# Dependency of context-based Word Sense Disambiguation from representation and domain complexity

Paola Velardi
Dipartimento di Scienze dell'Informazione
University "La Sapienza"
Roma
Velardi@dsi.uniroma1.it

Alessandro Cucchiarelli
Istituto di Informatica
University of Ancona
Ancona
alex@inform.unian.it

## Abstract

Word Sense Disambiguation (WSD) is a central task in the area of Natural Language Processing. In the past few years several context-based probabilistic and machine learning methods for WSD have been presented in literature. However, an important area of research that has not been given the attention it deserves is a formal analysis of the parameters affecting the performance of the learning task faced by these systems. Usually performance is estimated by measuring precision and recall of a specific algorithm for specific test sets and environmental conditions. Therefore, a comparison among different learning systems and an objective estimation of the difficulty of the learning task is extremely difficult.

In this paper we propose, in the framework of Computational Learning theory, a formal analysis of the relations between accuracy of a context-based WSD system, the complexity of the context representation scheme, and the environmental conditions (e.g. the complexity of language domain and concept inventory).

## 1 Introduction

In the literature (see Computational Linguistics (1998) for some recent results), there is a rather vast repertoire of supervised and unsupervised learning algorithms for WSD, most of which are based on a formal characterization of the surrounding context of a word or linguistic concept[1], and a function $f$ to compute the membership of a word to a category, given its context in running texts.

Despite the rich literature, none of these algorithms exhibit an "acceptable" performance with reference to the needs of real-world computational task (e.g. Information Retrieval, Information Extraction, Machine Translation etc.), except for particularly straightforward cases.

A very interesting WSD experiment is Senseval (1998), a large-scale exercise in evaluating WSD programs. One of the objectives of this experiment was to identify correlations between performance of the various systems and the parameters of the WSD task. Though the scoring of systems appears sensitive to certain factors, such as the degree of polysemy and the entropy of sense distributions, these correlations could not be consistently observed. There are words with fewer senses (e.g. *bet, consume, generous*) causing troubles to most systems, while there are words with a very high polysemy and entropy (e.g. *shake*) on which all systems obtain good performance. The justification that the Senseval coordinator Adam Kilgariff provides for *shake* is very interesting in the light of what we will discuss later in this paper: "The items (means contexts) for *shake* involve multi-word expressions, such as *shake one's head*. (...) Over 50% of the items for *shake* involve some multi-word expression or other." In other words, the contexts for *shake* are very

[1] The inventory of linguistic concepts is usually extracted from on-line resources like WordNet, the Longman dictionary (LDOCE), or HECTOR.

repetitive in the training set, therefore all systems could *easily learn* a sense discrimination model.

Furthermore, in Senseval (but also in other reported evaluations experiments) it appears that performances for individual words/concepts are extremely uneven within the same system. This scarce homogeneity of results suggests that performance is not solely related with the "cleverness" of a given learning algorithm.

Clearly, the performances of WSD systems are related to a variety of parameters, but the formal nature of these dependencies is not fully understood.

The Senseval experiment highlighted the necessity of a more accurate analysis of the correlations between performance of WSD systems and the parameters that may affect this task. In absence, a comparison of the various WSD algorithms and an estimation of their performance under different environmental conditions is extremely difficult.

In the next sections we briefly present a computational model of learning, called PAC theory (Anthony and Biggs (1997), Kearns and Vazirani (1994), Valiant (1984)), and we then show that this theory may be used to determine the formal relations between performance of context-based WSD models and environmental conditions, such as the complexity of the context representation scheme, and the the complexity of language domain and concept inventory.

