A Functional Analysis of Historical Memory Retrieval Bias in the Word Sense Disambiguation Task

Nate Derbinsky and John E. Laird
University of Michigan
2260 Hayward St.
Ann Arbor, MI 48109-2121
{nlderbin, laird}@umich.edu

Abstract

Effective access to knowledge within large declarative memory stores is one challenge in the development and understanding of long-living, generally intelligent agents. We focus on a sub-component of this problem: given a large store of knowledge, how should an agent's task-independent memory mechanism respond to an ambiguous cue, one that pertains to multiple previously encoded memories. A large body of cognitive modeling work suggests that human memory retrievals are biased in part by the recency and frequency of past memory access. In this paper, we evaluate the functional benefit of a set of memory retrieval heuristics that incorporate these biases, in the context of the word sense disambiguation task, in which an agent must identify the most appropriate word meaning in response to an ambiguous linguistic cue. In addition, we develop methods to integrate these retrieval biases within a task-independent declarative memory system implemented in the Soar cognitive architecture and evaluate their effectiveness and efficiency in three commonly used semantic concordances.

Introduction

One challenge in cognitive architecture research is to develop long-term memory systems that are capable of extracting diverse, useful experiences from agent interactions with the world; store large amounts of this information for long periods of time; and later retrieve this knowledge when it is relevant to making decisions and taking action. There is evidence that extending agents with long-term memory supports many functional cognitive capabilities (Nuxoll and Laird 2007); however, maintaining and querying large memories poses significant computational challenges that currently make it impossible to task these agents with real-world problems.

The focus of this paper is on one specific challenge facing long-term memory: given a large store of knowledge, and an ambiguous cue that matches multiple stored memories, how does the system efficiently determine which memory to retrieve? Anderson and Schooler (1991), positing that human memory optimally solves this problem with respect to the history of past memory access, have developed and validated memory models that are widely used in the cognitive modeling community. However, existing computational implementations of these models do not scale to tasks that require access to large bodies of knowledge (Douglass, Ball, and Rodgers 2009).

Previous work (Derbinsky, Laird, and Smith 2010) developed and evaluated techniques to efficiently support queries of large declarative memory stores; however, that work supported only a limited class of bias in the case of ambiguous cues, and did not evaluate the functional benefits of biases within that space.

Our goal is to develop a suite of possible memory retrieval heuristics and evaluate their effectiveness and efficiency in a variety of tasks to determine which heuristic is best suited to be used for memory retrieval in a cognitive architecture. In this paper, we explore an initial set of memory retrieval heuristics that incorporate recency and frequency of memory access. We evaluate their effectiveness and scaling in the word sense disambiguation (WSD) task, an important and well-studied problem in the Natural Language Processing community (Navigli 2009). We are not attempting to solve the WSD problem per se, but instead (1) provide evidence that the WSD task is an appropriate benchmark for evaluating and comparing memory models and (2) evaluate the effectiveness of long-term memory systems that efficiently bias retrievals towards regularities of past access.

We begin by introducing the word sense disambiguation task, including an analysis of the WordNet (Miller 1995), SemCor (Miller et al. 1993), and Senseval (Edmonds and Kilgarriff 2002) data sets. We then present results of how baseline algorithms perform on this task, including the relative advantage of a memory-based approach that incorporates the recency and frequency of past word sense assignment. We then evaluate two models of memory bias that combine recency and frequency, including base-level activation (Anderson et al. 2004), a commonly used model based upon the rational analysis of memory (Anderson and Schooler 1991). We motivate and describe an approximation of the base-level model that theoretically scales to large bodies of knowledge and we present empirical results of an agent implemented in Soar (Laird 2008), which disambiguates words with this model. We conclude with a discussion of future work.

Word Sense Disambiguation

The English language contains *polysemous* words, those that have multiple, distinct meanings, or *senses*, which are interpreted differently based upon the context in which they occur. Consider the following sentences:

a. Deposit the check at the *bank*.
b. After canoeing, they rested at the *bank*.

