Research Paper ■

A Multi-aspect Comparison Study of Supervised Word Sense Disambiguation

HONGFANG LIU, PHD, VIRGINIA TELLER, PHD, CAROL FRIEDMAN, PHD

Abstract **Objective:** The aim of this study was to investigate relations among different aspects in supervised word sense disambiguation (WSD; supervised machine learning for disambiguating the sense of a term in a context) and compare supervised WSD in the biomedical domain with that in the general English domain.

Methods: The study involves three data sets (a biomedical abbreviation data set, a general biomedical term data set, and a general English data set). The authors implemented three machine-learning algorithms, including (1) naïve Bayes (NBL) and decision lists (TDLL), (2) their adaptation of decision lists (ODLL), and (3) their mixed supervised learning (MSL). There were six feature representations (various combinations of collocations, bag of words, oriented bag of words, etc.) and five window sizes (2, 4, 6, 8, and 10).

Results: Supervised WSD is suitable only when there are enough sense-tagged instances with at least a few dozens of instances for each sense. Collocations combined with neighboring words are appropriate selections for the context. For terms with unrelated biomedical senses, a large window size such as the whole paragraph should be used, while for general English words a moderate window size between 4 and 10 should be used. The performance of the authors' implementation of decision list classifiers for abbreviations was better than that of traditional decision list classifiers. However, the opposite held for the other two sets. Also, the authors' mixed supervised learning was stable and generally better than others for all sets.

Conclusion: From this study, it was found that different aspects of supervised WSD depend on each other. The experiment method presented in the study can be used to select the best supervised WSD classifier for each ambiguous term.

■ **J Am Med Inform Assoc.** 2004;11:320–331. DOI 10.1197/jamia.M1533.

Word sense disambiguation (WSD) is the problem of tagging the appropriate sense of a given word in a context. Resolving sense ambiguity is one of the most important problems in natural language processing (NLP) and is essential for any kind of text-understanding task such as information extraction, information retrieval, or message understanding.[1,2] Despite the wide range of approaches investigated, including expert rules and supervised or unsupervised machine-learning techniques, currently there is no large-scale, broad-coverage, and highly accurate WSD system.[3]

One of the encouraging approaches to WSD is supervised machine learning.[2] Given an ambiguous word *W*, a supervised WSD classifier is obtained by applying supervised machine learning on a collection of sense-tagged instances of *W*, called a sense-tagged corpus of *W* STC(*W*), in which the sense of *W* in each instance has been tagged. Supervised approaches tend to achieve better performance than other WSD approaches.[4,5] However, supervised WSD suffers from the lack of a large, broad-coverage, sense-tagged corpus. Currently, there are two lines of research tackling this problem[6,7]: (1) design efficient sampling methods to lower the cost of sense tagging or (2) use a machine-readable dictionary (MRD) and a large collection of raw text to obtain a raw sense-tagged corpus. In several previous reports,[8,9] we have shown that methods based on MRD could be used to obtain a large, sense-tagged corpus for ambiguous biomedical abbreviations. We also did a comparison study in one report,[8] but it was limited (with one data set specialized in the medical reports domain, a couple of existing machine-learning algorithms, and a few feature representations that did not include collocations). The main focus of that report was to propose an unsupervised WSD method and not to compare different supervised WSD classifiers. Results showed that further exploration of supervised WSD was warranted. In this report, we investigated the relations among different aspects of supervised WSD and compared WSD in the biomedical domain with that in the general English domain.

Here, we first present background and related work about supervised WSD. We then describe our experiment on supervised WSD involving several aspects: data set, feature

Affiliations of the authors: Department of Information Systems, University of Maryland at Baltimore County, Baltimore, MD (HL); Department of Computer Science, Hunter College, City University of New York, New York, NY (VT); Department of Biomedical Informatics, Columbia University, New York, NY (CF).

Supported in part by NLM grant LM06274 and NSF grant NSF 0312250.

Correspondence and reprints: Hongfang Liu, PhD, Department of Information Systems, University of Maryland at Baltimore County, 1000 Hilltop Circle, Baltimore, MD 21250; e-mail: <hfliu@umbc.edu>.

Received for publication: 01/09/04; accepted for publication: 03/16/04.

representation, window size, and supervised machine-learning algorithm. Finally, the results are presented and discussed.

Construction of Supervised WSD Classifiers

Figure 1 shows the process of constructing a supervised WSD classifier for *W* given a sense-tagged corpus, STC(*W*). The input to the process is STC(*W*), and the output is a WSD classifier, which can disambiguate *W*. The first component transfers each instance in STC(*W*) to a feature vector. The second component uses a supervised learning algorithm to learn the disambiguation knowledge that forms a WSD classifier for *W*.

Feature Representation

Supervised WSD approaches require transforming each sense-tagged instance into a feature vector. Different kinds of features have been exploited.[2]

- **Local Co-occurring Words:** Co-occurring words in the context of an ambiguous word *W* in a fixed window are critical to WSD. For example, in the sentence "A spokesman said Healthvest has paid two of the three banks it owed interest in October," words such as *paid* and *banks* tend to indicate that in this sentence *interest* has the sense money paid for the use of money and not other senses, such as readiness to give attention or activity that one gives attention to (refer to Table 3 for sense definitions for *interest*).
- **Collocations:** A collocation refers to a short sequence of ordered words that occur together more often than by chance. It is also important for the sense determination of *W*. For example, in the phrase "the interest of," the sense of *interest* is the advantage, advancement, or favor sense of *interest* even though words *the* and *of* are usually included in the stop word list for word indexing of information retrieval systems.
- **Derived Features:** Derived features are obtained from words surrounding *W* in a window of a fixed size, taking into consideration the orientation and/or distance from *W*. A derived feature may also contain further linguistic knowledge, such as part of speech (POS) tags, semantic categories (e.g., classes in *Roget's Thesaurus*) or stemming techniques, which assign a common feature to inflected forms of a root (e.g., *discharged*, *discharging*, and *discharges* are treated as the same feature *discharge*).

Supervised Learning Algorithms

Several supervised learning algorithms have been adapted to built WSD classifiers: naïve Bayes learning,[4] neural networks,[10,11] decision list,[12] instance-based learning,[13,14] and inductive logic programming.[15] In the following, we provide background information about several supervised learning algorithms that we implemented or adapted. Readers can refer to our previous studies,[8,9] method section, and the survey paper of Marquez[16] for further detail. Note that naïve Bayes learning, decision list learning, and instance-based learning were already adapted for WSD.[8] We found that instance-based learning took a long time to process while the performance of naïve Bayes learning and decision list learning was not distinguishable given the same feature representation and window size.

Naïve Bayes Learning (NBL)[17] is widely used in machine learning due to its efficiency and its ability to combine evidence from a large number of features. An NBL classifier chooses the category with the highest conditional probability for a given feature vector, while the computation of conditional probabilities is based on the naïve Bayes assumption that the presence of one feature is independent of another given the category. Training a naïve Bayes classifier consists of estimating the prior probabilities for different categories as well as the probabilities of each category for each feature.

The **Decision List Method (DLL)**[12] is equivalent to simple case statements in most programming languages. In a DLL classifier, a sequence of tests is applied to each feature vector. If a test succeeds, then the sense associated with that test is returned. If the test fails, then the next test in the sequence is applied. This continues until the end of the list, where a default test simply returns the majority sense. Learning a decision list classifier consists of generating and ordering individual tests based on the characteristics of the training data.

