A Perspective on Word Sense Disambiguation Methods and Their Evaluation

Philip Resnik
Dept. of Linguistics/UMIACS
University of Maryland
College Park, MD 20742
resnik@umiacs.umd.edu

David Yarowsky
Dept. of Computer Science/CLSP
Johns Hopkins University
Baltimore, MD 21218
yarowsky@cs.jhu.edu

Abstract

In this position paper, we make several observations about the state of the art in automatic word sense disambiguation. Motivated by these observations, we offer several specific proposals to the community regarding improved evaluation criteria, common training and testing resources, and the definition of sense inventories.

1 Introduction

Word sense disambiguation (WSD) is perhaps the great open problem at the lexical level of natural language processing. If one requires part-of-speech tagging for some task, it is now possible to obtain high performance off the shelf; if one needs morphological analysis, software and lexical data are not too hard to find. In both cases, performance of state-of-the-art systems is respectable, if not perfect, and the fundamentals of the dominant approaches (noisy channel models for tagging, two-level morphology) are by now well understood. For word sense disambiguation, we have a far longer way to go.

2 Observations

Observation 1. Evaluation of word sense disambiguation systems is not yet standardized. Evaluation of many natural language processing tasks including part-of-speech tagging and parsing has become fairly standardized, with most reported studies using common training and testing resources such as the Brown Corpus and Penn Treebank. Performance measures include a fairly well recognized suite of metrics including crossing brackets and precision/recall of non-terminal label placement. Several researchers (including Charniak, Collins and Magerman) have facilitated contrastive evaluation of their parsers by even training and testing on identical segments of the Treebank. Government funding agencies have accelerated this process, and even the task of anaphora resolution has achieved an evaluation standard under the MUC-6 program.

In contrast, most previous work in word sense disambiguation has tended to use different sets of polysemous words, different corpora and different evaluation metrics. Some clusters of studies have used common test suites, most notably the 2094-word *line* data of Leacock et al. (1993), shared by Lehman (1994) and Mooney (1996) and evaluated on the system of Gale, Church and Yarowsky (1992). Also, researchers have tended to keep their evaluation data and procedures somewhat standard across their own studies for internally consistent comparison. Nevertheless, there are nearly as many test suites as there are researchers in this field.

Observation 2. The potential for WSD varies by task. As Wilks and Stevenson (1996) emphasize, disambiguating word senses is not an end in itself, but rather an intermediate capability that is believed — but not yet proven — to improve natural language applications. It would appear, however, that different major applications of language differ in their potential to make use of successful word sense information. In information retrieval, even perfect word sense information may be of only limited utility, largely owing to the implicit disambiguation that takes place when multiple words within a query match multiple words within a document (Krovetz and Croft, 1992). In speech recognition, sense information is potentially most relevant in the form of word equivalence classes for smoothing in language models, but smoothing based on equivalence classes of contexts (e.g. (Bahl et al., 1983; Katz, 1987)) has a far better track record than smoothing based on classes of words (e.g. (Brown et al., 1992)).

The potential for using word senses in machine translation seems rather more promising. At the level of monolingual lexical information useful for high quality machine translation, for example, there is good reason to associate information about syntactic realizations of verb meanings with verb senses rather than verb tokens (Dorr and Jones, 1996a; 1996b). And of course unlike machine translation or speech recognition, the *human* process followed in completing the task takes explicit account of word senses, in that translators make use of correspondences in bilingual dictionaries organized according to word senses.

Observation 3. Adequately large sense-tagged data sets are difficult to obtain. Availability of data is a significant factor contributing to recent advances in part-of-speech tagging, parsing, etc. For the most successful approaches to such problems, correctly annotated data are crucial for training learning-based algorithms. Regardless of whether or not learning is involved, the prevailing evaluation methodology requires correct test sets in order to rigorously assess the quality of algorithms and compare their performance.

