, one capable of integrating knowledge automatically from article-length simplified English texts. This system will create baseline cases from which analogies are drawn using a combination of top-down abduction (Tomai & Forbus, 2009) and sense choices that lead to measured improvements in reasoning performance (i.e., accuracy in question answering). Both of these techniques will be more expensive, and hence we hypothesize that analogical word sense disambiguation should provide a significant boost in performance for systems that learn by reading.

Finally, we note that in applying analogical processing to word sense disambiguation, we did not need to modify the SME, MAC/FAC, or SAGE systems. This provides additional evidence as to the generality of analogical processing, and highlights its potential for use in other cognitive systems and other settings.

Acknowledgements

This research was supported by Grant N00014-07-1-0919 from the Office of Naval Research. We thank Tom Hinrichs and C. J. McFate for many useful conversations and suggestions.

References

Allen, J. F. (1994). *Natural language understanding (2nd Ed.)* Redwood City, CA: Benjamin/Cummings.

Barker, K., Agashe, B., Chaw, S., Fan, J., Glass, M., Hobbs, J., Hovey, E., Israel, D., Kim, D., Mulkar, R., Patwardhan, S., Porter, B., Tecuci, D., & Yeh, P. (2007). Learning by reading: A prototype system, performance baseline, and lessons learned. *Proceedings of the Twenty-Second AAAI Conference on Artificial Intelligence* (pp. 280-286). Vancouver, BC: AAAI Press.

Barbella, D., & Forbus, K. (2011). Analogical dialogue acts: Supporting learning by reading analogies in instructional texts. *Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence.* (pp. 1429-1435). San Francisco, CA: AAAI Press.

Buckley, S. (1979). *Sun up to sun down*. New York: McGraw-Hill.

Carlson, A., Betteridge, B., Kisiel, B., Settles, E. R., Hruschka, E. R., & Mitchell, T. M. (2010). Toward an architecture for never-ending language learning. *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence* (pp. 1306-1313). Atlanta, GA: AAAI Press.

Clark, P., & Harrison, P. (2009). Large-scale extraction and use of knowledge from text. *Proceedings of the Fifth International Conference on Knowledge Capture* (pp. 153-160). Redondo Beach, CA: ACM.

Clark, P., & Harrison, P. (2010). BLUE-Lite: A knowledge-based lexical entailment system for RTE6. *Proceedings of the Third Text Analysis Conference.* Gaithersburg, MD: National Institute of Standards and Technology.

Clark, P., Harrison, P., Jenkins, T., Thompson, J., & Wojcik, R. (2005). Acquiring and using world knowledge using a restricted subset of English. *Proceedings of the Eighteenth International Florida Artificial Intelligence Research Society Conference* (pp. 506-511). Clearwater Beach, FL: AAAI Press.

Curtis J., Baxter D., Wagner P., Cabral J., Schneider D., Witbrock M. (2009). Methods of rule acquisition in the TextLearner system. *Proceedings of the AAAI Spring Symposium on Learning by Reading and Learning to Read* (pp. 22-28). AAAI Press.

Daelemans, W., Van Den Bosch, A., & Zavrel, J. (1999). Forgetting exceptions is harmful in language learning. *Machine Learning, 34*, 11-41.

Decadt, B., Hoste, V., Daelemans, W., & van den Bosch, A. (2004). GAMBL, genetic algorithm optimization of memory-based WSD. *Proceedings of the Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text* (pp. 108-112). Barcelona, Spain: Association for Computational Linguistics.

Dehghani, M., Tomai, E., Forbus, K., & Klenk, M. (2008). An integrated reasoning approach to moral decision-making. *Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence* (pp. 1280-1286). Chicago, IL: AAAI Press.

Etzioni, O., Cafarella, M. , Downey, D. , Popescu, A. , Shaked, T. , Soderland, S. , Weld, D., & Yates, A. (2005). Unsupervised named-entity extraction from the Web: An experimental study. *Artificial Intelligence, 165*¸ 91-134.

Falkenhainer, B., Forbus, K., & Gentner, D. (1989). The Structure-Mapping Engine: Algorithm and examples. *Artificial Intelligence, 41*, 1-63.

Fan, J., Kalyanpur, A., Gondek, D.C., & Ferrucci, D.A. (2012). Automatic knowledge extraction from documents. *IBM Journal of Research & Developement, 56*, 5:1-5:10.

Forbus, K., Ferguson, R. & Gentner, D. (1994). Incremental structure-mapping. *Proceedings of the Sixteenth Annual Conference of the Cognitive Science Society* (pp. 313-316). Atlanta, GA: Lawrence Erlbaum.

Forbus, K., Gentner, D. & Law, K. (1995). MAC/FAC: A model of similarity-based retrieval. *Cognitive Science, 19*, 141-205.

Forbus, K, Klenk, M. & Hinrichs, T. (2009). Companion Cognitive Systems: Design goals and lessons learned so far. *IEEE Intelligent Systems, 24*, 36-46.

Forbus, K., Riesbeck, C., Birnbaum, L., Livingston, K., Sharma, A., & Ureel, L. (2007). Integrating natural language, knowledge representation and reasoning, and analogical processing to learn by leading. *Proceedings of the Twenty-Second AAAI Conference on Artificial Intelligence* (pp. 1542-1547). Vancouver, BC: AAAI Press.

