Unsupervised Sense Disambiguation Using Bilingual Probabilistic Models

Indrajit Bhattacharya
Dept. of Computer Science
University of Maryland
College Park, MD,
USA
indrajit@cs.umd.edu

Lise Getoor
Dept. of Computer Science
University of Maryland
College Park, MD,
USA
getoor@cs.umd.edu

Yoshua Bengio
Dept. IRO
Université de Montréal
Montréal, Québec,
Canada
bengioy@IRO.UMontreal.CA

Abstract

We describe two probabilistic models for unsupervised word-sense disambiguation using parallel corpora. The first model, which we call the Sense model, builds on the work of Diab and Resnik (2002) that uses both parallel text and a sense inventory for the target language, and recasts their approach in a probabilistic framework. The second model, which we call the Concept model, is a hierarchical model that uses a concept latent variable to relate different language specific sense labels. We show that both models improve performance on the word sense disambiguation task over previous unsupervised approaches, with the Concept model showing the largest improvement. Furthermore, in learning the Concept model, as a by-product, we learn a sense inventory for the parallel language.

1 Introduction

Word sense disambiguation (WSD) has been a central question in the computational linguistics community since its inception. WSD is fundamental to natural language understanding and is a useful intermediate step for many other language processing tasks (Ide and Veronis, 1998). Many recent approaches make use of ideas from statistical machine learning; the availability of shared sense definitions (e.g. WordNet (Fellbaum, 1998)) and recent international competitions (Kilgarrif and Rosenzweig, 2000) have enabled researchers to compare their results. Supervised approaches which make use of a small hand-labeled training set (Bruce and Wiebe, 1994; Yarowsky, 1993) typically outperform unsupervised approaches (Agirre et al., 2000; Litkowski, 2000; Lin, 2000; Resnik, 1997; Yarowsky, 1992; Yarowsky, 1995), but tend to be tuned to a specific corpus and are constrained by scarcity of labeled data.

In an effort to overcome the difficulty of finding sense-labeled training data, researchers have begun investigating unsupervised approaches to word-sense disambiguation. For example, the use of parallel corpora for sense tagging can help with word sense disambiguation (Brown et al., 1991; Dagan, 1991; Dagan and Itai, 1994; Ide, 2000; Resnik and Yarowsky, 1999). As an illustration of sense disambiguation from translation data, when the English word *bank* is translated to Spanish as *orilla*, it is clear that we are referring to the *shore* sense of bank, rather than the *financial institution* sense.

The main inspiration for our work is Diab and Resnik (2002), who use translations and linguistic knowledge for disambiguation and automatic sense tagging. Bengio and Kermorvant (2003) present a graphical model that is an attempt to formalize probabilistically the main ideas in Diab and Resnik (2002). They assume the same semantic hierarchy (in particular, WordNet) for both the languages and assign English words as well as their translations to WordNet synsets. Here we present two variants of the graphical model in Bengio and Kermorvant (2003), along with a method to discover a cluster structure for the Spanish senses. We also present empirical word sense disambiguation results which demonstrate the gain brought by this probabilistic approach, even while only using the translated word to provide disambiguation information.

Our first generative model, the *Sense Model*, groups semantically related words from the two languages into *senses*, and translations are generated by probabilistically choosing a sense and then words from the sense. We show that this improves on the results of Diab and Resnik (2002).

Our next model, which we call the *Concept Model*, aims to improve on the above sense structure by modeling the senses of the two languages separately and relating senses from both languages through a higher-level, semantically less precise *concept*. The intuition here is that not all of the senses that are possible for a word will be relevant for a concept. In other words, the distribution over the senses of a word *given* a concept can be expected to have a lower entropy than the distribution over the senses of the word in the language as a whole. In this paper, we look at translation data as a resource for identification of semantic concepts. Note that actual translated word pairs are not always good matches semantically, because the translation process is not on a word by word basis. This introduces a kind of noise in the translation, and an additional hidden variable to represent the shared meaning helps to take it into account. Improved performance over the Sense Model validates the use of concepts in modeling translations.