The occurrences of the word *bank* in the two sentences clearly denote different meanings: 'financial institution' and 'side of a body of water,' respectively. Word sense disambiguation is the ability to identify the meaning of words in context in a computational manner (Navigli 2009). The task of WSD is critical to the field of NLP and various formulations have been studied for decades.

Our interest is in general competence across a variety of domains, and so we adopt the *all-words* WSD formulation, where the system is expected to disambiguate all open-class words in a text (i.e. nouns, verbs, adjectives, and adverbs). As input, the agent receives a sequence of sentences from a text, each composed of a sequence of words. However, as the focus of this work is memory, not unsupervised natural language processing, we supplement the input with the following two sources of additional structure, each of which is not uncommon in the WSD literature. First, each input word is correctly tagged with its contextually appropriate part-of-speech. Second, the agent is assumed to have access to a static machine-readable dictionary (MRD), such that each lexical word/part-of-speech pair in the input corresponds to a list of word senses within the MRD. For each sense, the MRD contains a textual definition, or *gloss*, and an annotation frequency from a training corpus. Thus, for each input word, the agent's task is to select an appropriate sense from the MRD, from which there may be multiple equally valid options for the given linguistic context.

Data Sets

To evaluate our work, we make use of three *semantic concordances*: each a textual corpus and lexicon linked such that every substantive word in the text is linked to its appropriate sense in the lexicon (Miller et al. 1993).

We begin with SemCor, the biggest and most widely used sense-tagged corpus, which includes 352 texts from the Brown corpus (Kucera and Francis 1967). We use the 186 Brown corpus files that have all open-class words annotated, which includes more than 185,000 sense references to version 3 of WordNet (Miller 1995). WordNet 3, the most utilized resource for WSD in English, includes more than 212,000 word senses. To prevent over-fitting in our results, we also utilize the Senseval-2 and Senseval-3 all-words corpora (Edmonds and Kilgarriff 2002), linked with WordNet 3. These data sets are nearly two orders of magnitude smaller than SemCor, comprising only 2,260 and 1,937 sense references, respectively.[1]

Task Analysis

In our formulation of the WSD task, an agent is provided a lexical word/part-of-speech pair and must select an appropriate sense, for the current context, from amongst a static list defined by the MRD. Let $\boldsymbol{s}$ represent the set of candidate senses from the MRD and let $\boldsymbol{a}$ represent the set of appropriate sense assignments in this context.

Given an arbitrary input word, an important measure that characterizes the difficulty of the task is $|\boldsymbol{s}|$, the cardinality of the set of candidate senses, referred to as *polysemy level* in the literature. However, as some open-class words in all of our data sets are tagged with more than one appropriate sense (0.33% in SemCor; 4.25% Senseval-2; 1.91% Senseval-3), it is also important to consider $|\boldsymbol{a}|$, the cardinality of the set of appropriate sense assignments. In this section, we characterize these measures across each of the data sets, and resolve the joint distribution of these values to derive expected task performance, given a random sense selection strategy.

To begin, we define a derived joint measure, *certainty*:

$$certainty = \frac{|\boldsymbol{a}|}{|\boldsymbol{s}|}$$

For a given lexical word/part-of-speech pair, since the set of suitable sense assignments is non-empty ($|\boldsymbol{a}| > 0$) and comprises a subset of the full set of candidate senses ($\boldsymbol{a} \subseteq \boldsymbol{s}$ and $|\boldsymbol{a}| \leq |\boldsymbol{s}|$), the range of *certainty* is (0, 1], where a value of 1 is, intuitively, unambiguous (any selection from amongst the candidate set is appropriate) and as a value becomes closer to 0, it becomes increasingly ambiguous (an appropriate selection is increasingly rare).

Given this nomenclature, Figure 1 represents the distribution of *certainty* within the SemCor data set, plotting the cumulative proportion of corpus words against

[1] The SemCor, Senseval-2, and Senseval-3 data sets are available to download at [http://www.cse.unt.edu/~rada/downloads.html]. WordNet is available to download at [http://wordnet.princeton.edu].