Instance-based Learning (IBL)[18] has appeared in several areas with different names: exemplar-based, case-based, and memory-based. It is a form of supervised learning from instances based on keeping a full memory of training instances and classifying new instances using the most similar training instances. Instance-based classifiers can be used without training if the similarity measure between two instances is local, i.e., the similarity between two instances is completely determined by their associated feature vectors. Sometimes instance-based classifiers include a training phase, in which a set of representative instances (to reduce the number of training instances presented to the classifier) and/or a similarity measure between two instances (to include distributional information in the similarity measure) are

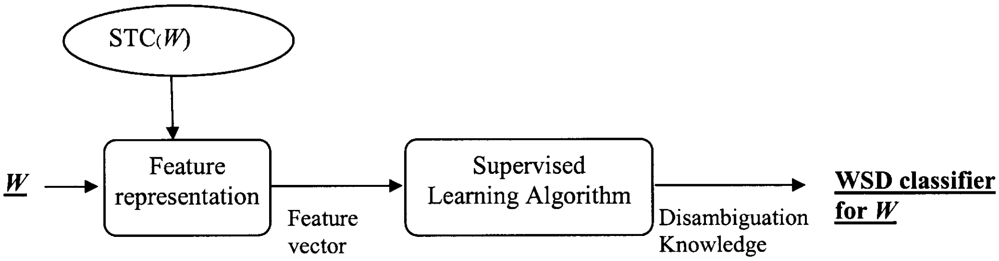

Figure 1. The processing phases for constructing a WSD classifier for *W*.

chosen. A critical component of instance-based classifiers is the similarity measure.

Related Work and Current Status

There are several studies on supervised WSD. Bruce and Wiebe [4] applied the Bayesian algorithm and chose features based on their "informative" nature. They tested their work on the *interest* corpus and achieved a precision of 79%. Towell and Voorhees[11] constructed a WSD classifier that combined the output of a neural network that learns topical context with the output of a network that learns local context to distinguish among the senses of highly ambiguous words. The precision of the classifier was tested on three words, the noun *line*, the verb *serve*, and the adjective *hard*; the classifier had an average precision of 87%, 90%, and 81%, respectively. The WSD system of Yarowsky[12] used the decision list method on features that consisted of both POS tags and oriented distances of the surrounding words. He claimed that the system had a precision of 99% when evaluated automatically for the accent restoration task in Spanish and French. Ng and Lee[13,14] described a WSD system that uses the instance-based method with multiple kinds of features. An ambiguous term in an instance was assigned to the sense of its most similar instance in the training set in the initial version; later the sense was determined by a fixed number of the most similar instances. They reported the systems had a precision of 87.4% for the *interest* corpus[13] and 58.7% and 75.2% for the two test sets derived from the Defence of Science Organization in Singapore (DSO) corpus.[14]

Currently, there is little agreement on feature representation, preference of window sizes (i.e., the number of neighboring words that should be included as sources for deriving features), and the best choice of supervised learning algorithms for WSD applications. It is generally believed that nouns require a larger window than verbs.[19] Larger values of window sizes capture dependencies at longer range but also dilute the effect of the words closer to the term. Leacock et al.[20] tested a window size of 50, while Yarowsky[12] argued that a small window size of 3 or 4 had better performance. A small window size has the advantage of requiring less system space and running time. Leacock et al.[20] showed that various supervised learning algorithms tended to perform roughly the same when given the same evidence. Mooney[21] reported that naïve Bayes learning gave the best performance on disambiguating the *line* corpus among seven learning algorithms tested. Ng and Lee[13] reported that the performance of instance-based classifiers was comparable to naïve Bayes classifiers. Yarowsky[12] stated that decision list classifiers had at least as good performance as naïve Bayes classifiers with the same evidence and also had the advantage of easy interpretation, easy modification, and easy implementation.

Most previous comparison studies of supervised WSD classifiers isolated different aspects, such as data set, feature representation, window size, and supervised learning algorithm. In our previous study,[8] we used only one data set specializing in the biomedical domain. This research is a multi-aspect comparison study of supervised WSD aimed at investigating the relations among these aspects. In addition, this report adapted two new techniques of machine-learning algorithms for the purpose of WSD. A new technique involved combining NBL with IBL, and a new technique consisted of a modification to DLL.

Experimental Methods

The experiment involved four aspects: data set, feature representation, window size, and machine-learning algorithm. Each aspect is described in more detail below.

Data Sets

There are three data sets used in the experiment. The first data set, **ABBR**, contains 15 ambiguous abbreviations. The gold standard instances for **ABBR** were constructed utilizing the fact that authors sometimes define abbreviations when they are first introduced in documents using parenthesized expressions [e.g., *Androgen therapy prolongs complete remission in acute myeloblastic leukemia (AML)*] and that these abbreviations have the same senses within the same documents (for details, refer to Liu et al.[9]). Note that in Liu et al.,[9] there were 34 ambiguous three-letter abbreviations. We excluded 18 of them in which the majority sense in the gold standard set was more than 90% to avoid the comparison biases. The abbreviation EMG was also excluded because three of four associated senses (i.e., *exomphalos macroglossia gigantism*, *electromyography, electromyographs*, *electromyogram*) were closely related. Additionally, for terms with more than 3,000 sense-tagged instances, we randomly chose about 3,000 instances to avoid the bias these terms could introduce when computing the overall performance. Table 1 gives details about the set.

The second set, **MED**, contains 22 ambiguous general biomedical terms. The gold standard instances were derived manually by Weeber et al.[22] There were 50 terms in that study. We excluded 12 that Weeber et al. considered problematic, as well as 16 terms in which the majority sense occurred with over 90% of instances. Note that in the study by Weeber et al.[22] an occurrence of an inflected variant (e.g., *discharged* is considered as an occurrence of *discharge*) of an ambiguous word was considered to be an ambiguous occurrence. In our study, however, only occurrences of the exact same ambiguous term were considered as occurrences of that term and were included in the **MED** set. Table 2 provides details.

The third set, **ENG**, contains four ambiguous general English words (i.e., *line*, *interest*, *hard*, and *serve*), which have been used frequently in the general English domain to evaluate the performance of WSD. We downloaded the sense-tagged instances of these words from Ted Peterson's Web site (http://www.d.umn.edu/~tpederse/data.html). The sense-tagged instances for *interest* were created by Bruce and Wiebe.[4] Each instance of *interest* was tagged with one of six possible Longman Dictionary of Contemporary English (LDOCE) senses. The sense-tagged instances for *line*, *hard*, and *serve* were contributed by Leacock et al.[20,23] and senses were defined using WordNet. Table 3 describes the set.

Feature Representations and Window Sizes

Six different feature representations were studied for a given window size *n*, which will be referred to as "A," "B," "C," "D," "E," and "F," respectively. Each word was normalized to unify derivable forms associated with the same word, and all numbers were unified to the string XXX. Four feature representations (i.e., "A," "B," "C," and "E") depended on window sizes, while the window sizes for feature representations "D" and "F" were constant. Let "$\ldots w_{Ln}\ldots w_{L2}w_{L1}Ww_{R1}w_{R2}\ldots w_{Rn}\ldots$" be the context of consecutive words around the term *W* to be disambiguated.