An interesting by-product of the Concept Model is a semantic structure for the secondary language. This is automatically constructed using background knowledge of the structure for the primary language and the observed translation pairs. In the model, words sharing the same sense are synonyms while senses under the same concept are semantically related in the corpus. An investigation of the model trained over real data reveals that it can indeed group related words together.

It may be noted that predicting senses from translations need not necessarily be an end result in itself. As we have already mentioned, lack of labeled data is a severe hindrance for supervised approaches to word sense disambiguation. At the same time, there is an abundance of bilingual documents and many more can potentially be mined from the web. It should be possible using our approach to (noisily) assign sense tags to words in such documents, thus providing huge resources of labeled data for supervised approaches to make use of.

For the rest of this paper, for simplicity we will refer to the primary language of the parallel document as English and to the secondary as Spanish. The paper is organized as follows. We begin by formally describing the models in Section 2. We describe our approach for constructing the senses and concepts in Section 3. Our algorithm for learning the model parameters is described in Section 4. We present experimental results in Section 5 and our analysis in Section 6. We conclude in Section 7.

2 Probabilistic Models for Parallel Corpora

We motivate the use of a probabilistic model by illustrating that disambiguation using translations is possible even when a word has a unique translation. For example, according to WordNet, the word *prevention* has two senses in English, which may be abbreviated as *hindrance* (the act of hindering or obstruction) and *control* (by prevention, e.g. the control of a disease). It has a single translation in our corpus, that being *prevención*. The first English sense, *hindrance*, also has other words like *bar* that occur in the corpus and all of these other words are observed to be translated in Spanish as the word *obstrucción*. In addition, none of these other words translate to *prevención*. So it is not unreasonable to suppose that the intended sense for *prevention* when translated as *prevención* is different from that of *bar*. Therefore, the intended sense is most likely to be *control*. At the very heart of the reasoning is probabilistic analysis and independence assumptions. We are assuming that senses and words have certain occurrence probabilities and that the choice of the word can be made independently once the sense has been decided. This is the flavor that we look to add to modeling parallel documents for sense disambiguation. We formally describe the two generative models that use these ideas in Subsections 2.2 and 2.3.

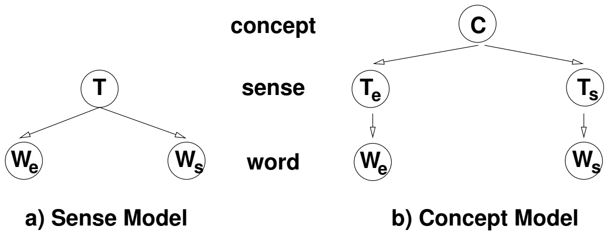

Figure 1: Graphical Representations of the a) Sense Model and the b) Concept Model

2.1 Notation

Throughout, we use uppercase letters to denote random variables and lowercase letters to denote specific instances of the random variables. A translation pair is (W_e, W_s) where the subscript e and s indicate the primary language (English) and the secondary language (Spanish). $W_e \in \{w_{e_1}, \ldots, w_{e_n}\}$ and $W_s \in \{w_{s_1}, \ldots, w_{s_m}\}$. We use the shorthand $P(w_e)$ for $P(W_e = w_e)$.

2.2 The Sense Model

The Sense Model makes the assumption, inspired by ideas in Diab and Resnik (2002) and Bengio and Kermorvant (2003), that the English word W_e and the Spanish word W_s in a translation pair share the same precise sense. In other words, the set of sense labels for the words in the two languages is the same and may be collapsed into one set of senses that is responsible for both English and Spanish words and the single latent variable in the model is the sense label $T \in \{t_1, \ldots, t_k\}$ for both words W_e and W_s. We also make the assumption that the words in both languages are conditionally independent given the sense label. The generative parameters θ^g for the model are the prior probability $P(t)$ of each sense t and the conditional probabilities $P(w_e|t)$ and $P(w_s|t)$ of each word w_e and w_s in the two languages given the sense. The generation of a translation pair by this model may be viewed as a two-step process that first selects a sense according to the priors on the senses and then selects a word from each language using the conditional probabilities for that sense. This may be imagined as a factoring of the joint distribution: $P(W_e, W_s, T) = P(T)P(W_e|T)P(W_s|T)$. Note that in the absence of labeled training data, two of the random variables W_e and W_s are observed, while the sense variable T is not. However, we can derive the possible values for our sense labels from WordNet, which gives us the possible senses for each English word W_e. The Sense model is shown in Figure 1(a).