Unfortunately, of the few sense-annotated corpora currently available, virtually all are tagged collections of a single ambiguous word such as *line* or *tank*. The only broad-coverage annotation of all the words in a subcorpus is the WordNet semantic concordance (Miller et al., 1994). This represents a very important contribution to the field, providing the first large-scale, balanced data set for the study of the distributional properties of polysemy in English. However, its utility as a training and evaluation resource for supervised sense taggers is currently somewhat limited by its token-by-token sequential tagging methodology, yielding too few tagged instances of the large majority of polysemous words (typically fewer than 10 each), rather than providing much larger training/testing sets for a selected subset of the vocabulary. In addition, sequential annotation forces annotators to repeatedly refamiliarize themselves with the sense inventories of each word, slowing annotation speed and lowering intra- and inter-annotator agreement rates. Nevertheless, the WordNet semantic hierarchy itself is a central training resource for a variety of sense disambiguation algorithms and the existence of a corpus tagged in this sense inventory is a very useful complementary resource, even if small.

The other major potential source of sense-tagged data comes from parallel aligned bilingual corpora. Here, translation distinctions can provide a practical correlate to sense distinctions, as when instances of the English word *duty* translated to the French words *devoir* and *droit* correspond to the monolingual sense distinction between *duty*/OBLIGATION and *duty*/TAX. Current offerings of parallel bilingual corpora are limited, but as their availability and diversity increase they offer the possibility of limitless "tagged" training data without the need for manual annotation.

Given the data requirements for supervised learning algorithms and the current paucity of such data, we believe that unsupervised and minimally supervised methods offer the primary near-term hope for broad-coverage sense tagging. However, we see strong future potential for supervised algorithms using many types of aligned bilingual corpora for many types of sense distinctions.

Observation 4. The field has narrowed down approaches, but only a little. In the area of part-of-speech tagging, the noisy channel model dominates (e.g. (Bahl and Mercer, 1976; Jelinek, 1985; Church, 1988)), with transformational rule-based methods (Brill, 1993) and grammatico-statistical hybrids (e.g. (Tapanainen and Voutilainen, 1994)) also having a presence. Regardless of which of these approaches one takes, there seems to be consensus on what makes part-of-speech tagging successful:

- The inventory of tags is small and fairly standard.
- Context outside the current sentence has little influence.
- The within-sentence dependencies are very local.
- Prior (decontextualized) probabilities dominate in many cases.
- The task can generally be accomplished successfully using only tag-level models without lexical sensitivities besides the priors.
- Standard annotated corpora of adequate size have long been available.

Table 1: Some properties of the POS tagging task.

In contrast, approaches to WSD attempt to take advantage of many different sources of information (e.g. see (McRoy, 1992; Ng and Lee, 1996; Bruce and Wiebe, 1994)); it seems possible to obtain benefit from sources ranging from local collocational clues (Yarowsky, 1993) to membership in semantically or topically related word classes (Yarowsky, 1992; Resnik, 1993) to consistency of word usages within a discourse (Gale et al., 1992); and disambiguation seems highly lexically sensitive, in effect requiring specialized disambiguators for each polysemous word.

3 Proposals

Proposal 1. A better evaluation criterion. At present, the standard for evaluation of word sense disambiguation algorithms is the "exact match" criterion, or simple accuracy:

$$\% \text{ correct} = 100 \times \frac{\#\text{ exactly matched sense tags}}{\#\text{ assigned sense tags}}$$

Despite its appealing simplicity, this criterion suffers some obvious drawbacks. For example, consider the context:

... bought an **interest** in Lydak Corp. ... (1)

and assume the existence of 4 hypothetical systems that assign the probability distribution in Table 2 to the 4 major senses of *interest*.