Table 1 ▪ Information about the Abbreviation Set ABBR

W	SID	Sense Definition	N	W	SID	Sense Definition	N
APC	APC_1	antigen-presenting cells	1,356	MAC	MAC_1	membrane attack complex	231
	APC_2	adenomatous polyposis coli	430		MAC_2	macrophage	40
	APC_3	atrial premature complexes	8		MAC_3	mycobacterium avium complex	535
	APC_4	aphidicholin	37		MAC_4	macandrew alcoholism scale	18
	APC_5	activated protein c	479		MAC_5	monitored anesthesia care	19
ASP	ASP_1	antisocial personality	54		MAC_6	mental adjustment to cancer	19
	ASP_2	asparaginase	17	MAS	MAS_1	mccune albright syndrome	31
	ASP_3	aspartic acid	8		MAS_2	meconium aspiration syndrome	81
	ASP_4	ankylosing spondylitis	2	MCP	MCP_1	metacarpophalangeal joint	8
	ASP_5	aspartate	60		MCP_2	multicatalytic protease	9
BPD	BPD_1	borderline personality disorder	208		MCP_3	metoclopramide	157
	BPD_2	bronchopulmonary dysplasia	465		MCP_4	monocyte chemoattractant protein	185
	BPD_3	biparietal diameter	233		MCP_5	membrane cofactor protein	102
BSA	BSA_1	body surface area	354	PCA	PCA_1	para chloroamphetamine	210
	BSA_2	bovine serum albumin	2,808		PCA_2	passive cutaneous anaphylaxis	376
DIP	DIP_1	desquamative interstitial pneumonia	31		PCA_3	patient controlled analgesia	507
	DIP_2	distal interphalangeal	81		PCA_4	posterior communicating artery	5
FDP	FDP_1	fructose diphosphate	8		PCA_5	posterior cerebral artery	112
	FDP_2	formycin diphosphate	2		PCA_6	principal component analysis	343
	FDP_3	fibrinogen degradation product	382	PCP	PCP_1	p chlorophenylalanine	1
	FDP_4	flexor digitorum profundus	39		PCP_2	pentachlorophenol	341
LAM	LAM_1	lipoarabinomannan	103		PCP_3	phencyclidine	1,071
	LAM_2	lymphangiomyomatosis	22		PCP_4	pneumocystis carinii pneumonia	812
	LAM_3	leukocyte adhesion molecule	2	PEG	PEG_1	polyethylene glycols	52
	LAM_4	lymphangioleiomyomatosis	56		PEG_2	percutaneous endoscopic gastrostomy	18
RSV	RSV_1	respiratory syncytial virus	1,335	PVC	PVC_1	polyvinylchloride	473
	RSV_2	rous sarcoma virus	619		PVC_2	premature ventricular contraction	98
Total # of abbreviations: 15				Total # senses: 69		Total # instances: 24,407	

The first and fifth columns are abbreviations. The second and sixth columns are their sense identifications (SID). The third and seventh columns are corresponding sense definitions, and the fourth and eighth columns (N) are their numbers of sense-tagged instances in the training set.

Referring to this context, feature representations are described as follows:

- **A**—All words with their corresponding oriented distances within the window, i.e., $Ln/w_{Ln}, \ldots, L2/w_{L2}, L1/w_{L1}, R1/w_{R1}, R2/w_{R2}, \ldots$, and Rn/w_{Rn}, where L is for left, R is for right, and the number is for the distance.
- **B**—All words with their corresponding orientations within the window, i.e., $L/w_{Ln}, \ldots, L/w_{L2}, L/w_{L1}, R/w_{R1}, R/w_{R2}, \ldots$, and R/w_{Rn}.
- **C**—All words within the window, i.e., $w_{Ln}, \ldots, w_{L2}, w_{L1}, w_{R1}, w_{R2}, \ldots$, and w_{Rn}.
- **D**—All words with their corresponding orientation within a window of size 3 and the three nearest two-word collocations, i.e., L/w_{L3}, L/w_{L2}, L/w_{L1}, R/w_{R1}, R/w_{R2}, R/w_{R3}, $L2L1/w_{L2}\_w_{L1}$, $L1R1/w_{L1}\_w_{R1}$, and $R1R2/w_{R1}\_w_{R2}$.
- **E**—Features in representations "C" and "D."
- **F**—Features in representation "D" and all words in the context except *W*.

For purposes of illustration, features of *CSF* in **Instance 1** with a window size of 3 are shown in Table 4.

Instance 1. *At the same time, other researchers explored CSF parameters in multiple sclerosis, treatment of experimental optic neuritis, corticosteroid treatment of multiple sclerosis, and variations and mimickers of optic neuritis.*

Supervised Learning Algorithms Implemented

We experimented with four different supervised learning methods, including naïve Bayes learning, traditional decision list learning, our mixed supervised learning, and our implementation of decision list learning; the last two were our adaptation of naïve Bayes learning and tradition decision list learning. The first two algorithms were introduced in the background section; our detailed adaptation of the algorithms is presented below.

Naïve Bayes Learning

We used the Witten-Bell discounting technique[25] to avoid zero probability in the algorithm. Witten-Bell discounting is based on a simple intuition about zero-frequency events: the probability of seeing a zero-frequency feature is estimated by the probability of seeing a feature for the first time. Let *N* be the occurrences of all features in the training set, *T* be the number of different features appearing in the training set, and *Z* be the number of different features that have zero-frequency in the universe. The frequency of unseen features is

$$\frac{T}{Z} \times \frac{N}{(N+T)}$$

However, *Z* is not known in the WSD problem. We used

$$\frac{T}{100 \times (N+T)}$$

as the frequency of unseen features by assuming $Z = 100 \times N$.

Traditional Decision List Learning

We used the algorithm that was implemented by Yarowsky.[12] Each individual feature *f* consists of a test. All tests are ordered according to their log–likelihood ratios:

Table 2 ▪ Detailed Information for General Biomedical Terms MED

w	SID	Sense Definition	N	w	SID	Sense Definition	N
COLD	M1	cold temperature (Natural Phenomenon)	86	MOLE	M1	mol (Quantitative Concept)	83
	M2	common cold (Disease)	6		M2	mole the mammal	1
	M3	chronic obstructive airway disease	1		None	none of above	16
	M5	cold sensation (Qualitative Concept)	2	MOSAIC	M1	spatial mosaic (Spatial Concept)	45
	None	none of above	5		M2	mosaic (Organism Attribute)	52
DEGREE	M1	degree <1> (Qualitative Concept)	63		None	none of above	3
	M2	degree <2> (Intellectual Product)	2	NUTRITION	M1	nutrition (Organism Attribute)	45
	None	none of above	35		M2	science of nutrition	16
DEPRESSION	M1	mental depression	85		M3	feeding and dietary regimens	28
	None	none of above	15		None	none of above	11
DISCHARGE	M1	discharge (Body Substance)	1	PATHOLOGY	M1	pathology (Occupation or Discipline)	14
	M2	patient discharge	74		M2	pathology <3> (Pathologic Function)	85
	None	none of above	25		None	none of above	1
EXTRACTION	M1	extraction (Laboratory Procedure)	82	REDUCTION	M1	reduction—action (Health Care Activity)	2
	M2	extraction, NOS (Therapeutic Procedure)	5		M2	reduction (chemical) (Natural Phenomenon)	9
	None	none of above	13		None	none of above	89
FAT	M1	obese build (Organism Attribute)	2	REPAIR	M1	repair–action	52
	M2	fatty acid glycerol esters (Lipid)	71		M2	wound healing	16
	None	none of above	27		None	none of above	32
GROWTH	M1	growth <1> (Organism Function)	37	SCALE	M2	intellectual scale	65
	M2	growth <2> (Functional Concept)	63		None	none of above	35
IMPLANTATION	M1	Blastocyst implantation, natural	17	SEX	M1	coitus (Organism Function)	15
	M2	implantation procedure	81		M2	sex (Individual Behavior)	5
	None	none of above	2		M3	gender (Organism Attribute)	80
JAPANESE	M1	Japanese language	6	ULTRASOUND	M1	ultrasonography	84
	M2	Japanese population	73		M2	ultrasonic shockwave	16
	None	none of above	21	WEIGHT	M1	weight (Qualitative Concept)	24
LEAD	M1	lead (Element)	27		M2	body weight (Organism Attribute)	29
	M2	lead measurement, quantitative	2		None	none of above	47
	None	none of above	71	WHITE	M1	white color	41
MAN	M1	male (Organism Attribute)	58		M2	Caucasoid race	49
	M2	men (Population Group)	1		None	none of above	10
	M3	Homo sapiens (Population Group)	33				
Total # of terms: 22		Total # senses: 66		Total # instances: 2,200			