2.3 The Concept Model

The assumption of a one-to-one association between sense labels made in the Sense Model may be too simplistic to hold for arbitrary languages. In particular, it does not take into account that translation is from sentence to sentence (with a shared meaning), while the data we are modeling are aligned single-word translations (W_e, W_s), in which the intended meaning of W_e does not always match perfectly with the intended meaning of W_s. Generally, a set of m related senses in one language may be translated by one of n related senses in the other. This many-to-many mapping is captured in our alternative model using a second level hidden variable called a *concept*. Thus we have three hidden variables in the Concept Model — the English sense T_e, the Spanish sense T_s and the concept C, where $T_e = \{t_{e_1}, \ldots, t_{e_k}\}$, $T_s = \{t_{s_1}, \ldots, t_{s_j}\}$ and $C = \{c_1, \ldots, c_l\}$.

We make the assumption that the senses T_e and T_s are independent of each other given the shared concept C. The generative parameters θ^g in the model are the prior probabilities $P(c)$ over the concepts, the conditional probabilities $P(t_e|c)$ and $P(t_s|c)$ for the English and Spanish senses given the concept, and the conditional probabilities $P(w_e|t_e)$ and $P(w_s|t_s)$ for the words w_e and w_s in each language given their senses. We can now imagine the generative process of a translation pair by the Concept Model as first selecting a concept according to the priors, then a sense for each language given the concept, and finally a word for each sense using the conditional probabilities of the words. As in Bengio and Kermorvant (2003), this generative procedure may be captured by factoring the joint distribution using the conditional independence assumptions as $P(W_e, W_s, T_e, T_s, C) = P(C)P(T_e|C)P(W_e|T_e)P(T_s|C)P(W_s|T_s)$. The Concept model is shown in Figure 1(b).

3 Constructing the Senses and Concepts

Building the structure of the model is crucial for our task. Choosing the dimensionality of the hidden variables by selecting the number of senses and concepts, as well as taking advantage of prior knowledge to impose constraints, are very important aspects of building the structure.

If certain words are not possible for a given sense, or certain senses are not possible for a given concept, their corresponding parameters should be 0. For instance, for all words w_e that do not belong to a sense t_e, the corresponding parameter $\theta_{w_e|t_e}$ would be permanently set to 0. Only the remaining parameters need to be modeled explicitly.

While model selection is an extremely difficult problem in general, an important and interesting option is the use of world knowledge. Semantic hierarchies for some languages have been built. We should be able to make use of these known taxonomies in constructing our model. We make heavy use of the WordNet ontology to assign structure to both our models, as we discuss in the following subsections. There are two major tasks in building the structure — determining the possible sense labels for each word, both English and Spanish, and constructing the concepts, which involves choosing the number of concepts and the probable senses for each concept.

3.1 Building the Sense Model

Each word in WordNet can belong to multiple synsets in the hierarchy, which are its possible senses. In both of our models, we directly use the WordNet senses as the English sense labels. All WordNet senses for which a word has been observed in the corpus form our set of English sense labels. The Sense Model holds that the sense labels for the two domains are the same. So we must use the same WordNet labels for the Spanish words as well. We include a Spanish word w_s for a sense t if w_s is the translation of any English word w_e in t.

3.2 Building the Concept Model

Unlike the Sense Model, the Concept Model does not constrain the Spanish senses to be the same as the English ones. So the two major tasks in building the Concept Model are constructing the Spanish senses and then clustering the English and Spanish senses to build the concepts.