Sense	System 1	2	3	4
(1) monetary (e.g. on a loan)	.47	.85	.28	1.00
(2) stake or share ⇐ *correct*	.42	.05	.24	.00
(3) benefit/advantage/sake	.06	.05	.24	.00
(4) intellectual curiosity	.05	.05	.24	.00

Table 2: Probability distributions assigned by four hypothetical systems to the example context (1) above.

Each of the systems assigns the *incorrect* classification (sense 1) given the correct sense 2 (*a stake or share*). However System 1 has been able to nearly rule out senses 3 and 4 and assigns reasonably high probability to the correct sense, but is given the same penalty as other systems that either have ruled out the correct sense (systems 2 and 4) or effectively claim ignorance (system 3).

If we intend to use the output of the sense tagger as input to another probabilistic system, such as a speech recognizer, topic classifier or IR system, it is important that the sense tagger yield probabilities with its classifications that are as accurate and robust as possible. If the tagger is confident in its answer, it should assign high probability to its chosen classification. If it is less confident, but has effectively ruled out several options, the assigned probability distribution should reflect this too.

A solution to this problem comes from the speech community, where *cross-entropy* (or its related measures perplexity and Kullback-Leibler distance) are used to evaluate how well a model assigns probabilities to its predictions. The easily computable formula for cross entropy is

$$-\frac{1}{N}\sum_{i=1}^{N}\log_2 \Pr_{\mathcal{A}}(cs_i|w_i, \text{context}_i)$$

where N is the number of test instances and $\Pr_{\mathcal{A}}$ is the probability assigned by the algorithm $\mathcal{A}$ to the correct sense, cs_i of polysemous word w_i in context$_i$. Crucially, given the hypothetical case above, the sense disambiguation algorithm in System 1 would get much of the credit for assigning high probability, even if not the highest probability, to the correct sense. Just as crucially, an algorithm would be penalized heavily for assigning very low probability to the correct sense,[1] as illustrated below:

Illustration of Cross-Entropy Calculation	System 1	2	3	4
$\Pr_{\mathcal{A}}(cs_i\|w_i, \text{context}_i)$	.42	.05	.24	.00
$-log_2\Pr_{\mathcal{A}}(cs_i\|w_i, \text{context}_i)$	1.25	4.32	2.05	∞

In aggregate, optimal performance is achieved under this measure by systems that assign as accurate a probability estimate as possible to their classifications, neither too conservative (System 3) nor too overconfident (Systems 2 and 4).

This evaluation measure does not necessarily obviate the exact match criterion, and the two could be used in conjunction with each other since they make use of the same test data. However, a measure based on cross-entropy or perplexity would provide a fairer test, especially for the common case where several fine-grained senses may be correct and it is nearly impossible to select exactly the sense chosen by the human annotator.

Finally, not all classification algorithms return probability values. For these systems, and for those that yield poorly estimated values, a variant of the cross entropy measure without the log term ($\frac{1}{N}\sum_{i=1}^{N}\Pr_{\mathcal{A}}(cs_i|w_i, \text{context}_i)$) can be used to measure improvement in restricting and/or roughly ordering the possible classification set without excessive penalties for poor or absent probability estimates. In the latter case, when the assigned tag is given probability 1 and all other senses probability 0, this measure is equivalent to simple % correct.

Proposal 2. Make evaluation sensitive to semantic/communicative distance between subsenses.

Current WSD evaluation metrics also fail to take into account semantic/communicative distance between senses when assigning penalties for incorrect labels. This is most evident when word senses are nested or arranged hierarchically, as shown in the example sense inventory for *bank* in Table 3.

[1] The extreme case of assigning 0 probability to the correct sense is given a penalty of ∞ by the cross-entropy measure.