## 2 A relation between sample size and complexity of learning task

Formally, the problem of example-based learning of WSD models can be stated as follows:

1 Given a class C of concepts $C_i$ (where C is either a hierarchy or a "flat" concept inventory),

2 Given a context-based *representation class* H for a concept class C, where H: $\Sigma^* \rightarrow C$ and $\Sigma$ is a finite alphabet of symbols (e.g. words or word tags),

3 Given an input space $X \subseteq \Sigma^*$ of encodings of instances in the learner's world, e.g. feature vectors representing contexts around words $w_j$, where $w_j$ is a member of $C_i$,

4 Given a training sample S of length m:

$$S = ((x_1, b_1) \ldots (xm, bm)) \quad x_i \in X, b_i \in \{0,1\}$$

where $b_i=1$ if $x_i$ is a positive example of $C_i$,

characterize formally a function $h(C_i) \in H$ that assigns a word w to a concept $C_i$, given the sentence context x of w. The hypothesis may have the form of a Hidden Markov Model with estimated transition probabilities, a decision list, a cluster of points in a representation space, a logic formula, etc.

The complexity of this learning task is related to several aspects, such as selecting an appropriate representation space H, an appropriate grain for the concept inventory C, and finally, a sufficiently representative training sample S.

As first, H must be a "reasonable" representation space for C. Quite intuitively, if we represent a linguistic concept as the set of possible morphologic tags pairs in a ±1 window, we will not be able to predict much, simply because surrounding morphologic tags are not sufficient to determine the semantic category of a word.

On the opposite, if we select an overly complex representation model, including irrelevant features, we run through the so called *overfitting* problem.

Thirdly, some of the features used in a representation may be dependent from other features, and again the model would result unnecessarily complex.

The problem of noise and overfitting are well known in the area of Machine Learning (Russell and Norvig (1999)), therefore we will not discuss the matter in detail here. An analysis of this issue as applied to probabilistic WSD learners may be found in Bruce and Wiebe (1999).

For the purpose of this paper, we assume that the representation space H is optimized with respect to the choice of the relevant model parameters. Our objective will be to determine the size of S, given H and C, and given certain performance objectives.

As we said, the aim of a WSD learning process, when instructed with a sequence S of examples in X, is to produce an hypothesis h which, in some sense, "corresponds" to the concept under consideration. Because S is a *finite* sequence, only concepts with a finite number of positive examples can be learned with total success, i.e. the learner can output an hypothesis h= $C_i$ . In general, and this is the case for linguistic concepts, we can only hope that h is a *good approximation* of $C_i$.. In our problem at hand, it is worth noticing that even humans may provide only approximate definitions of linguistic concepts!

The theory of Probably Approximately Correct (PAC) learning, a relatively recent field at the borderline between Artificial Intelligence and Information Theory, states the conditions under which h reaches this objective, i.e. the conditions under which a computer derived hypothesis h 'probably' represents $C_i$ 'approximately'.

**Definition 1 (PAC learning).** Let C be a concept class over X. Let D be a fixed probability distribution over the instance space X, and EX($C_i$,D) be a procedure reflecting the probability distribution of the population we whish to learn about. We say that C is **PAC learnable** if there exists an algorithm L with the following property: For every $C_i \in C$, for every distribution D on X, and for all $0<\varepsilon<1/2$ and $0<\delta<1/2$, if L is given access to EX($C_i$,D) and inputs ε and δ, then with probability at least (1-δ), L outputs a hypothesis h for concept $C_i$, satisfying error(h)<ε. The parameters ε and δ have the following meaning: ε is the probability that the learner produces a generalization of the sample that does not coincide with the target concept, while δ is the probability, given D, that a particularly unrepresentative (or noisy) training sample is drawn. The objective of PAC theory is to predict the performance of learning systems by deriving a lower bound for m, as a function of the performance parameters ε and δ.

