A Theoretical Analysis of Context-Based Learning Algorithms for Word Sense Disambiguation

Paola Velardi[1] and **Alessandro Cucchiarelli**[2]

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

Word Sense Disambiguation (WSD) is perhaps the most central and difficult task in the area of Natural Language Processing. The problem of WSD is one of identifying the semantic category of an ambiguous word in a sentence context, for example, the financial institution sense of bank in: " *A survey by the Federal Reserve's 12 district banks and the latest report by the National Association of Purchasing Management blurred that picture of the economy.*"

All interesting, large-scale applications of NLP, e.g. Information Retrieval, Filtering and Extraction, Machine Translation and Summarization, etc., suffer performance limitations originated by their limited ability to discriminate the relevant senses of word occurrences in running texts.

Linguistic concepts are rather vague - the notion that the word "bank" belongs to such categories as *human organization* (the financial institution sense) and *location* (the bank-river sense) is more or less intuitive, but in no way it is possible to characterize a linguistic concept in a rigorous way through a mathematical expression. Linguistic concepts are a convention, and even one on which there is little assent.

In NLP, linguistic concepts are often defined as *clusters of words sharing some properties* that can be systematically observed in spoken or written language. A property is a regularity related to the way words are used, or to the internal structure of the entities they represent. In purely context-based algorithms the idea is that, if a group of words share certain properties, this must be reflected by some observable regularity in the use we make of these words in texts.

More semantically oriented approaches use a "deeper" notion of word sense. Sense definitions are manually created using some formal representation language, or automatically extracted from on-line dictionary definitions.

In both cases, the resulting taxonomy, or concept inventory, maintains a considerable degree of "fuzziness", though it may result an acceptable convention for the purpose of certain interesting computational tasks.

In the literature (see [3] for a collection of 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, and a function *f* to compute the membership of a word to a category, given its context in running texts.

A recent large-scale exercise in evaluating WSD programs is Senseval [7].

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[3] of sense distributions, these correlations could not be consistently observed. There are words with fewer senses causing troubles to most systems, while there are words with a very high polysemy and entropy on which all systems obtain good performance.

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 [1][5][8], 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.

2 LEARNING APPROXIMATE DEFINITIONS OF LINGUISTIC CONCEPTS

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

[1] Dipartimento di Scienze dell'Informazione, University of Roma 'La Sapienza', Via Salaria 113, I-00198 Roma, Italy, email: velardi@dsi.uniroma1.it

[2] Istituto di Informatica, University of Ancona, Via Brecce Bianche, I-60131 Ancona, Italy, email:alex@inform.unian.it

[3] A high entropy indicates an even distribution of sense probabilities in the analyzed sublanguage

i) Given a class C of concepts C_i (where C is either a hierarchy or a "flat" concept inventory),
ii) Given a context-based *representation class* H for a concept class C, where H: $\sum * \rightarrow C$ and $\sum$ is a finite alphabet of symbols (e.g. words or word tags),
iii) Given an input space $X \subseteq \sum *$ of encodings of instances in the learner's world, e.g. feature vectors representing contexts around words wj, where wj is a member of C_i,
iv) Given a training sample S of length m :

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

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

formally characterize 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.

Firstly, H must be an "adequate" 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 other hand, if we select an overly complex representation model, including irrelevant features, we run through the so called *overfitting* problem.

Finally, some of the features used in a representation may be dependent on 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 [7], 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 [2].

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 the training set S, given H, C, a learning algorithm L and 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.. With 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 on 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 which does not coincide with the target concept, while δ is the probability, given D, that a particularly unrepresentative 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.

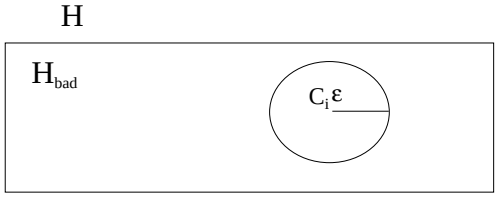

Figure 1. ε-sphere around the "true" function C_i

Figure 1 (from [6]) illustrates the "intuitive" meaning of PAC definition. After seeing m examples, the probability that H_{bad} includes consistent hypotheses is:

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

And we want:

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

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:

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

The inequality (1) establishes a sort of worst-case general bound, but unfortunately this bound turns out to have limited utility in our WSD application.