I	*Bank*		- REPOSITORY
	I.1		Financial Bank
			I.1a - the institution
			I.1b - the building
	I.2		General Supply/Reserve/Inventory
II	*Bank*		- GEOGRAPHICAL
	II.1		Shoreline
	II.2		Ridge/Embankment
III	*Bank*		- ARRAY/GROUP/ROW

Table 3: Example sense inventory for *bank*

An erroneous classification between close siblings in the sense hierarchy should be given relatively little penalty, while misclassifications across homographs should receive a much greater penalty. The penalty matrix $distance(\text{subsense}_1, \text{subsense}_2)$ could capture simple hierarchical distance (e.g. (Resnik, 1995; Richardson et al., 1994)), derived from a single semantic hierarchy such as WordNet, or be based on a weighted average of simple hierarchical distances from multiple sources such as sense/subsense hierarchies in several dictionaries. A very simple example of such a distance matrix for the *bank* sense hierarchy is given in Table 4.

	I.1a	I.1b	I.2	II.1	II.2	III
I.1a	0	1	2	4	4	4
I.1b	1	0	2	4	4	4
I.2	2	2	0	4	4	4
II.1	4	4	4	0	1	4
II.2	4	4	4	1	0	4
III	4	4	4	4	4	0

Table 4: Example distance/cost matrix for *bank*

Penalties could also be based on general pairwise *functional communicative distance*: errors between subtle sense differences would receive little penalty while gross errors likely to result in misunderstanding would receive a large penalty. Such communicative distance matrices could be derived from several sources. They could be based on psycholinguistic data, such as experimentally derived estimates of similarity or confusability (Miller and Charles, 1991; Resnik, 1995). They could be based on a given task, e.g. in speech synthesis only those sense distinction errors corresponding to pronunciation distinctions (e.g. bass-/bæs/ vs. bass-/beɪs/) would be penalized. For the machine-translation application, only those sense differences lexicalized differently in the target language would be penalized, with the penalty proportional to communicative distance.[2] In general such a distance matrix could support arbitrary communicative cost/penalty functions, dynamically changible according to task.

[2]Such distance could be based on the weighted % of all languages that lexicalize the two subsenses differently.

There are several ways in which such a (hierarchical) distance penalty weighting could be utilized along with the cross-entropy measure. The simplest is to minimize the mean distance/cost between the assigned sense (as_i) and correct sense (cs_i) over all N examples as an independent figure of merit:

$$\frac{1}{N}\sum_{i=1}^{N}\text{distance}(cs_i, as_i)$$

However, one could also use a metric such as the following that measures efficacy of probability assignment in a manner that penalizes probabilities assigned to incorrect senses weighted by the communicative distance/cost between that incorrect sense and the correct one:

$$\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{S_i}\text{distance}(cs_i, s_j) \times \Pr_{\mathcal{A}}(s_j|w_i, \text{context}_i)$$

where for any test example i, we consider all S_i senses (s_j) of word w_i, weighting the probability mass assigned by the classifier $\mathcal{A}$ to incorrect senses ($\Pr_{\mathcal{A}}(s_j|w_i, \text{context}_i)$) by the communicative distance or cost of that misclassification.[3]

Note that in the special case of sense tagging without probability estimates (all are either 0 or 1), this formula is equivalent to the previous one (simple mean distance or cost minimization).

Proposal 3. A framework for common evaluation and test set generation. Supervised and unsupervised sense disambiguation methods have different needs regarding system development and evaluation. Although unsupervised methods may be evaluated (with some limitations) by a sequentially tagged corpus such as the WordNet semantic concordance (with a large number of polysemous words represented but with few examples of each), supervised methods require much larger data sets focused on a subset of polysemous words to provide adequately large training and testing material. It is hoped that US and international sources will see fit to fund such a data annotation effort. To facilitate discussion of this issue, the following is a proposed framework for providing this data, satisfying the needs of both supervised and unsupervised tagging research.

[3]Although this function enumerates over all S_i senses of w_i, because $\text{distance}(cs_i, cs_i) = 0$ this function only penalizes probability mass assigned to incorrect senses for the given example.