Figure 1 (from Russell and Norvig (1999)) illustrates the "intuitive" meaning of PAC definition. After seeing m examples, the probability that $H_{bad}$ includes consistent hypotheses is:

$P(H_{bad} \supseteq H_{cons}) \leq | H_{bad} |(1-\varepsilon)^m \leq |H|(1-\varepsilon)^m$

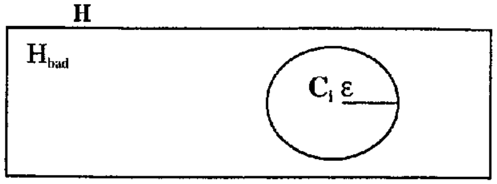


**Figure 1 : ε-sphere around the "true" function $C_i$**

And we want this to be:

$|H|(1-\varepsilon)^m \leq \delta$

we hence obtain a lower bound for the number of examples we need to submit to the learner in order to obtain the required accuracy:

(1) $m \geq \frac{1}{\varepsilon}\left(\ln\frac{1}{\delta}+\ln|H|\right)$

The inequality (1) establishes a sort of worst-case general bound, relating the size of the learning set with the complexity of the representation space |H|. Unfortunately this bound turns out to have limited utility in practical applications.

For example, if the hypothesis space for a linguistic concept $C_i$ is the classic "bag of words", i.e. a set of at least k "typical" context words selected by a probabilistic learner, after observing m samples of the ±n words around words $w \in C_i$

(e.g. $x = (w_{-n}, w_{-n+1}, .. w, \ldots w_{n-1}, w_n)$ )

then H is any choice of $\leq k \leq |V|$ words over |V| elements, where |V| ($\approx 10^5$) is the size of the vocabulary. We then have:

$|H| = 1 + \binom{|V|}{1} + \ldots + \binom{|V|}{k} \leq 2^{|V|}$

the above expression, used in inequality (1), produces an overly high bound for m, that can be hardly pursued especially in case the learning algorithm is supervised!

In PAC literature, the bound for m is often derived "ad hoc" for specific algorithms, in order to exploit knowledge on the precise learning conditions.

It is also worth noticing that PAC literature has mostly a theoretical emphasis, and most applications concentrated on the field of neural networks and natural learning systems (Hanson, Petsche, Kearns, Rivest (1994)). To the knowledge of the authors, the utility of this theory in the area of computer learning of natural language has not been explored.

In the following, we will derive a probabilistic expression for m in the track of (1), for the case of a *context-based WSD probabilistic learner*, a learning method that includes a rather wide class of algorithms in the area of WSD. We believe that adapting our analysis to other example-based WSD systems will not require a significant effort. This relation allows it to establish, upon an a-priori analysis of the chosen conceptual model and of the language domain, a more precise relation between performance, complexity of the learning algorithm, and environmental conditions (e.g. complexity of the language domain).

Our objective is to show that an a-priori analysis of the learning model and language domain may help to tune precisely a WSD experiment and allows a more uniform comparison between different WSD systems.

## 3. A formal estimate of accuracy for context-based probability WSD models

A probabilistic context-based WSD learner may be described as follows:

Let X be a space of feature vectors:

$f_k=( f(a_1{}^i=v_1,a_2{}^i=v_2,\ldots a_n{}^i=v_n)\in\Re^n,\ b_k^i))$,

$b_k^i=1$ if $f_k$ is a positive example of $C_i$ under H.

Each vector describes the context in which a word $w\in C_i$ is found, with variable degree of complexity. For examples, arguments may be any combination of plain words and their morphologic, syntactic and semantic tags.

We assume that arguments are **not** statistically independent (in case they are, the representation of a concept is more simple, see Bruce and Wiebe, (1999)).

An example (Cucchiarelli, Luzi and Velardi (1998)) is the case in which $f_k$ represents a syntactic relation between $w\in C_i$ and another word in its context. For example, given the compound *district banks* the following feature is generated as an example of the category *organization*:

((N_N district bank), organization(bank))

We further assume that observations of contexts are *noisy*, and the noise may be originated by several factors, such as tags ambiguity, and semantic ambiguity of the word whose context is observed.

In the above feature vector, the syntactic tag (first argument) could be wrong because of syntactic ambiguity and limited coverage of available parsers, and the ambiguous word *bank* could not be, in a specific context, an instance of the category *organization*, though it is in the example above.