Francis, W. N. & Kučera, H. (1964). *Manual of information to accompany a standard corpus of present-day edited American English, for use with digital computers*. Providence, RI: Department of Linguistics, Brown University. Revised 1971. Revised and amplified 1979.

Friedman, S., & Forbus, K. (2008). Learning causal models via progressive alignment & qualitative modeling: A simulation. *Proceedings of the Thirtieth Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

Gale, W. A., Church, K., & Yarowsky, D. (1992). Estimating upper and lower bounds on the performance of word-sense disambiguation programs. *Proceedings of the Thirtieth Annual Meeting of the Association for Computational Linguistics* (pp. 249–256). Newark, DE: Association for Computational Linguistics.

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science, 7*, 155-170.

Gondek,D., Lally, A., Kalyanpur, A. Murdock, J., Duboue, P., Zhang, L., Pan, Y., Qui, Z., & Welty, C. (2012). A framework for merging and ranking of answers in DeepQA. *IBM Journal of Research & Devevelopment, Volume 56, Issue 3:4* (pp. 14:1-14:12).

Grishman, R., Macleod, C., & Wolff, S. (1993). *The COMLEX Syntax Project*. Ft. Belvoir: Defense Technical Information Center.

Halstead, D., & Forbus, K. (2005). Transforming between propositions and features: Bridging the gap. *Proceedings of the Twentieth National Conference on Artificial Intelligence* (pp. 777-782). Pittsburgh, PA: AAAI Press/The MIT Press

Hearst, M. (1991). Noun homograph disambiguation using local context in large text corpora. In *Proceedings of the Seventh Annual Conference of the UW Centre for the New OED and Text Research: Using Corpora*. Waterloo, ON: University of Waterloo.

Hinrichs, T., & Forbus, K. (2012). Learning qualitative models by demonstration. *Proceedings of the Twenty-sixth AAAI Conference on Artificial Intelligence* (pp. 207-213). Toronto, ON: AAAI Press.

Hobbs, J. R. (2004). Abduction in natural language understanding. In L. Horn & G. Ward, (Eds.), *Handbook of pragmatics*. Padstow, England: Blackwell Publishing.

Kamp, H., & Reyle, U. (1993). *From discourse to logic: Introduction to model-theoretic semantics of natural language*. Boston, MA: Kluwer Academic.

Kim, D., & Porter, B. (2009). Kleo: A bootstrapping learning by reading system. *Proceedings of the AAAI Spring Symposium on Learning by Reading and Learning to Read* (pp. 44-49). Menlo Park, CA: AAAI Press.

Kuehne, S. E. (2004). *Understanding natural language descriptions of physical phenomena*. Doctoral dissertation, Computer Science Department, Northwestern University, Evanston, IL.

Kuehne, S., & Forbus, K. (2004*).* Capturing QP-relevant information from natural language text. *Proceedings of the Eighteenth International Qualitative Reasoning Workshop*. Evanston, IL.

Kuehne, S., Forbus, K., Gentner, D., & Quinn, B. (2000). SEQL: Category learning as progressive abstraction using structure mapping. *Proceedings of the Twenty-Second Annual Conference of the Cognitive Science Society* (pp. 770-775). Philadelphia, PA: Cognitive Science Society.

Lockwood, K., & Forbus, K. (2009). Multimodal knowledge capture from text and diagrams. *Proceedings of the Fifth International Conference on Knowledge Capture* (pp. 65-72). Redondo Beach, CA: AAAI Press.

Macleod, C., Grisham, R., & Meyers, A. (1998). *COMLEX syntax reference manual, Version 3.0*.

Navigli, R. (2009). Word sense disambiguation: A survey. *ACM Computing Surveys, 41*, 10:1-10:69.

Paul, D. B., & Baker, J. M. (1992). The design for the Wall Street Journal-based CSR corpus. *Proceedings of the Workshop on Speech and Natural Language* (pp. 357-362). Stroudsburg, PA: Association for Computational Linguistics.

Pantel, P., Bhagat, R., Coppola, B., Chklovski, T., & Hovy, E. (2007). ISP: Learning inferential selection preferences. *Proceedings of the 2007 Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics* (pp. 131-138). Rochester, NY: Association for Computational Linguistics.

Pradhan, S., Loper, E., Dligach, D., & Palmer, M. (2007). SemEval-2007 task 17: English lexical sample, SRL and all words. *Proceedings of the Fourth International Workshop on Semantic Evaluations* (pp. 87-92). Stroudsburg, PA: Association for Computational Linguistics.

Tomai, E., & Forbus, K. (2009). EA NLU: Practical language understanding for cognitive modeling. *Proceedings of the 22nd International Florida Artificial Intelligence Research Society Conference*. Sanibel Island, FL: AAAI Press.

Van Durme, B., & Schubert, L. (2008). Open knowledge extraction through compositional language processing. *Symposium on Semantics in Systems for Text Processing* (pp. 239-254). Stroudsburg, PA: Association for Computational Linguistics.

Yarowsky, D. (1995). Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics* (189-196). Stroudsburg, PA: Association for Computational Linguistics.