1. Select/Collect a very large (e.g., N = 1 billion words), diverse unannotated corpus.

2. Select a sense inventory (e.g. WordNet, LDOCE) with respect to which algorithms will be evaluated (see Proposal 4).

3. Pick a subset of $R < N$ (e.g., 100M) words of unannotated text, and release it to the community as a training set.

4. Pick a smaller subset of $S < R < N$ (e.g., 10M) words of text as the source of the test set. Generate the test set as follows:

   (a) Select a set of M (e.g., 100) ambiguous words that will be used as the basis for the evaluation, *without* telling the research community what those words will be.
   (b) For each of the M words, annotate all available instances of that word in the test corpus. Make sure each annotator tags all instances of a *single* word, e.g. using a concordance tool, as opposed to going through the corpus sequentially.
   (c) For each of the M words, compute evaluation statistics using individual annotators against other annotators.
   (d) For each of the M words, go through the cases where annotators disagreed and make a consensus choice, by vote if necessary.

5. Instruct participants in the evaluation to "freeze" their code; that is, from this point on no changes may be made.

6. Have each participating algorithm do WSD on the full S-word test corpus.

7. Evaluate the performance of each algorithm considering *only* instances of the M words annotated as the basis for the evaluation. Compare exact match, cross-entropy, and inter-judge reliability measures (e.g. Cohen's κ) using annotator-vs-annotator results as an upper bound.

8. Release this year's S-word test corpus as a *development* corpus for those algorithms that require supervised training, so they can participate from now on, being evaluated in the future via cross-validation.

9. Go back to Step 3 for next year's evaluation.

There are a number of advantages to this paradigm, in comparison with simply trying to annotate large corpora with word sense information. First, it combines an emphasis on broad coverage with the advantages of evaluating on a limited set of words, as is done traditionally in the WSD literature. Step 4a can involve any form of criteria (frequency, level of ambiguity, part of speech, etc.) to narrow down to set of candidate words, and then employ random selection among those candidates. At the same time, it avoids a common criticism of studies based on evaluating using small sets of words, namely that there is not enough attention being paid to scalability. In this evaluation paradigm, algorithms must be able to sense tag *all* words in the corpus meeting specified criteria, because there is no way to know in advance which words will be used to compute the figure(s) of merit.

Second, the process avoids some of the problems that arise in using exhaustively annotated corpora for evaluation. By focusing on a relatively small set of polysemous words, much larger data sets for each can be produced. This focus will also allow more attention to be paid to selecting and vetting comprehensive and robust sense inventories, including detailed specifications and definitions for each. Furthermore, by having annotators focus on one word at at time using concordance software, the initial level of consistency is likely to be far higher than that obtained by a process in which one jumps from word to word by going sequentially through a text, repeatedly refamiliarizing oneself with different sense inventories at each word. Finally, by computing inter-annotator statistics blindly and *then* allowing annotators to confer on disagreements, a cleaner test set can be obtained without sacrificing trustworthy upper bounds on performance.

Third, the experience of the Penn Treebank and other annotation efforts has demonstrated that it is difficult to select and freeze a comprehensive tag set for the entire vocabulary in advance. Studying and writing detailed sense tagging guidelines for each word is comparable to the effort required to create a new dictionary. By focusing on only 100 or so polysemous words per evaluation, the annotating organization can afford to do a multi-pass study of and detailed tagging guidelines for the sense inventory present in the data for each target word. This would be prohibitively expensive to do for the full English vocabulary. Also, by utilizing different sets of words in each evaluation, such factors as the level of detail and the sources of the sense inventories may change without worrying about maintaining consistency with previous data.