Probabilistic learners usually associate to uncertain information a measure of the confidence the system has in that information. Therefore, we assume that each feature $f_k$ is associated to a concept $C_i$ with a *confidence* $\phi(i,k)$.

The confidence may be calculated in several ways, depending upon the type of selected features for $f_k$. For example, the Mutual Information measures the strength of a correlation between co-occurring arguments, and the Plausibility (Cucchiarelli, Luzi and Velardi (1998)) assigns a weight to a feature vector, depending upon the degree of ambiguity of its arguments and the frequency of its observations in a corpus. We assume here that $\phi$ is adjusted to be a probability, i.e. $\Sigma_i\phi(i,k)=1$. The factor $\phi(i,k)$ represents hence an estimate of the probability that $f_k$. is indeed a context of $C_i$.

Under these hypotheses, a representation $h\in H$ for a concept $C_i$ is the following:

$h(C_i):\{f^i_1..f^i_{mi}\}$

(2) $f_k \rightarrow h(C_i)$ iff $\phi(i,k) > \gamma$

A concept is hence represented by a set of features with associated probabilities[2]. Policy (2) establishes that only features with a probability higher than a threshold $\gamma$ are assigned to a category model.

Given an unknown word w' occurring in a context represented by $f'_k$, the WSD algorithm assigns w' to the category in C that maximizes the similarity between $f'_k$ and one of its members. Again, see Cucchiarelli, Luzi and Velardi (1998) and Bruce and Wiebe, (1999) for examples of similarity functions.

[2] Note that in case of statistical independence among the features in a vector, a model for a concept would be a set of features, rather than feature vectors, but most of what we discuss in this section would still apply with simple changes.

Given the above, the probabilistic WSD model for a category $C_i$ may fail because:

1 $C_i$ includes *false positives* (fp), e.g. feature vectors erroneously assigned to $C_i$

2 There are *false negatives* (fn), i.e. feature vectors erroneously discarded because of a low value $\phi(i,k)$

3 The context $f'_k$ of the word w' has never been observed around members of $C_i$, nor it is *similar* (in the precise sense of similarity established by a given algorithm) to any of the vectors in the contextual models.

We then have[3]:

(3) P(w' is misclassified on the basis of $f'_k$)=
$P(f'_k \in fp \text{ in } C_i)+P(f'_k \in fn \text{ outside } C_i)+P(f'_k \text{is unseen in } C_i)$

Let:

m be the total number of feature vectors extracted from a corpus

$m^k$ the total number of occurrences of a feature $f_k$

$m_i^k$ the number of times the context $f_k$ occurred with a word w' member of $C_i$

Notice that $\Sigma_i m_i^k \neq m^k$, since, because of ambiguity, a context may be assigned to more than one concept (or to none).

We can then estimate the three probabilities in expression (3) as follows:

$$(3.1)\ \hat{P}\ (\text{fp in } C_i)= \sum_{\phi(i,k)>\gamma} \frac{m_i^k}{m}(1-\phi(i,k))$$

$$(3.2)\ \hat{P}\ (\text{fn outside } C_i)= \sum_{\phi(i,k)\leq\gamma} \frac{m_i^k}{m}\phi(i,k)$$

$$(3.3)\ \hat{P}\ (\text{unseen in } C_i)=$$
$$(\frac{1}{m}\sum_{\forall m^k=1} m^k)\cdot(\frac{1}{m}\sum_k m_i^k)\cdot(\bar{\phi}(i))=\frac{\beta}{m}\sum_k m_i^k\phi(i,k$$

The third probability is computed as the product of three estimated factors: the probability $\beta$ of unseen contexts[4] in the corpus, the probability of extracting contexts around members of $C_i$, and the average confidence of a feature vector in $C_i$.