The first and fifth columns are general biomedical terms. The second and sixth columns are their sense identifications (SID). The third and seventh columns are corresponding sense definitions, and the fourth and eighth columns (N) are their numbers of sense-tagged instances in the training set. Note that we did not list senses that have no occurrence in the sense-tagged corpus. Also, general biomedical terms may have senses that are not biomedical; all meanings not presented in the UMLS are designated with one sense "None."

$$\log\left(\frac{Occu(s,f)}{Occu(f)-Occu(s,f)}\right)$$

where s is the majority sense that co-occurs with f, *Occu(f)* is the number of occurrences of f, and *Occu(s,f)* is the number of occurrences of f appearing in instances of W that are associated with the sense s. The default test returns the majority sense. For features (*f*) that co-occur with only one sense, a smoothing factor 0.1 is added to the total occurrences of f.

Our Decision List Learning

In the traditional implementation of decision list learning, a smoothing factor is added to the occurrence of features that occur with only one sense. However, it is not clear what the suitable smoothing factor is. In **our implementation of decision list learning (ODLL)**, we separated features that co-occur with only one sense from others to avoid the estimation of a smoothing factor. Two sets of tests are derived during the learning. The first set consists of features that co-occur with only one sense and are ordered according to the following formula:

$$\log\left(\frac{Occu(f)}{Occu(s)}\right)$$

where *Occu(s)* is the number of occurrences of the sense s. The second set consists of features (*f*) that co-occur with multiple senses and are ordered according to their log–likelihood ratio:

$$\log\left(\frac{Occu(s,f)}{Occu(f)-Occu(s,f)}\right)$$

Given a novel instance, the first set is applied first; if the sense cannot be determined by the first set, the second set is then applied, and the default test returns the majority sense.

Mixed Supervised Learning

After observing that the existence of instances with rare senses deteriorates naïve Bayesian classifiers, we implemented our **mixed supervised learning algorithm**

Table 3 ■ Information about General English Word Set ENG

Word	SID	Sense Definition	N
HARD	HARD$_1$	difficulty	3,455
	HARD$_2$	laborious, heavy	502
	HARD$_3$	contradiction to soft	376
INTEREST	INTEREST$_1$	readiness to give attention	361
	INTEREST$_2$	quality to causing attention to be given to	11
	INTEREST$_3$	activity, etc., that one gives attention to	66
	INTEREST$_4$	advantage, advancement or favor	178
	INTEREST$_5$	a share in a company or business	500
	INTEREST$_6$	money paid for the use of money	1,252
LINE	LINE$_1$	cord	373
	LINE$_2$	division	374
	LINE$_3$	formation	349
	LINE$_4$	phone	429
	LINE$_5$	product	2,218
	LINE$_6$	text	404
SERVE	SERVE$_1$	work for or be a servant to	1,814
	SERVE$_2$	be sufficient; be adequate, either in quality or quantity	1,272
	SERVE$_3$	do duty or hold offices; serve in a specific function	853
	SERVE$_4$	provide (usually but not necessarily food)	439
Total # of words: 4		Total # senses: 19	Total # instances: 15,983

The first column lists each word. The second and third columns are associated sense identifications (SID) and sense definitions. The last column (N) is the number of sense-tagged instances for associated sense in the training set.

(MSL), which contains a naïve Bayesian classifier and an instance-based classifier. The mixed supervised learning algorithm can be stated as follows:

- Split the training set into two parts, I and II, where part I contains instances associated with senses that have at least ten associated instances and occur more than 1% of the total number of instances; all remaining instances are included in part II.
- Build a naïve Bayes classifier trained on part I and an instance-based classifier trained on part II.
- For a novel instance, the instance-based classifier predicts its sense with the majority vote sense of all instances with a relatively high similarity (≥0.5). If there is such a sense, return the predicted sense; else return the predicted sense of the naïve Bayes classifier.

The similarity measure for the instance-based classifier is weighted. Assume the numbers of non-zero feature values for two instances are T1 and T2, then a weight of 2/(T1+T2+1) is assigned to each collocation, 1.5/(T1+T2+1) to each oriented word, and 1/(T1+T2+1) to each of the other features. Note that if there are no instances in part II of the training set, our mixed supervised learning algorithm is the same as naïve Bayes learning.

Table 4 ■ Six Options of Feature Representation for the Instance *"At the same time, other researchers explored CSF parameters in multiple sclerosis, treatment of experimental optic neuritis, corticosteroid treatment of multiple sclerosis, and variations and mimickers of optic neuritis."*

FP	Features	Example (window size = 3)
A	Words with oriented distance within the window	*L3/other, L2/researcher, L1/explore, R1/parameter, R2/in, R3/multiple*
B	Words with orientation within the window	*L/other, L/researcher, L/explore, R/parameter, R/in, R/multiple*
C	Words within the window	*other, researcher, explore, parameter, in, multiple*
D	Three collocations, oriented words within a window size 3	*L/researcher, L2L1/researcher_explore, L/explore, L1R1/explore_parameter, R/parameter, R1R2/parameter_in, R/in*
E	Features in C and D	*L/researcher, L2L1/researcher_explore, L/explore, L1R1/explore_parameter, R/parameter, R1R2/parameter_in, R/in, other, researcher, explore, parameter, in, multiple*
F	Features in D and all other words	*L/researcher, L2L1/researcher_explore, L/explore, L1R1/explore_parameter, R/parameter, R1R2/parameter_in, R/in, at, the, . . . of, optic, neuritis*

FP = feature representation.

Evaluation Methods

For each ambiguous word *W* in the three data sets, we derived 88 WSD classifiers for *W*: eight were represented using a pair (**ml**, **fp$_1$**), and 80 were represented by a tuple (**ml**, **fp$_2$**, **ws**). The aspect **ml** is a supervised learning algorithm with four choices: naïve Bayes learning (NBL), traditional decision list learning (TDLL), our implementation of decision list learning (ODLL), and our mixed supervised learning (MSL); **fp$_1$** and **fp$_2$** are feature representation aspects, where **fp$_1$** has two values "D" and "F," and **fp$_2$** has four values "A," "B," "C," and "E" (refer to Table 4 for these feature representations). The aspect **ws** is the window size with five values (2, 4, 6, 8, and 10). For multiple occurrences of *W* in an instance, features were derived for each occurrence, and the final feature vector was presented to learning algorithms containing all derived features.

We applied the ten-fold cross-validation method to measure the performance (i.e., measures were averaged over the results of the ten folds), in which the performance of classifiers was measured using precision (i.e., the ratio of the number of instances tagged correctly to the number of instances in the training set). Note that we assigned the majority sense to instances that failed to be tagged by classifiers. We controlled each ten-fold partition so that the same partition was used to evaluate each classifier.