For example, if the hypothesis space for a linguistic concept C_i is the widely used "*bag of words*" model, 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**,…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:

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

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 L 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 almost all applications concentrated on the field of neural networks and natural learning systems [9]. 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 PROBABILISTIC WSD MODELS

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

Let X be a space of feature vectors:

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

$$b^i_k=1 \text{ if } f_k \text{ is a positive example of } C_i \text{ under H.}$$

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

We assume that arguments are statistically *dependent*, and that a concept is represented as the set of its "typical" context vectors (in case arguments are assumed independent, the representation of a concept is more simple, see [2]).

An example [4] is the case in which f_k is tuple representing a syntactic relation between a word w and another word w_n in its context:

$$f_k: ((\text{synt\_rel\_type}, w_n, w)\ \text{is\_a}(C_i,w))$$

For example, given the compound *district banks* the following feature is generated as an example of the category *organization*: ((N_N district *bank*), is_a(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 *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 [4] 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 ϕ is adjusted to be a probability, i.e. $\sum_i\phi(i,k)=1$. The factor $\phi(i,k)$ represents hence an *estimate* of the probability $Pr(f_k\in C_i)$.

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

$$h(C_i):\{f^i_1..f^i_{mi}\}$$
$$f_k\rightarrow h(C_i)\ \text{iff}\ \phi(i,k) > \gamma \qquad (2)$$

Policy (2) establishes that only feature vectors with a probability higher than a threshold γ 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 which maximizes the similarity between f'_k and one of its members. Again, see [4] and [2] for examples of similarity functions.

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 is true positive for C_i, but was never observed around members of C_i, nor was *similar* (in the precise sense of similarity established by a given algorithm) to any of the vectors in the contextual models.

We then have[4]:

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

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^k_i the number of times the context f_k occurred with a word w' member of C_i

Notice that $\sum_i m^k_i \neq m^k$, 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:

$$\hat{P}(\text{fp})= \sum_{\phi(i,k)>\gamma} \frac{m^k_i}{m}\left(1-\phi(i,k)\right) \qquad (3.1)$$

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

$$\hat{P}(\text{uns. and pos.})= \left(\frac{1}{m}\sum_{\forall m^k=1} m^k\right)\cdot\left(\frac{1}{m}\sum_{\phi(i,k)>\gamma} m^k_i\phi(i,k)\right) \qquad (3.3)$$

The third probability estimate is expressed as the joint

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

probability of extracting a previously unseen context[5], and of extracting positive examples of C_i. Since in (3.1) $(1-\phi(i,k))<(1-\gamma)$, in (3.2) $\phi(i,k)<\gamma$, and in (3.3) $\gamma<\phi(i,k)\leq 1$, we obtain the upper bound:

$$\text{P(w' is misclassified on the basis of f'}_k) \leq \frac{M_i - N_i}{m}(1-\gamma) + \frac{N_i}{m}\gamma + \beta_m \frac{M_i}{m} \quad (4)$$

where (M_i-N_i) is the number of vectors in $h(C_i)$.

We can then impose that (4) $<\varepsilon$, and determine the bound for m. Notice that (4) does not depend on δ. In a noisy learning model the probability of unrepresentative examples is replaced by the probability of noisy examples. In our model we assume that f_k is a positive example for C_i if $\phi(i,k) > \gamma$, therefore we can estimate the noise rate by evaluating the conditional probability on sample data.

$$P(f_k \text{ is fp} / \phi(i,k) > \gamma) \quad (5)$$

Classic methods such as Chernoff bounds [5] may be applied to obtain good approximations for the probabilities (3), (4) and (5) above. Notice however that in order to obtain a given accuracy of estimate, Chernoff bounds (and other methods) impose again bounds on the number of tagged examples needed to compute sufficiently accurate estimates.

Therefore, even in the case of untrained probabilistic learning models, there is the need of a certain amount of tagged examples to verify the validity of certain hypothesis.