Classic methods such as Chernoff bounds may be applied to obtain good approximations for the three probabilities above. Notice however that in order to obtain a given accuracy of estimate, Chernoff bounds (and other methods) again impose a bound on the number of observed examples (Kearns and Vazirani (1994))

Since in (3.1) $(1-\phi(i,k))<\gamma$, in (3.2) $\phi(i,k))>\gamma$, and in (3.3) $\phi(i,k))\leq 1$, we obtain the bound:

$$P(w' \text{ is misclassified on the basis of } f'_k)=$$
$$\leq \frac{M_i - N_i}{m}(1-\gamma)+\frac{N_i}{m}\gamma+\beta_m\frac{M_i}{m}$$

The expression (3) establishes interesting dependencies between the accuracy of a context-based probabilistic WSD model and certain environmental conditions.

## 3.1 Dependency upon the corpus and linguistic concepts

In a complex language domain (e.g. newspaper articles) linguistic phenomena are far less repetitive than in a restricted language (e.g. airline reservations). However, even in a relatively unrestricted domain certain categories are used in a more narrow sense.

Let us consider the probabilistic context-based algorithm in Cucchiarelli, Luzi and Velardi (1998), where a feature is defined by:

$f^k$: (syntactic_relation, w1, $w_i$) (e.g. (N_N district *bank*))

$f^k$ →$C_i$ if $w_i$ reaches the hyperonym $C_i$ in the WordNet on-line taxonomy, and $\phi(i,k) > \gamma$

Using the 1 million word Wall Street Journal corpus, we estimated the following probabilities (3.3) of unseen feature vectors (m in this experiment is $O(10^5)$):

P(unseen in *artifact*)=0,7692

P(unseen in *person*)= 0,7161

P(unseen in psychological feature)=0.8598

[3] In the expression 3) the three events are clearly mutually exclusive.

[4] We here assume for simplicity that the similarity function is an identity. A multinomial or a more complex function must be used in case contexts are considered similar if, for example, co-occurring words have some common hyperonym. See Cucchiarelli, Luzi and Velardi (1998) for examples.

The linguistic concepts *artifact*, *person* and *psychological feature* are three hyperonyms of the on-line WordNet taxonomy. The above figures show that the more "vague" concept *psychological feature* occurs in more variable contexts, though the distribution of words in the three categories is approximately even.

### 3.2 Dependency on the representation model

The representation model H also affects the estimates of erroneous classifications. For example, if we modify the contextual model by removing the information on $w_i$ (that is to say, the feature vectors in the contextual model now only includes the syntactic relation type and the co-occurring word w1), we obtain the following values for the probabilies (3.3):

P(unseen in *artifact*)=0,1778<br>
P(unseen in *person*)= 0,1714<br>
P(unseen in *psychological feature*)=0,2139

The probability of "unseens" in this simpler model is considerably lower (we removed an attribute, $w_i$, that assumes values over V), but clearly, the probability of false positives and false negatives increases.

The motivation is that we now assume that a context for a word belonging (also to) $C_i$ is a valid context for *any* word in that category. Regardless of the specific adopted formula for $\phi(i,k)$, the confidence $\phi(i,k)$ in such a generalization depends on the number of different words $w_i$ in occurring in a given context $f^k$. If this number is low, or is just 1, then the value of $\phi(i,k)$ must be low, accordingly. The selected threshold $\gamma$ then determines the different contribution of false positives and false negatives to the total model accuracy.

A preliminary experiment is illustrated in Figure 2. The figure computes (1-P(fp in $C_i$) for the category *artifact*, as a function of m and $\phi(i,k)$, evaluated on a test set of 78 words.

The figure shows that when $\gamma$ is $\geq$0,5 the number of false positives is rather low, after observing sufficient examples.

On the other side, P(fn outside $C_i$) (not shown here for sake of space) has a specular behavour. For $\gamma$=0,9, the probability of false negative is as low as 0,6.

## 4. Conclusion

By no means the work presented in this paper needs more investigation, especially on the experimental side. However, we believe that learnability analysis of WSD models has strong practical implications.