Results

The overall running time for the experiment was about 13 hours; we did not keep track of the running time for each classifier. The overall performances of different classifiers for sets ABBR, MED, and ENG are listed in Tables 5, 6, and 7, respectively. Table 8 lists the best classifier for each word from these sets.

Table 5 ■ Overall Performance of Different Classifiers for Abbreviations

		Precision for Abbreviations ABBR (95% confidence interval) %			
FP	WS	NBL	MSL	TDLL	ODLL
A	2	90.9 (±1.1)	91.8 (±1.1)	90.7 (±1.2)	90.6 (±1.2)
	4	91.8 (±1.1)	93.4 (±1.0)	91.9 (±1.1)	92.1 (±1.1)
	6	91.6 (±1.1)	94.1 (±0.9)	91.4 (±1.1)	92.0 (±1.1)
	8	90.8 (±1.1)	93.9 (±0.9)	91.3 (±1.1)	91.9 (±1.1)
	10	90.4 (±1.2)	94.1 (±0.9)	91.0 (±1.1)	91.8 (±1.1)
B	2	91.4 (±1.1)	92.0 (±1.1)	90.9 (±1.1)	90.7 (±1.2)
	4	94.5 (±0.9)	94.8 (±0.9)	93.3 (±1.0)	93.0 (±1.0)
	6	95.5 (±0.8)	95.7 (±0.8)	93.8 (±1.0)	94.0 (±0.9)
	8	96.0 (±0.8)	96.1 (±0.8)	93.9 (±0.9)	94.2 (±0.9)
	10	96.3 (±0.7)	96.4 (±0.7)	94.1 (±0.9)	94.3 (±0.9)
C	2	91.7 (±1.1)	92.1 (±1.1)	91.3 (±1.1)	91.0 (±1.1)
	4	94.6 (±0.9)	94.7 (±0.9)	93.5 (±1.0)	93.1 (±1.0)
	6	95.9 (±0.8)	95.9 (±0.8)	94.1 (±0.9)	94.1 (±0.9)
	8	96.3 (±0.7)	96.3 (±0.7)	94.3 (±0.9)	94.4 (±0.9)
	10	96.9 (±0.7)	96.8 (±0.7)	94.6 (±0.9)	94.7 (±0.9)
E	2	90.0 (±1.2)	93.1 (±1.0)	91.7 (±1.1)	92.3 (±1.1)
	4	94.6 (±0.9)	95.8 (±0.8)	93.9 (±0.9)	94.6 (±0.9)
	6	96.2 (±0.8)	96.8 (±0.7)	94.7 (±0.9)	95.4 (±0.8)
	8	97.2 (±0.7)	97.4 (±0.6)	95.1 (±0.9)	95.8 (±0.8)
	10	97.6 (±0.6)	97.7 (±0.6)	95.3 (±0.8)	96.2 (±0.8)
D	NA	84.0 (±1.5)	91.0 (±1.1)	91.7 (±1.1)	92.0 (±1.1)
F	NA	98.6 (±0.5)	98.5 (±0.5)	95.7 (±0.8)	96.7 (±0.7)

The machine learning algorithm has four choices: NBL, MSL, TDLL, and ODLL. The feature representation (FP) has six options: A, B, C, D, E, and F, where A, B, C, and E have five different window sizes 2, 4, 6, 8, and 10. The 95% confidence interval for each precision value (p) in the table is p ± 0.5~1.5%.

Table 6 ■ Overall Performance of Different Classifiers for General Medical Terms

		Precision for General Biomedical Terms MED (95% confidence interval) %			
FP	WS	NBL	MSL	TDLL	ODLL
A	2	59.7 (±6.5)	72.0 (±5.9)	70.9 (±6.0)	71.4 (±6.0)
	4	46.6 (±6.6)	69.6 (±6.1)	74.6 (±5.8)	73.0 (±5.9)
	6	37.1 (±6.4)	67.1 (±6.2)	74.3 (±5.8)	71.0 (±6.0)
	8	32.5 (±6.2)	65.2 (±6.3)	74.8 (±5.7)	71.4 (±6.0)
	10	30.7 (±6.1)	63.5 (±6.4)	74.7 (±5.7)	69.8 (±6.1)
B	2	62.2 (±6.4)	71.8 (±5.9)	74.8 (±5.7)	75.3 (±5.7)
	4	58.1 (±6.5)	71.1 (±6.0)	77.0 (±5.6)	76.5 (±5.6)
	6	57.6 (±6.5)	72.9 (±5.9)	77.2 (±5.5)	76.9 (±5.6)
	8	56.4 (±6.6)	71.7 (±6.0)	76.5 (±5.6)	74.3 (±5.8)
	10	56.0 (±6.6)	71.8 (±5.9)	77.1 (±5.6)	73.6 (±5.8)
C	2	63.6 (±6.4)	72.2 (±5.9)	76.0 (±5.6)	75.4 (±5.7)
	4	60.8 (±6.5)	71.8 (±5.9)	77.1 (±5.6)	76.4 (±5.6)
	6	64.0 (±6.3)	75.6 (±5.7)	77.6 (±5.5)	77.1 (±5.6)
	8	64.6 (±6.3)	76.4 (±5.6)	77.1 (±5.6)	75.5 (±5.7)
	10	64.5 (±6.3)	76.1 (±5.6)	76.9 (±5.6)	74.4 (±5.8)
E	2	49.4 (±6.6)	68.7 (±6.1)	75.2 (±5.7)	75.2 (±5.7)
	4	51.6 (±6.6)	71.9 (±5.9)	77.7 (±5.5)	76.5 (±5.6)
	6	53.9 (±6.6)	72.3 (±5.9)	76.4 (±5.6)	76.2 (±5.6)
	8	57.7 (±6.5)	74.0 (±5.8)	77.4 (±5.5)	77.0 (±5.6)
	10	56.9 (±6.5)	74.7 (±5.7)	77.3 (±5.5)	76.6 (±5.6)
D	NA	45.1 (±6.6)	66.6 (±6.2)	75.0 (±5.7)	75.9 (±5.7)
F	NA	63.0 (±6.4)	77.8 (±5.5)	78.0 (±5.5)	74.0 (±5.8)

The machine learning algorithm has four choices: NBL, MSL, TDLL, and ODLL. The feature representation (FP) has six options: A, B, C, D, E, and F, where A, B, C, and E have five different window sizes 2, 4, 6, 8, and 10. The 95% confidence interval for each precision value (p) in the table is p ± 5.5~6.6%.