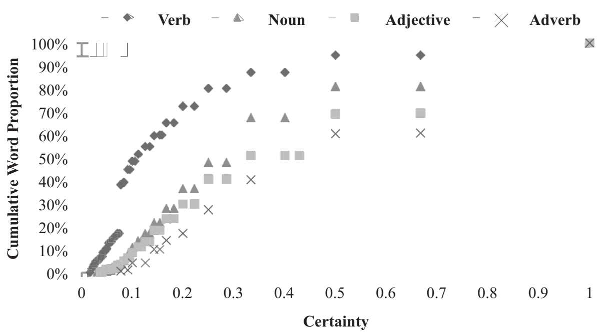

Figure 1. SemCor Cumulative Word Proportion vs. Certainty.

certainty. Both this plot and the descriptive analysis below aggregate the distribution with respect to part-of-speech; while our work does not investigate methods that are differentially sensitive to part-of-speech, we see distinctions in this distribution, which may be useful to future work. We only plot and textually analyze SemCor, the largest data set, but Table 1 summarizes pertinent outcomes for all three semantic concordances.

We first draw out the proportion of words with *certainty* value 1, those that require no disambiguation, by reading the second plotted point from the right for each part-of-speech. While for adverbs and adjectives this statistic is about 39% and 31%, respectively, for nouns and verbs it is about 19% and 5%. Across the entirety of SemCor, this statistic is 19.54%, establishing the absolute minimum for task performance in this formulation of WSD using SemCor. We next assess the median *certainty* for each part-of-speech by reading the x-axis as each part-of-speech intersects 50% on the y-axis. For adverbs, the median is $^1/_2$; for adjectives and nouns it is $^1/_3$; for verbs it is $^1/_9$; and the overall median *certainty* in SemCor is $^1/_4$. Finally, the average *certainty*, and thus the expected performance given a random selection strategy, is 38.73%.

Baseline Task Performance Results

To contextualize the performance results of a memory-based approach to the WSD task, we first implemented a set of baseline algorithms from the WSD literature. The results from these baselines are summarized in Table 2, including *random* selection, derived as expected performance in the previous section. Note that all algorithms we implement select a sense for all input words, and thus precision and recall are identical for all results, so we simply report them jointly as "task performance."

	SemCor	Senseval-2	Senseval-3
Unambiguous	19.54%	19.69%	15.02%
Median Certainty	0.25	0.25	0.2
Minimum Certainty	0.0169	0.0204	0.0169
Expected Performance	38.73%	40.56%	32.98%

Table 1. Semantic Concordance Task Analysis Summary.

	SemCor	Senseval-2	Senseval-3
Random	38.73%	40.56%	32.98%
Frequency Bias	76.39%	65.56%	65.41%
Lesk	63.40%	58.17%	53.46%
Simplified Lesk	65.52%	56.28%	53.66%

Table 2. Baseline Task Performance Results.

WordNet includes, for each word sense, an annotation frequency from the Brown corpus and the first baseline selection policy, *frequency bias*, exploits this information by choosing the most frequent sense for each lexical word/part-of-speech input pair. As the SemCor textual corpus is a subset of the Brown corpus, we expected this resource to be highly informative and, unsurprisingly, this algorithm yields nearly twice the performance of pure random selection. As the Senseval data sets were not derived from the Brown corpus, it is unsurprising that the absolute performance advantage of this heuristic is not as great when applied to these corpora. However, the relative improvement for Senseval-3 is greater than that of SemCor (98.33% versus 97.24%), which likely reflects the increased difficulty of Senseval-3 (see Table 1). Incorporating a frequency bias is not uncommon in the WSD literature, sometimes termed *commonest*, but it does tend to suffer, as found here, when the frequency distribution of the MRD is not representative of the corpus.