Fourth, both unsupervised and supervised WSD algorithms are better accommodated in terms of the amount of data available. Unsupervised algorithms

Target Word	WordNet Sense #	English description	Spanish	French	German	Italian	Japanese
interest (noun)	1	monetary (e.g. on loan)	interés, rédito	intérêt	Zinsen	interesse	rishi, risoku
	2	stake/share	interés, participación	intérêt participation	Anteil	interesse	riken
	3,4	intellectual curiosity	interés,	intérêt	Interesse	interesse	kanshin, kyōmi
	5	benefit, advantage	provecho, interés, beneficio	intérêt	Interesse	interesse	rieki
drug (noun)	1a	medicine	medicamento, droga	medicament	Medikament, Arzheimittel	medicina	kusuri
	1b	narcotic	narcótica droga	drogue	Drogue, Rauschgift	droga	mayaku
bank	1	shoreline	ribera, orilla	banc, rive	Ufer	sponda,riva	kishi
(noun)	2	embankment	loma, cuesta	talus, terasse	Erdwall	muccio	teibō
	3	financial inst.	banco	banque	Bank	banca	ginkō
	4	supply/reserve	banco	banque	Bank	banca	ginkō
	5	bank building	banco	banque	Bank	banca	ginkō
	6	array/row	hilera, batería	rang, batterie	Reihe	batteria	retsu
fire (t. verb)	1	dismiss from job	despedir, echar	renvoyer	feuern	licenziare	kubi ni shimasu
	2	arouse, provoke	excitar, enardecer	enflammer, animer	beflügeln entzünden	accendere infiammare	kōfun saseru
	4	discharge weapn	disparar	lâcher	abfeuern	sparare	happō s.
	5	bake pottery	cocer	cuire	brennen	cuocere	yaku

Table 5: Mapping between cross-linguistic sense labels and established lexicons

can be given very large quantities of training data: since they require no annotation the value of R can be quite large. And although supervised algorithms are typically plagued by sparse data, this approach will yield much larger training and testing sets per word, facilitating the exploration and development of data intensive supervised algorithms.

Proposal 4. A multilingual sense inventory for evaluation. One of the most fraught issues in applied lexical semantics is how to define word senses. Although we certainly do not propose a definitive answer to that question, we suggest here a general purpose criterion that can be applied to existing sources of word senses in a way that, we suggest, makes sense both for target applications and for evaluation, and is compatible with the major sources of available training and test data.

The essence of the proposal is to restrict a word sense inventory to those distinctions that are typically *lexicalized cross-linguistically.* This cuts a middle ground between restricting oneself to homographs within a single language, which tends toward a very coarse-grained distinction, and an attempt to express all the fine-grained distinctions made in a language, as found in monolingual dictionaries. In practice the idea would be to define a set of target languages (and associated bilingual dictionaries), and then to require that any sense distinction be realized lexically in a minimum subset of those languages. This would eliminate many distinctions that are arguably better treated as regular polysemy. For example, *table* can be used to refer to both a physical object and a group of people:

(1) a. The waiter put the food on the table.
b. Then he told another table their food was almost ready.
c. He finally brought appetizers to the table an hour later.

In German the two meanings can actually be lexicalized differently (*Tisch* vs. *Tischrunde*). However, as such sense distinctions are typically conflated into a single word in most languages, and because even German can use *Tisch* in both cases, one could plausibly argue for a common sense inventory for evaluation that conflates these meanings.

A useful reference source for both training and evaluation would be a table linking sense numbers in established lexical resources (such as WordNet or LDOCE) with these crosslinguistic translation distinctions. An example of such a map is given in Table 5. A comparable mapping could readily be extracted semi-automatically from bilingual dictionaries or from the EuroWordNet effort (Bloksma et al., 1996) which provides both semantic hierarchies and interlingual node linkages, currently for the languages Spanish, Italian, Dutch and English. We note that the table follows many lexical resources, such as the original WordNet, in being organized at the top level according to parts of speech. This seems to us a sensible approach to take for sense inventories, especially in light of Wilks and Stevenson's (1996) observation that part-of-speech tagging accomplishes much of the work of semantic disambiguation, at least at the level of homographs.