Our results showed that supervised WSD achieved the best performance on the set of abbreviations, the second best on the set of general English terms, and the worst for general biomedical terms. For abbreviations, naïve Bayes learning and our mixed supervised learning achieved the best performance when using feature representation "F," with an overall precision of more than 98%. For general biomedical terms, the traditional decision list achieved the best performance with

Table 7 ■ Overall Performance of Different Classifiers for General English Words

		Precision for General English Words ENG (%)				
FS	WS	SVM	NBL	MSL	TDLL	ODLL
A	2	84.4	83.8	83.9	83.0	81.9
	4	86.0	85.2	85.5	83.5	81.0
	6	85.6	83.8	84.4	82.6	79.0
	8	85.4	82.9	83.7	82.1	77.4
	10	84.9	81.5	82.5	81.4	75.8
B	2	84.3	84.3	81.6	82.4	81.1
	4	86.8	85.5	85.6	82.7	79.9
	6	87.1	84.9	85.0	81.8	78.2
	8	87.0	83.3	83.4	80.3	75.8
	10	86.9	82.7	82.7	79.5	74.2
C	2	84.0	82.9	83.0	80.8	79.6
	4	86.9	83.6	83.6	81.4	78.3
	6	87.2	83.6	83.7	80.6	77.0
	8	86.8	82.5	82.5	79.2	75.0
	10	86.5	81.7	81.8	78.8	74.2
E	2	85.9	84.8	83.6	84.9	83.8
	4	88.8	88.9	89.3	86.1	84.9
	6	89.5	90.4	90.6	85.8	84.5
	8	89.7	90.5	90.6	85.2	84.1
	10	89.6	90.8	90.8	85.2	83.8
D	NA	85.4	83.2	82.6	85.5	84.5
F	NA	89.2	89.2	89.3	83.8	82.6

The machine-learning algorithm has four choices: NBL, MSL, TDLL, and ODLL. The feature representation (FP) has six options: A, B, C, D, E, and F, where A, B, C, and E have five different window sizes 2, 4, 6, 8, and 10. The 95% confidence interval for each precision value (p) in the table is p ± 1.4~2.1%.

feature representation "F," with an overall precision around 75%. For general English words, naïve Bayes learning and our mixed supervised learning achieved the best performance when using feature representation "E" with a window size of 10, with an overall precision of 90.8%. However, naïve Bayes learning with feature representation "D" or feature representation "A" in some cases had the worst performance for each set. For example, naïve Bayes learning with feature representation "D" had a precision of 84.0% compared with more than 90% achieved by other classifiers for abbreviations.

Table 8 indicates that (1) abbreviations usually achieved the best performance using a larger window size with naïve Bayes learning or our mixed supervised learning; (2) there is no particular preference for feature representation, window size, or machine-learning algorithm for general biomedical terms; and (3) our mixed supervised learning achieved the best performance with feature representation "E" associated with a large window size (≥6) for *hard*, *interest*, and *serve*, and naïve Bayes learning achieved the best performance using feature representation "F" for the noun *line*.

Comparisons between the decision list learning algorithms and naïve Bayes learning with our mixed supervised learning for different word sets are shown in Figures 2 and 3, which also show the overall performance of classifiers with different window sizes. These figures indicate that naïve Bayes learning was unstable and varied dramatically for different feature representations. For example, naïve Bayes learning had the worst performance for feature representation "D" but had the best performance for feature representation "F" when testing on abbreviations.

Table 8 ■ Best Classifier for Each Word in Three Data Sets

Word	Best Classifier	Precision (%)
ANA	{F,0,MSL|NBL}	100.0
APC	{F,0,NBL}	99.0
ASP	{C,10,MSL}	90.8
BPD	{E,8,MSL|NBL}	98.4
	{F,0,MSL|NBL}	
BSA	{F,0,MSL|NBL}	99.5
DIP	*	100.0
FDP	{F,0,MSL|NBL}	98.8
LAM	{C,10,MSL|NBL}	
MAC	{F,0,MSL|NBL}	98.4
MAS	{C,10,MSL|NBL}	100.0
	{F,0,MSL|NBL|ODLL}	
MCP	{E,10,NBL}	99.1
PCA	{F,0,MSL|NBL}	99.4
PCP	{E,10,MSL}	98.2
PEG	{F,0,MSL|NBL}	96.7
PVC	{F,0,MSL|NBL}	99.3
	{E,10,MSL|NBL}	
RSV	{F,0,MSL|NBL}	98.4
LINE	{F,0,MSL|NBL}	90.0
SERVE	{E,10,MSL|NBL}	91.8
INTEREST	{E,6,MSL}	91.9
HARD	{E,8|10,MSL|NBL}	92.1
COLD	{C,2,ODLL}	90.9
DEGREE	{E,4,MSL|TDLL}	98.2
DEPRESSION	{A,8,ODLL}	88.8
DISCHARGE	{E,10,MSL}	90.8
	{B,8,ODLL}	
EXTRACTION	{C,8,TDLL}	89.7
FAT	{A|B,4,ODLL}	85.9
GROWTH	{F,0,MSL|NBL}	72.2
IMPLANTATION	{A,8,TDLL}	90.0
JAPANESE	{F,0,TDLL}	79.8
	{B,2,MSL}	
LEAD		91.0
MAN	{C,2,MSL}	91.0
MOLE	{F,0,MSL|TDLL}	91.1
MOSAIC	{F,0,MSL}	87.8
NUTRITION	{C,10,TDLL}	58.1
	{C,8,MSL|NBL}	
PATHOLOGY	{A,10,TDLL}	88.2
REDUCTION	{E,2,ODLL|MSL|TDLL}	91.0
REPAIR	{E,8,TDLL}	76.1
SCALE	{E,6,MSL|NBL}	90.9
SEX	{A,4,ODLL},	89.9
	{E,8|6|10,ODLL}	
ULTRASOUND	{D,3,ODLL}	87.8
WEIGHT	{C,8,TDLL|NBL}	78.0
	{E,6,NBL}	
WHITE	{E|C,4,ODLL}	75.6

Note that when there is no rare sense, our mixed supervised learning is the same as naïve Bayes learning.

For a fixed window size "ws" and a fixed feature representation option "fp" (p-values were computed using one-tailed paired t-test):

- The performance of our implementation of decision list classifiers for abbreviations was better than that of traditional decision list classifiers ($p < 0.0013$). However, the opposite held for the other two sets ($p < 0.001$).