The quantitative and (preliminary) experimental results of Section 2 put in evidence that :

- In order to acquire statistically stable contextual models of linguistic concepts, the dimension of the analyzed corpora must be considerably high. Paradoxically, untrained probabilistic systems are in better shape in this regard. Very large repositories of language samples can be now obtained from the WWW.
- The experimental setting (i.e. size of the training set) must be tuned for each category and language domain, because the variability of contextual behavior may be significantly different, depending on domain complexity, e.g. the type and grain of the selected category, and the more or less restricted language domain
- it is possible and indeed advisable, for a given WSD algorithm, to determine in a formal way the relation between expected accuracy of the WSD model and the domain and representation complexity. This would allow a better comparison among systems, and an a-priori tuning of the parameters of the disambiguation model.

## References

Anthony M. and Biggs, N. (1997) *Computational Learning Theory* Cambridge University Press, 1997

Bruce R. and Wiebe J., (1999) *Decomposable Modeling in Natural Language Processing*, Computational Linguistics vol. 25, N. 2. 199

Computational Linguistics (1998) *Special Issue on Word Sense Disambiguation*, Vol. 24 (1) March 1988

Cucchiarelli A. Luzi D. and Velardi P. (1998) *Automatic Semantic Tagging of Unknown Proper Names* Proc. of joint 36 ° ACL-17 ° COLING, Montreal, August 1998

Hanson S.J., Petsche T., Kearns M., Rivest R.L. (1994) *Computational Learning Theory and Natural Learning Systems,* Vol. II, MIT Press, 1994

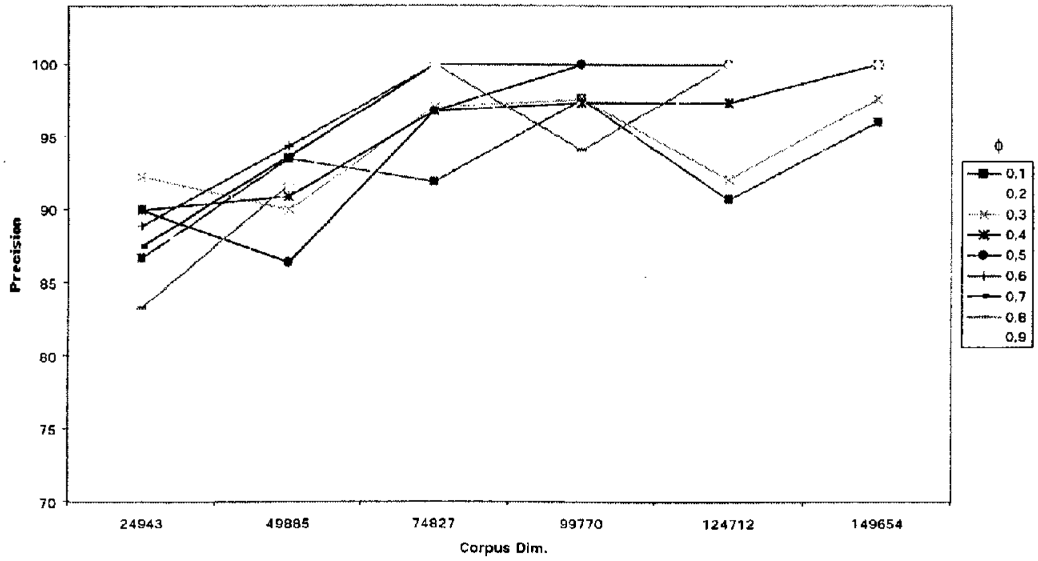
Kearns M.J. and Vazirani U.V. (1994*) An Introduction to Computational Learning Theory* MIT Press, 1994

Russell S.J and Norvig P (1999). *Chapter 18: Learning from Observations* in: *Artificial Intelligence: a modern approach* Prentice-hall 1999

Senseval (1998) homepage: http://www.itri.brighton.ac.uk/events/senseval/

Valiant L. (1984) *A Theory of Learnable* Communications of the ACM, 27(11), 1984

Figure 2: (1-P(fp)) vs. Corpus Dim. For the category Artifact


Precision
Corpus Dim.
ϕ