The remaining baselines were two variants of the Lesk algorithm for word sense disambiguation (Lesk 1986). The Lesk algorithm is a commonly used baseline metric (Vasilescu, Langlais, and Lapalme 2004) that assumes that words in a given "neighborhood" (such as a sentence) tend to share a common topic, and thus biases sense selection based upon shared terms in sense definitions and context. We explored the classic algorithm, with constant-sized neighborhood windows, as well as a "simplified" Lesk algorithm (Kilgarriff and Rosenzweig 2000), which defines word context as simply the terms in the neighborhood, as opposed to their definitions. The performance of the Lesk family of algorithms is known to be highly sensitive to the exact wording of sense definitions, and so it is common to supplement Lesk with heuristics and additional sources of semantic meaning, such as in Banerjee and Pedersen (2002). Thus, for both classic and simplified Lesk, we evaluated four supplemental heuristics: (1) the use of a *stop* list, which excludes definition terms that are common to the target language, such as "a" and "the"; (2) excluding example sentences from sense definitions, to avoid uninformative overlapping terms; (3) the use of the Porter Stemming (Porter 2006) algorithm to strip word suffixes, to facilitate overlap of words with common linguistic roots; and (4) a bias towards the corpus frequency information, applied in the case of equivalent sense evaluation. We evaluated the combinatorial set of these parameters across both algorithms. The maximum results for both classic and simplified algorithms occurred using the stop list, pruned definitions, and frequency bias, but not the Stemming algorithm. For the classic algorithm, we achieved maximum task performance with a neighborhood size of 2. However, as summarized in Table 2, the Lesk variants consistently underperformed, compared to *frequency bias*.

These baseline results are specific to our implementation and data sets, and are not intended for representation of or comparison to modern NLP techniques, but instead provide a reasonable baseline for our memory-based results.

Evaluating a Memory-based Approach

The main thrust of our research is to understand and develop memory mechanisms that effectively support agents in a variety of tasks, while computationally scaling as the amount of stored knowledge grows to be very large. Thus far in this paper we have characterized one important task, word sense disambiguation, including an analysis of WSD across three data sets and the types of performance we can expect from baselines that do not adapt their sense selection policy to the task instance. In this section, we describe and evaluate a simple approach to the WSD task that is available to those agents with a declarative memory.

Our approach is as follows: given a lexical word/part-of-speech input, the agent cues its memory for a sense that satisfies these constraints and selects the first retrieved result. We make two assumptions in our evaluation of this approach. First, the agent's long-term memory is preloaded with *at least* the contents of the data set's MRD, which affords the agent the potential to perform perfectly on the task, as it is not constrained by a limited vocabulary. In the case of our data sets, this assumption requires that the agent's memory mechanism scale computationally to at least the knowledge contained in WordNet. Later we return to explore the computational feasibility of this assumption.

Our second assumption is that immediately after the agent attempts to disambiguate a word, it is supplied with the set of appropriate senses for that input. This evaluation paradigm is important to isolate the effect of memory bias: it eliminates unintended divergent learning, which could occur sans truthful feedback and might obfuscate results.

In our evaluation, the agent's *a priori* knowledge and approach to the WSD task remains constant. However, we experimentally alter the agent's memory retrieval mechanism, changing how correct sense assignments in the past bias future retrievals. We investigate the degree to which the recency and frequency of past assignments inform future retrievals. Our set of experimental heuristics is motivated by the rational analysis of memory (Anderson and Schooler 1991), which demonstrated, across a set of linguistic tasks, reliable relationships between recency and frequency of past events and future memory accesses. When integrated within a memory system, these retrieval heuristics are independent of the WSD task, and can thus be applied and evaluated on additional WSD data sets, as well as tasks beyond WSD or linguistic settings.

We begin by evaluating recency and frequency biases individually, and then proceed to explore two candidate memory models that combine these heuristics. For each heuristic, we apply a greedy selection strategy, retrieving the word sense with the greatest bias value, and selecting randomly from amongst ties.