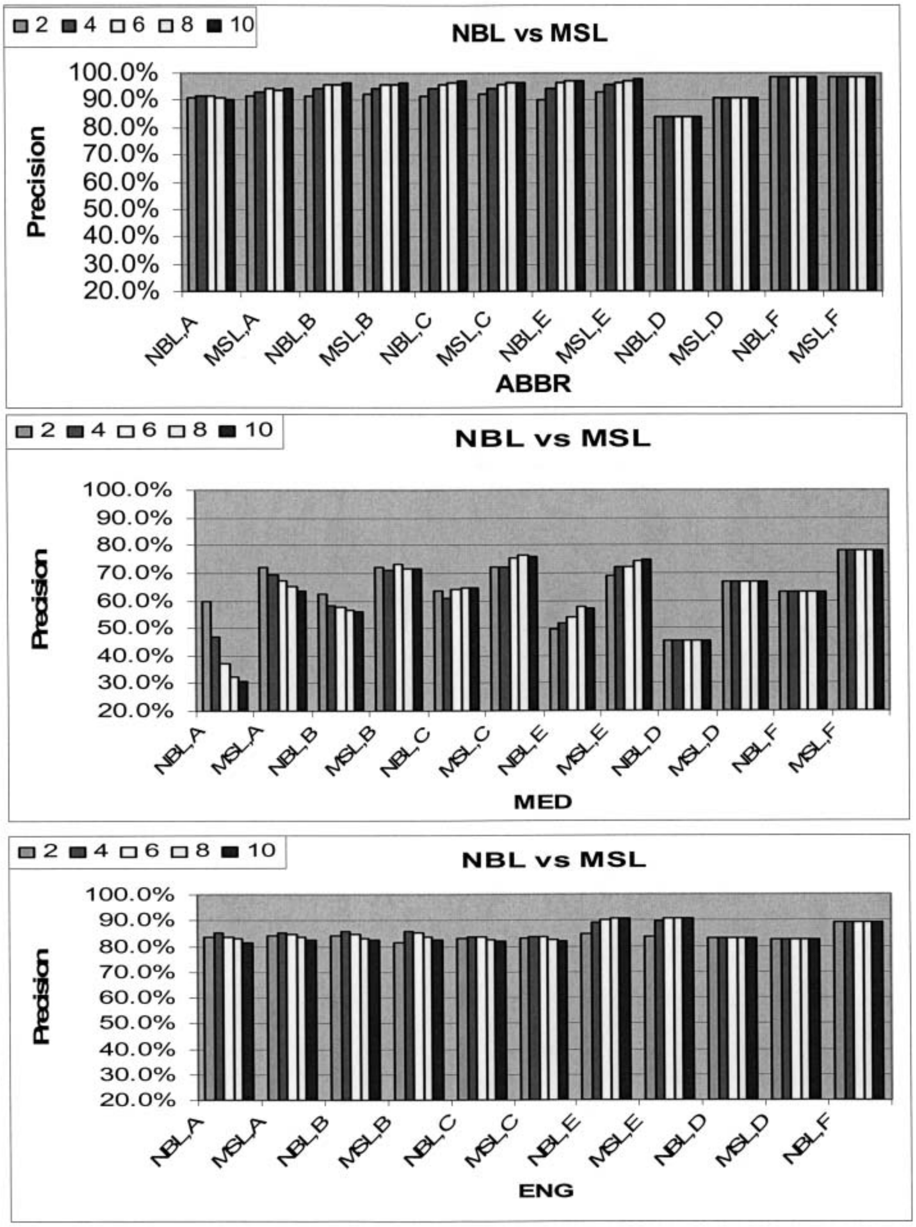

Figure 2. The comparison of naïve Bayes learning and our supervised mixed learning. ABBR stands for abbreviations, MED for general biomedical terms, and ENG for general English words.

- The performance of mixed supervised learning classifiers for all sets was generally superior to that of naïve Bayes classifiers ($p < 0 .001$).
- The performance of naïve Bayes classifiers was much worse than other classifiers for general biomedical terms ($p < 0.0001$).

Discussion

Note that the results obtained in this study for general English words cannot be compared with results reported in other studies[4,13,20,21,25] for not using the common evaluation method. But some of our findings are consistent with findings reported in these studies. For example, we found that naïve Bayes learning achieved the best performance on disambiguating the word *line*, which is consistent with the finding of Mooney.[21]

We believe that supervised WSD is suitable when there are enough sense-tagged instances. For example, the best classifier for *ASP* (with a total of 141 gold standard instances) achieved a precision of 90.8%, while the precision of the best classifier for *APC* (with a total of 2,310 gold standard instances) achieved a precision of 99.0% even though they have the same number of senses (i.e., five). There are at most 100 instances for each general biomedical term with an average of 33.3 instances, while averages of other sets are at least several hundred instances. All supervised WSD classifiers performed with a precision of less than 80% for general biomedical terms, while most classifiers achieved around 90% for general English words and more than 90% for abbreviations. The overall performances of supervised WSD classifiers differ among data sets. Almost all classifiers

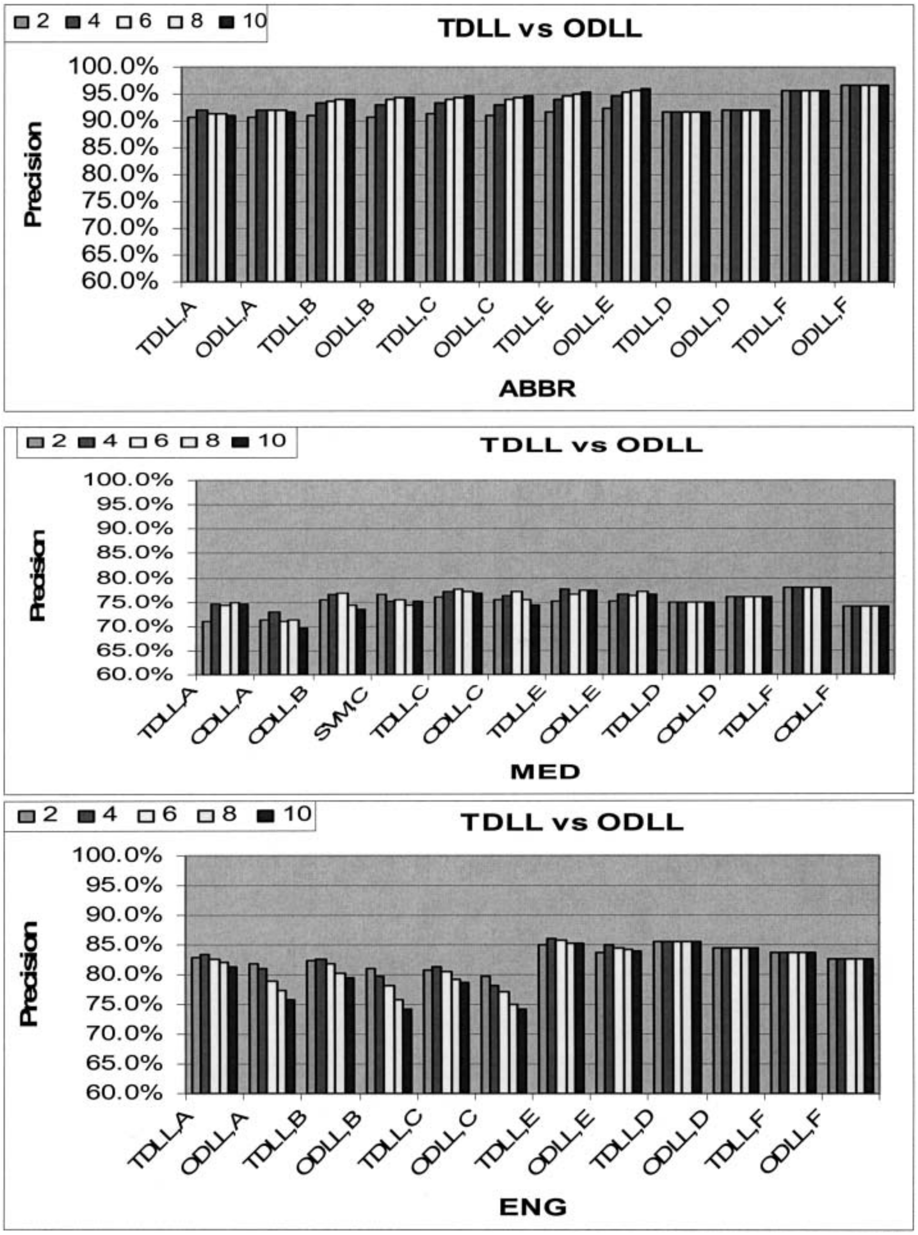

Figure 3. The comparison of traditional decision list learning and our decision list learning. ABBR stands for abbreviations, MED for general biomedical terms, and ENG for general English words.

achieved better overall performance for set ABBR than those for sets MED and ENG. One possible reason is that senses of abbreviations are domain-specific, and most of them are generally quite unrelated, which makes the disambiguation task easier than others. For example, the abbreviation *PCA* has six senses that are comparable to the general English words *line* and *interest*. The best classifier for *PCA* achieves a precision of more than 99%, while the best classifiers for *line* and *interest* achieve precisions of lower than 93%. The best choice of window size is also related to data sets. For example, the performance of classifiers for abbreviations increases when the window size increases, but the performance for general English words usually decreases when the window size increases after a size of 4.