4 PRELIMINARY EXPERIMENTAL ANALYSIS

A convincing experimental evaluation of the probabilistic models derived so far is rather demanding, since it requires the preparation of manually tagged test sets for different semantic categories, different language domains, and different contextual and category representation models. Such an evaluation represents our long-term objective and is already in progress.

In [10] we present a preliminary analysis to evaluate the effectiveness of bound (4) to predict the performance of the WSD method [4].

In this section we briefly discuss the dependencies between the accuracy of a context-based probabilistic WSD model and certain "environmental" conditions.

4.1 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 [4], where a feature is defined by:

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

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

[5] We assume here 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 [4] for examples.

Using the one million word Wall Street Journal corpus, we computed the following probabilities of unseen feature vectors:

P(unseen in *artifact*)=0.7692
P(unseen in *person*)= 0.7161
P(unseen in *psychological feature*)=0.8598

The linguistic concepts *artifact*, *person* and *psychological feature* are three hyperonyms of the on-line WordNet taxonomy, a widely used linguistic resource. The above figures show that the "vaguer" concept *psychological feature* occurs in rather sparce contexts, though the distribution of word senses in the three categories is approximately even.

4.2 Dependency on the representation model

The representation model for 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 w_1), we obtain the following:

P(unseen in *artifact*)=0.1778
P(unseen in *person*)= 0.1714
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 γ then determines the different contribution of false positives and false negatives to the total model accuracy.

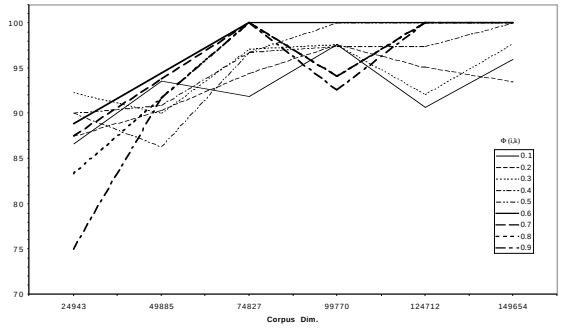

Figure 2. (1-P(fp in C_i)) for the Wnet category artifact

A preliminary experiment is illustrated in Figure 2. The figure plots (1-P(fp in C_i)) for the WordNet category *artifact*, as a function of m and $\phi(i,k)$, evaluated on a test set of 100 words.

The experimental setting is that in which the information on w_i is removed from the contextual model.

The figure shows that when γ is ≥ 0.5 the number of false positives is rather low, after observing sufficient examples.

On the other hand, P(fn outside C_i) (not shown here for sake of space) has a specular behavior. For γ=0.9, the probability of false

negative is as low as 0.6. As expected, with such a high threshold, the contextual model for *artifact* is highly precise, but has a very low recall.

5 CONCLUSIONS

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

The preliminary results of Sections 3 and 4 put in evidence that:

- In order to acquire statistically stable contextual models of linguistic concepts in an untrained setting, the dimension of the analyzed corpora must be considerably high. Paradoxically, untrained probabilistic systems are in better shape in this regard: large repositories of language samples can be now obtained from the WWW.
- Even in untrained systems, a certain amount of samples must be manually tagged to test the system and to accurately estimate the rate of noise during learning
- 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 upon the type and grain of the selected category, and on the 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 environmental and experimental settings. This would allow a better comparison among systems, and an a-priori tuning of the parameters of the disambiguation model.

REFERENCES

[1] Anthony M. and Biggs, N. *Computational Learning Theory* Cambridge University Press, 1997

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

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

[4] Cucchiarelli A. Luzi D. and Velardi P. *Automatic Semantic Tagging of Unknown Proper Names* Proc. of Joint 36° ACL-17° COLING, Montreal, August 1998

[5] Kearns M.J. and Vazirani U.V. *An Introduction to Computational Learning Theory* MIT Press, 1994

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

[7] Senseval homepage: http://www.itri.brighton.ac.uk/events/ senseval/

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

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

[10] Cucchiarelli A., Faggioli E., Velardi P. *Will Very Large Corpora Play For Semantic Disambiguation The Role That Massive Computing Power Is Playing For Other AI-Hard Problems?* LREC 2000, Athens