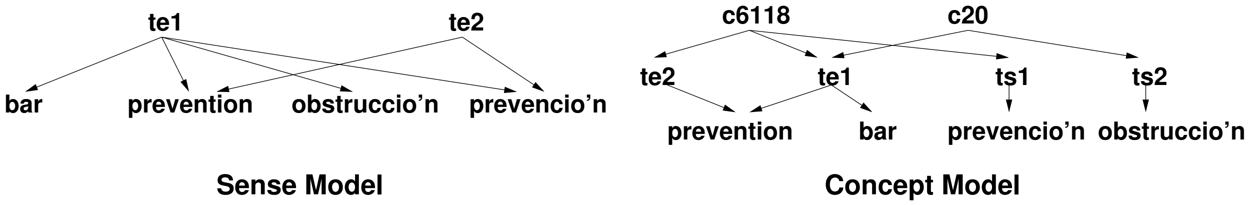

Figure 2: The Sense and Concept models for *prevention, bar, prevención* and *obstrucción*

For each Spanish word w_s, we have its set of English translations $\{w_{e_1}, \ldots, w_{e_k}\}$. One possibility is to group Spanish words looking at their translations. However, a more robust approach is to consider the relevant English senses for w_s. Each English translation for w_s has its set of English sense labels $S_{w_{e_i}}$ drawn from WordNet. So the relevant English sense labels for w_s may be defined as $S_{w_s} = \bigcup_i S_{w_{e_i}}$. We call this the *English sense map* or $sMap$ for w_s. We use the $sMap$s to define the Spanish senses. We may imagine each Spanish word to come from one or more Spanish senses. If each word has a single sense, then we add a Spanish sense t_s for each $sMap$ and all Spanish words that share that $sMap$ belong to that sense. Otherwise, the $sMap$s have to be split into frequently occurring subgroups. Frequently co-occurring subsets of $sMap$s can define more refined Spanish senses. We identify these subsets by looking at pairs of $sMap$s and computing their intersections. An intersection is considered to be a Spanish sense if it occurs for a significant number of pairs of $sMap$s. We consider both ways of building Spanish senses. In either case, a constructed Spanish sense t_s comes with its relevant set $\{t_{e_i}\}$ of English senses, which we denote as $sMap(t_s)$.

Once we have the Spanish senses, we cluster them to form concepts. We use the $sMap$ corresponding to each Spanish sense to define a measure of similarity for a pair of Spanish senses. There are many options to choose from here. We use a simple measure that counts the number of common items in the two $sMap$s.[1] The similarity measure is now used to cluster the Spanish senses t_s. Since this measure is not transitive, it does not directly define equivalence classes over t_s. Instead, we get a similarity graph where the vertices are the Spanish senses and we add an edge between two senses if their similarity is above a threshold. We now pick each connected component from this graph as a cluster of similar Spanish senses.

Now we build the concepts from the Spanish sense clusters. We recall that a concept is defined by a set of English senses and a set of Spanish senses that are related. Each cluster represents a concept. A particular concept is formed by the set of Spanish senses in the cluster and the English senses relevant for them. The relevant English senses for any Spanish sense is given by its $sMap$. Therefore, the union of the $sMap$s of all the Spanish senses in the cluster forms the set of English senses for each concept.

4 Learning the Model Parameters

Once the model is built, we use the popular EM algorithm (Dempster et al., 1977) for hidden variables to learn the parameters for both models. The algorithm repeatedly iterates over two steps. The first step maximizes the expected log-likelihood of the joint probability of the observed data with the current parameter settings θ^g. The next step then re-estimates the values of the parameters of the model. Below we summarize the re-estimation steps for each model.

4.1 EM for the Sense Model

$$P(T_i = t) = \frac{1}{N} \sum_{i=1}^{N} P(T = t | w_{e_i}, w_{s_i}, \theta^g)$$

$$P(W_{e_i} = e | T_i = t) = \frac{\sum_{w_{e_i}=e, i=1}^{N} P(T = t | w_{e_i}, w_{s_i}, \theta^g)}{\sum_e \sum_{W_{e_i}=e, i=1}^{N} P(T = t | w_{e_i}, w_{s_i}, \theta^g)}$$

$P(W_{s_i} = s | T_i = t)$ follows similarly.