Although cross-linguistic divergence is a significant problem, and 1-1 translation maps do not exist for all sense-language pairs, this table suggests how *multiple* parallel bilingual corpora for different language pairs can be used to yield sets of training data covering different subsets of the English sense inventory, that in aggregate may yield tagged data for all given sense distinctions when any one language alone may not be adequate.

For example, a German-English parallel corpus could yield tagged data for Senses 1 and 2 for *interest*, and the presence of certain Spanish words (provecho, beneficio) aligned with *interest* in a Spanish-English corpus will tag some instances of Sense 5, with a Japanese-English aligned corpus potentially providing data for the remaining sense distinctions. In some cases it will not be possible to find any language (with adequate on-line parallel corpora) that lexicalize some subtle English sense distinctions differently, but this may be evidence that the distinction is regular or subtle enough to be excluded or handled by other means.

Note that Table 5 is not intended for direct use in machine translation. Also note that when two word senses are in a cell they are not necessarily synonyms. In some cases they realize differences in meaning or contextual usage that are salient to the target language. However, at the level of sense distinction given in the table, they correspond to the same word senses in English and the presence of either in an aligned bilingual corpus will indicate the same English word sense.

Monolingual sense tagging of another language such as Spanish would yield a similar map, such as distinguishing the senses of the Spanish word *dedo*, which can mean 'finger' or 'toe'. Either English or German could be used to distinguish these senses, but not Italian or French, which share the same sense ambiguity.

It would also be helpful for Table 5 to include alignments between multiple monolingual sense representations, such as COBUILD sense numbers, LDOCE tags or WordNet synsets, to support the sharing and leveraging of results between multiple systems. This brings to the fore an existing problem, of course: different sense inventories lead to different algorithmic biases. For example, WordNet as a sense inventory would tend to bias an evaluation in favor of algorithms that take advantage of taxonomic structure; LDOCE might bias in favor of algorithms that can take advantage of topical/subject codes, and so forth. Unfortunately we have no solution to propose for the problem of which representation (if any) should be the ultimate standard, and leave it as a point for discussion.

4 Conclusions

The most important of our observations about the state of the art in word sense disambiguation is that it is still a hard, open problem, for which the field has not yet narrowed much. We have made several suggestions that we believe will help assess progress and advance the state of the art. In summary:

- We proposed that the accepted standard for WSD evaluation include a cross-entropy like measure that tests the accuracy of the probabilities assigned to sense tags and offers a mechanism for assigning partial credit.
- We suggested a paradigm for common evaluation that combines the benefits of traditional "interesting word" evaluations with an emphasis on broad coverage and scalability.
- We outlined a criterion that should help in determining a suitable sense inventory to use for comparison of algorithms, compatible with both hierarchical sense partitions and multilingually motivated sense distinctions.

References

L. Bahl and R. Mercer. 1976. Part-of-speech assignment by a statistical decision algorithm. In *International Symposium on Information Theory*, Ronneby, Sweden.

L. Bahl, F. Jelinek, and R. Mercer. 1983. A maximum likelihood approach to continuous speech recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-5(2):179–190.

L. Bloksma, P. Díez-Orzas and P. Vossen. 1996. User Requirements and Functional Specification of the EuroWordNet Project. `http://www.let.uva.nl/~ewn`.

E. Brill. 1993. *A Corpus-Based Approach to Language Learning*. Ph.D. thesis, Computer and Information Science, University of Pennsylvania.

P. Brown, V. Della Pietra, P. deSouza, J. Lai, and R. Mercer. 1992. Class-based n-gram models of natural language. *Computational Linguistics*, 18(4):467–480.

R. Bruce and J. Wiebe. 1994. Word-sense disambiguation using decomposable models. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, Las Cruces.

K. Church. 1988. A stochastic parts program and noun phrase parser for unrestricted texts. In *Proceedings of the Second Conference on Applied Natural Language Processing*, Austin, Texas.