Unlike the non-adaptive WSD baselines, memory-endowed agents have the potential to improve with added corpus exposure, and thus we performed 10 sequential runs for each experimental condition, where the agent attempts to disambiguate the entirety of the corpus during each run. We report performance on the first, second, and tenth run of each of the data sets: the first run affords direct comparison to baseline results, the second run illustrates speed of learning, given relatively little corpus exposure, and the tenth speaks to asymptotic performance. Table 3 reports expected performance, as opposed to the sample average of individual probabilistic runs, and thus even small differences should be considered relevant.

Individual Memory Bias Task Results

The first heuristic we evaluate is *recency*, which biases ambiguous retrievals towards the last selected sense for each input. This bias, drawing on the *one sense per discourse* heuristic (Gale, Church, and Yarowsky 1992), performs well if the same sense is used repeatedly in immediate succession, but does not improve performance after the first full exposure to the data set, as demonstrated by no difference in performance between runs 2 and 10 in Table 3 (top), independent of the semantic concordance.

The next heuristic is *frequency*, which biases ambiguous retrievals towards the sense that has been retrieved most often. This bias performs well if particular senses of words are generally more common than others in a corpus, as opposed to being highly dependent upon sentence context.

Finally, to establish an upper bound on the degree to which recency and frequency can individually contribute to WSD performance, we implemented an *oracle* bias. For each input, this heuristic scores both the *recency* and *frequency* algorithms described above and returns the result from the two algorithms that provided the best score.

We draw two conclusions from Table 3. First, under the assumptions of our evaluation, the run 10 results suggest that memory retrieval agents perform better than the baselines from Table 1, with the potential for additional reasoning mechanisms to improve performance further. And second, nearly all memory bias results for run 10 are

	SemCor			Senseval-2			Senseval-3		
	Run 1	Run 2	Run 10	Run 1	Run 2	Run 10	Run 1	Run 2	Run 10
Recency	72.34%	74.43%	74.43%	61.74%	84.02%	84.02%	54.32%	79.29%	79.29%
Frequency	71.69%	76.21%	76.53%	62.13%	88.89%	89.28%	54.85%	84.30%	84.86%
Oracle	79.51%	83.77%	84.08%	63.68%	89.93%	90.23%	57.25%	86.23%	86.77%
Base-level	74.45%	77.90%	78.47%	62.17%	87.01%	88.47%	54.41%	82.19%	83.84%
Threshold	73.58%	77.82%	78.11%	62.34%	88.99%	89.39%	55.11%	84.50%	85.06%

Table 3. Memory Bias Task Performance Results by Semantic Concordance: Individual (top), Oracle (middle), and Joint (bottom).

better than all baselines in the respective test set (excl. SemCor/*Recency*). This suggests that history of sense assignment, with relatively little corpus exposure, yields a performance benefit in the WSD task, an advantage that is not dependent upon MRD definition quality (unlike Lesk).

Exploring Joint Memory Bias Models

The top two rows in Table 3 present evidence that recency and frequency of word sense assignment can individually yield performance benefits in the WSD task. Additionally, the relative gain in the *oracle* results (up to nearly 8% in SemCor) indicates that there is room for improvement. However, applying these findings to a memory system requires a model of how these heuristics combine in a task-independent fashion to bias memory retrieval (recall that the *oracle* algorithm is not possible to implement, as it requires the memory system to evaluate correct sense assignments during word sense selection). In this section, we explore memory bias models that jointly incorporate recency and frequency of memory access.

We first consider base-level activation (Anderson et al. 2004), which computes the activation bias of a memory using an exponentially decaying memory access history:

$$\ln\left(\sum_{j=1}^{n} t_j^{-d}\right)$$

where n is the number of accesses of the memory, t_j is the time since the j^{th} access, and d is a free decay parameter.