Feature representations "E" or "F" together with large window sizes achieved the best performance for almost every abbreviation. However, feature representation "A," which contains words with their oriented distances, and feature representation "D," which contains all words with their corresponding orientation within a window of size 3 plus the three nearest two-word collocations, performed much worse than others for abbreviations. This difference may be because feature representations "A" and "D" failed to capture critical keywords that indicated their biomedical senses. Moreover, features derived using "A" were sparse compared with others, a problem not shared by "E" and "F."

The difference between feature representations "C" and "E" is the inclusion of collocations in "E." Classifiers associated with "E" outperform the corresponding classifiers using feature representation "C" given fixed value combinations of other aspects ($p < 0.001$). This indicates that the inclusion of collocations is important for supervised WSD. However, since collocations by themselves do not achieve good performance, the ideal feature representation would be a combination of collocations with other techniques such as bag of words and oriented words.

Naïve Bayes learning did not perform well when there were rare senses in the training set. NBL was also unstable with respect to feature representations and window sizes. Our mixed supervised learning, which combines naïve Bayes learning with instance-based learning, overcomes these disadvantages and achieves relatively better performance ($p < 0.001$).

The study shows that there is no single combination of feature representation, window size, and machine-learning algorithm that has a stable and relatively good performance for all ambiguous terms. The choice of the best classifier for each term depends on the number of sense-tagged instances for that term as well as its associated domain. The experimental method presented in the study can be used to select the best-supervised WSD classifier for each ambiguous term.

Note that in the study, we did not investigate the predication power of classifiers that were produced through bagging or boosting multiple weak classifiers.[26] One possible direction of future work is to apply bagging or boosting techniques to supervised WSD and see the potential improvement when using these techniques.

Limitations

The study has several limitations. We used a predetermined number 3,000 to reduce the number of instances for a specific term. However, different machine-learning algorithms exhibit different sensitivities to sample size. We plan to perform a sample sensitivity study to investigate the relation between machine-learning algorithm and sample size.

Additionally, we used only precision to measure the performance. Measures such as the multiclass receiver-operating characteristic (ROC) curve could be used, but it would be very complicated because the numbers of classes as well as the numbers of instances were different among terms in the same data set as well as from different data sets. It is caused by the nature of our task: different terms have different numbers of senses and frequency in the same domain, and the same term has different numbers of senses and frequency in different domains. It is also the reason why our comparisons among different data sets were not equally footed.

Conclusion

We conducted an experiment that compared feature representation, window size, and supervised learning algorithms and concluded that supervised WSD is suitable only when we have enough sense-tagged instances (with at least a few dozens of instances for each sense). Collocations combined with neighboring words are appropriate feature representations. For terms with unrelated biomedical senses, a large window size (e.g., the whole paragraph) should be used, while for general English words a moderate window size between 4 and 10 is sufficient. For abbreviations, our mixed supervised learning was stable and generally better than naïve Bayes learning, and our implementation of decision list learning performed better than traditional decision list learning.

This study shows clearly that the different aspects of supervised WSD depend on each other. The experiment method presented in the study can be used to select the best-supervised WSD classifier for each ambiguous term.

References ■

1. Ide N, Veronis J. Introduction to the special issue on word sense disambiguation: the state of the art. Computational Linguistics. 1998;24(1):1–40.
2. Ng HT, Zelle J. Corpus-based approaches to semantic interpretation in natural language processing. AI Magazine. 1997;winter:45–64.
3. Kilgarriff A, Rosenzweig J. Framework and results for English SENSEVAL. Comput Humanities. 1999;34:1–2.
4. Bruce R, Wiebe J. Word-sense disambiguation using decomposable models. Proceedings of the Thirty-Second Annual Meeting of the Association of Computational Linguistics. 1994: 139–46.
5. Ng HT. Getting serious about word-sense disambiguation. Proceedings of the ACL SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What and How? 1997:1–7.
6. Engelson SP, Dagan I. Minimizing manual annotation cost in supervised training from corpora. Proceedings of the Thirty-Fourth Annual Meeting of the Association of Computational Linguistics. 1996;34:319–26.
7. Fujii A, Inui K, Tokunaga T, Tanaka H. Selective sampling for example-based word sense disambiguation. Computational Linguistics. 1998;24(4):573–97.
8. Liu H, Lussier Y, Friedman C. Disambiguating ambiguous biomedical terms in biomedical narrative text: an unsupervised method. J Biomed Inform. 2001;34:249–61.
9. Liu H, Johnson SB, Friedman C. Automatic resolution of ambiguous terms based on machine learning and conceptual relations in the UMLS. J Am Med Inform Assoc. 2002;9:621–36.
10. Veronis J, Ide N. Very large neural networks for natural language processing. Proceedings of the European Conference on Artificial Intelligence. 1990:366–8.
11. Towell G, Voorhees EM. Disambiguating highly ambiguous words. Computational Linguistics. 1998;24(1):125–46.
12. Yarowsky D. Decision lists for lexical ambiguity resolution: application to accent restoration in Spanish and French. Proceedings of the Thirty-Second Annual Meeting of the Association of Computational Linguistics. 1994:88–95.
13. Ng HT, Lee HB. Integrating multiple knowledge sources to disambiguate word sense: an exemplar-based approach. Proceedings of the Thirty-Fourth Annual Meeting of the Association of Computational Linguistics. 1996:40–7.
14. Ng HT. Exemplar-based word sense disambiguation: some recent improvements. Proceedings of the Second Conference on Empirical Methods in Natural Language Processing. 1997:208–13.
15. Mooney R. Inductive logic programming for natural language processing: In: Muggleton S (ed). Inductive Logic Programming: Selected Papers from the 6th International Workshop. New York, NY: Springer Verlag, 1997.
16. Marquez L. Machine learning and natural language processing. Technical Report LSI-00-45-R, Departament de Llenguatges i Sistemes Informàtics, Universitat Politècnica de Catalunya, 2000.
17. Duda R, Hart P. Pattern Classification and Scene Analysis. New York, NY: John Wiley and Sons, 1973.
18. Aha D, Kibler D, Albert M. Instance-based learning algorithms. Machine Learning. 1991;7:33–66.
19. Jorgensen J. The psychological reality of word senses. J Psycholinguist Res. 1990;19(3):167–90.

20. Leacock C, Chodorow M, Miller G. Using corpus statistics and WordNet relations for sense identification. Computational Linguistics. 1998;24(1):147–65.
21. Mooney R. Comparative experiments on disambiguating word senses: an illustration of the role of bias in machine learning. Proceedings of the First Conference on Empirical Methods in Natural Language Processing. 1996:82–91.
22. Weeber M, Mork J, Aronson A. Developing a test collection for biomedical word sense disambiguation. Proc AMIA Symp. 2001: 746–50.
23. Leacock C, Towell G, Voorhees EM. Corpus-based statistical sense resolution. Proceedings of the Advanced Research Projects Agency (ARPA) Workshop on Human Language Technology. 1993:260–5.
24. Witten I, Bell T. The zero-frequency problem: estimating the probabilities of novel events in adaptive text compression. IEEE Trans Inf Theory. 1991;37:1085–94.
25. Escudero G, Marquez L, Rigau G. Naive Bayes and exemplar-based approaches to word sense disambiguation revisited. Proceedings of the 14th European Conference on Artificial Intelligence (ECAI). 2000:421–5.
26. Escudero G, Marquez L, Rigau G. Boosting applied to word sense disambiguation. Proceedings of the European Conference on Machine Learning. 2000:129–41.