B. Dorr and D. Jones. 1996a. Acquisition of semantic lexicons: Using word sense disambiguation to improve precision. In *Proceedings of the SIGLEX Workshop on Breadth and Depth of Semantic Lexicons*, Santa Cruz, CA.

B. Dorr and D. Jones. 1996b. Role of word sense disambiguation in lexical acquisition: Predicting semantics from syntactic cues. In *Proceedings of the International Conference on Computational Linguistics*, Copenhagen, Denmark.

W. Gale, K. Church, and D. Yarowsky. 1992. One sense per discourse. *Proceedings of the 4th DARPA Speech and Natural Language Workshop.*

W. Gale, K. Church, and D. Yarowsky. A method for disambiguating word senses in a large corpus. *Computers and the Humanities*, 26:415–439, 1992.

F. Jelinek. 1985. Markov source modeling of text generation. In J. Skwirzinski, editor, *Impact of Processing Techniques on Communication*. Dordrecht.

S. Katz. 1987. Estimation of probabilities from sparse data for the language model component of a speech recognizer. *IEEE Transactions on Acoustics, Speech and Signal Processing*, ASSP-35(3):400–401.

R. Krovetz and W. B. Croft. 1992. Lexical ambiguity and information retrieval. *ACM Transactions on Information Systems*, 10(2):115–141.

C. Leacock, G. Towell and E. Voorhees. 1993. Corpus-based statistical sense resolution. In *Proceedings, ARPA Human Language Technology Workshop*, pp. 260–265, Plainsboro, NJ.

J. Lehman. 1994. Toward the essential nature of statistical knowledge in sense resolution. In *Proceedings of the Twelfth National Conference on Artificial Intelligence*, pp. 734–471.

R. Mooney. 1996. Comparative experiments on disambiguating word senses: An illustration of the role of bias in machine learning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Philadelphia.

S. McRoy. 1992. Using multiple knowledge sources for word sense discrimination. *Computational Linguistics*, 18(1):1–30.

G. Miller and W. Charles. 1991. Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1):1–28.

G. Miller, M. Chodorow, S. Landes, C. Leacock, and R. Thomas. 1994. Using a semantic concordance for sense identification. In *Proceedings of the ARPA Human Language Technology Workshop*, San Francisco. Morgan Kaufmann.

H. Ng and H. Lee. 1996. Integrating multiple knowledge sources to disambiguate word sense: An exemplar-based approach. In *Proceedings of the 34th Annual Meeting of the Society for Computational Linguistics*, pp. 40-47, Santa Cruz, CA.

P. Resnik. 1993. *Selection and Information: A Class-Based Approach to Lexical Relationships*. Ph.D. thesis, University of Pennsylvania. (`ftp://ftp.cis.upenn.edu/pub/ircs/tr/93-42.ps.Z`).

P. Resnik. 1995. Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI-95)*. (cmp-lg/9511007).

R. Richardson, A. Smeaton, and J. Murphy. 1994. Using WordNet as a knowledge base for measuring semantic similarity between words. Working Paper CA-1294, Dublin City University, School of Computer Applications, Dublin, Ireland. `ftp://ftp.compapp.dcu.ie/pub/w-papers/1994/CA1294.ps.Z`.

P. Tapanainen and A. Voutilainen. 1994. Tagging accurately – don't guess if you know. In *Proceedings of ANLP '94*.

Y. Wilks and M. Stevenson. 1996. The grammar of sense: Is word-sense tagging much more than part-of-speech tagging? cmp-lg/9607028.

D. Yarowsky. 1992. Word-sense disambiguation using statistical models of Roget's categories trained on large corpora. In *Proceedings of COLING-92*, pp. 454–460, Nantes, France.

D. Yarowsky. 1993. One sense per collocation. *Proceedings of the ARPA Human Language Technology Workshop*, Morgan Kaufmann, pp. 266-271.