The second model, *threshold*, is a novel bias, which tests the hypothesis that the *recency* bias is proxy for meaning in local context (Gale et al. 1992): within a relatively small temporal frame, recent memories are informative of immediate context, and thus potentially more effective than globally learned frequency. Our model retrieves a memory according to the *recency* bias but, if the last memory access of that retrieval is older than a threshold, it resorts to the selection of the *frequency* heuristic.

We performed exploratory sweeps within each data set for model parameters and evaluated both models in the same fashion as the individual memory biases above (see Table 3, bottom two rows): both perform competitively, and *threshold* bests *recency* and *frequency* run 10 results across all data sets. The SemCor results used base-level decay parameter of 0.7 and threshold of 300, whereas both Senseval corpora used decay 0.4 and threshold 6.

Evaluating Memory Model Scalability

Our goal in this paper is to explore memory heuristics that are both effective and efficient in the WSD task. The results in Table 3 show that incorporating recency and frequency of past memory access to bias future retrievals, both individually and jointly, supports WSD task performance, as compared to non-memory baselines (see Table 2). We now evaluate the degree to which these heuristics can support efficient WSD memory retrievals.

The first attempt to access WordNet within a cognitive architecture illustrated the difficulty of the task: the declarative memory module of ACT-R could not encode even a third of WordNet. Douglass, Ball, and Rogers (2009) augmented ACT-R with a database and achieved 40 msec. retrievals, but with no memory bias. For context, prior work in dynamic environments, such as robotics and interactive computer games, has indicated that ~50 msec. is necessary for reactive control. Later, Derbinsky, Laird and Smith (2010) developed computational methods to efficiently support queries of large memory stores. They achieved sub-millisecond retrievals across all of WordNet, but supported only those bias models that are *locally efficient*: those for which updates can be completed in constant time and affect a constant number of elements.

The *recency* and *frequency* heuristics are *locally efficient*: both rely upon memory statistics that can be maintained incrementally and only update a single memory at a time. We implemented these heuristics within the semantic memory module of the Soar cognitive architecture (Laird 2008) and summarize run-time results across all data sets in Table 4. We report the maximum time in milliseconds, averaged over ten trials, for a Soar agent to retrieve a memory on a 2.8GHz Intel Core i7 processor, a measure of agent reactivity on modern commodity hardware. Both *recency* and *frequency* heuristics perform far faster than the 50 msec. threshold.

Unfortunately, neither joint model appears to be *locally efficient*. The *threshold* model conditionally requires two

	SemCor	Senseval-2	Senseval-3
Recency	0.85 msec.	0.82 msec.	0.80 msec.
Frequency	0.87 msec.	0.82 msec.	0.78 msec.

Table 4. Individual Bias Evaluation: Max. Query Time.

searches using two different heuristics. Consequently, we have not evaluated the scaling characteristics of this model. The base-level model includes a time decay component that changes frequently for all memories. We implemented a highly optimized version and were able to achieve 13.25 msec. retrievals in SemCor; however, this time is not bounded, growing with the store size. This heuristic, however, also affords a useful monotonicity: from the time that bias is calculated for a memory, that value is guaranteed to *over-estimate* the true bias until the memory is accessed again in the future. This characteristic affords a *locally efficient* approximation: the memory system updates activation only when a memory is accessed.

To evaluate this approximation, we consider query time, WSD task performance, which is comparable to results in Table 3, and model fidelity (a measure of interest to cognitive modelers), which we define as the smallest proportion of senses that the model selected within a run that matched the results of the original model.

We implemented this approximation within the semantic memory module of Soar and the results are summarized in Table 5 (d=0.5). To guarantee constant time bias calculation (Petrov 2006), we used a history of size 10. We also applied an incremental maintenance routine that updated memories that had not been accessed for 100 time steps, so as to avoid stale bias values. The query times across all data sets are far below our requirement of 50 msec. and an order of magnitude faster than the original. The run 10 WSD results were comparable to the original (see Table 3) and model fidelity was at or above 90% for all runs of all data sets. These results show that our approximation can support effective and efficient retrievals across large stores of knowledge.

Future Work

This paper has focussed on effectively and efficiently applying historical forms of memory bias in the word sense disambiguation task. We need to determine the extent to which these biases benefit agents in other tasks, especially those that are non-linguistic in nature. We also plan to explore the efficient incorporation of additional sources of bias, such as present context, which may lend functional benefits to WSD, amongst other problems. Finally, a great deal of work still needs to be done to understand how best to utilize high quality memory retrievals in context of other computational mechanisms in a cognitive architecture.

	SemCor	Senseval-2	Senseval-3
Max. Query Time	1.34 msec	1.00 msec	0.67 msec
Run 10 WSD Perf.	77.65%	89.03%	84.56%
Min. Model Fidelity	90.30%	95.70%	95.09%

Table 5. Base-level Approximation Evaluation.

References

Anderson, J. R., Schooler, L. J. 1991. Reflections of the Environment in Memory. *Psychological Science* 2 (6): 396-408.

Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., Qin, Y. 2004. An Integrated Theory of the Mind. *Psychological Review* 111 (4): 1036-1060.

Banerjee, S., Pedersen, T. 2002. An Adapted Lesk Algorithm for Word Sense Disambiguation Using WordNet. *Lecture Notes in Computer Science* 2276: 136-145.

Derbinsky, N., Laird, J. E., Smith, B. 2010. Towards Efficiently Supporting Large Symbolic Declarative Memories. In Proc. of the Tenth International Conference on Cognitive Modeling. Phil, PA.

Douglass, S., Ball, J. Rodgers, S. 2009. Large Declarative Memories in ACT-R. In Proc. of the Ninth International Conference on Cognitive Modeling. Manchester, UK.

Edmonds, P., Kilgarriff, A. 2002. Introduction to the Special Issue on Evaluating Word Sense Disambiguation Systems. *Natural Language Engineering* 8 (4): 279-291.

Gale, W. A., Church, K. W., Yarowsky, D. 1992. One Sense per Discourse. In Proc. of the Workshop on Speech and Natural Language, 233-237. Stroudsburg, PA.

Kilgarriff, A., Rosenzweig, J. 2000. English SENSEVAL: Report and Results. In Proc. of Language Resources and Evaluation.

Kucera, H., Francis, W. N. 1967. *Computational Analysis of Present-Day American English*. Brown Univ. Press, Prov., RI.

Laird, J. E. 2008. Extending the Soar Cognitive Architecture. In Proc. of the First Conference on Artificial General Intelligence, 224-235. Amsterdam, NL.: IOS Press.

Lesk, M. 1986. Automatic Sense Disambiguation Using Machine Readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone. In Proc. of Fifth Annual International Conference on Systems Documentation, 24-26. Toronto, Canada: ACM.

Miller, G. A. 1995. WordNet: A Lexical Database for English. *Communications of the ACM* 38 (11): 39-41.

Miller, G. A., Leacock, R., Tengi, R., Bunker, R. T. 1993. A Semantic Concordance. In Proc. of the Workshop on Human Language Technology, 303-308.

Navigli, R. 2009. Word Sense Disambiguation: A Survey. *ACM Computing Surveys* 41 (2): 1-69.

Nuxoll, A. M., Laird, J. E. 2007. Extending Cognitive Architecture with Episodic Memory. In Proc. of the Twenty-Second National Conference on Artificial Intelligence, 1560-1565. Vancouver, Canada.: AAAI Press.

Petrov, A. 2006. Computationally Efficient Approximation of the Base-Level Learning Equation in ACT-R. In Proc. of the Seventh International Conference on Cognitive Modeling. Trieste, Italy.

Porter, M. F. 2006. An Algorithm for Suffix Stripping. *Program: Electronic Library and Information Systems* 40 (3): 211-218.

Vasilescu, F., Langlais, P., Lapalme, G. 2004. Evaluating Variants of the Lesk Approach for Disambiguating Words. In Proc. of Language Resources and Evaluation.