4.2 EM for the Concept Model

$$P(C_i = k) = \frac{1}{N} \sum_{i=1}^{N} P(C_i = k | w_{e_i}, w_{s_i}, \theta^g)$$

$$P(T_{e_i} = l | C_i = k) = \frac{\sum_{i=1}^{N} P(C_i = k, T_{e_i} = l | w_{e_i}, w_{s_i}, \theta^g)}{\sum_{i=1}^{N} P(C_i = k | w_{e_i}, w_{s_i}, \theta^g)}$$

[1] Another option would be to use a measure of similarity for English senses, proposed in Resnik (1995) for two synsets in a concept hierarchy like WordNet. Our initial results with this measure were not favorable.

$$P(W_{e_i} = e|T_{e_i} = l) = \frac{\sum_{W_{e_i}=e,i=1}^{N} P(T_{e_i} = l|w_{e_i} = e, w_{s_i}, \theta^g)}{\sum_e \sum_{W_{e_i}=e,i=1}^{N} P(T_{e_i} = l|W_{e_i} = e, w_{s_i}, \theta^g)}$$

$P(T_{s_i} = m|C_i = k)$ and $P(W_{s_i} = s|T_{s_i} = m)$ follow similarly.

4.3 Initialization of Model Probabilities

Since the EM algorithm performs gradient ascent as it iteratively improves the log-likelihood, it is prone to getting caught in local maxima, and selection of the initial conditions is crucial for the learning procedure. Instead of opting for a uniform or random initialization of the probabilities, we make use of prior knowledge about the English words and senses available from WordNet. WordNet provides occurrence frequencies for each synset in the SemCor Corpus that may be normalized to derive probabilities $P_{wn}(t_e)$ for each English sense t_e. For the Sense Model, these probabilities form the initial priors over the senses, while all English (and Spanish) words belonging to a sense are initially assumed to be equally likely. However, initialization of the Concept Model using the same knowledge is trickier. We would like each English sense t_e to have $P_{init}(t_e) = P_{wn}(t_e)$. But the fact that each sense belongs to multiple concepts and the constraint $\sum_{t_e \in c} P(t_e|c) = 1$ makes the solution non-trivial. Instead, we settle for a compromise. We set $P_{init}(t_e|c) = P_{wn}(t_e)$ and $P(c) = \sum_{t_e \in c} P_{wn}(t_e)$. Subsequent normalization takes care of the sum constraints. For a Spanish sense, we set $P(t_s) = \sum_{t_e \in sMap(t_s)} P_{wn}(t_e)$. Once we have the Spanish sense probabilities, we follow the same procedure for setting $P(t_s|c)$ for each concept. All the Spanish and English words for a sense are set to be equally likely, as in the Sense Model. It turned out in our experiments on real data that this initialization makes a significant difference in model performance.

5 Experimental Evaluation

Both the models are generative probabilistic models learned from parallel corpora and are expected to fit the training and subsequent test data. A good fit should be reflected in good prediction accuracy over a test set. The prediction task of interest is the sense of an English word when its translation is provided. We estimate the prediction accuracy and recall of our models on Senseval data.[2] In addition, the Concept Model learns a sense structure for the Spanish language. While it is hard to objectively evaluate the quality of such a structure, we present some interesting concepts that are learned as an indication of the potential of our approach.

5.1 Evaluation with Senseval Data

In our experiments with real data, we make use of the parallel corpora constructed by Diab and Resnik (2002) for evaluation purposes. We chose to work on these corpora in order to permit a direct comparison with their results. The sense-tagged portion of the English corpus is comprised of the English "all-words" section of the SENSEVAL-2 test data. The remainder of this corpus is constructed by adding the Brown Corpus, the SENSEVAL-1 corpus, the SENSEVAL-2 English Lexical Sample test, trial and training corpora and the Wall Street Journal sections 18-24 from the Penn Treebank. This English corpus is translated into Spanish using two commercially available MT systems: Globalink Pro 6.4 and Systran Professional Premium. The GIZA++ implementation of the IBM statistical MT models was used to derive the most-likely word-level alignments, and these define the English/Spanish word co-occurrences. To take into account variability of translation, we combine the translations from the two systems for each English word, following in the footsteps of Diab and Resnik (2002). For our experiments, we focus only on nouns, of which there are 875 occurrences in our tagged data. The sense tags for the English domain are derived from the WordNet 1.7 inventory. After pruning stopwords, we end up with 16,186 English words, 31,862 Spanish words and 2,385,574 instances of 41,850 distinct translation pairs. The English words come from 20,361 WordNet senses.

Table 1: Comparison with Diab's Model

Model	Accuracy	Recall	Parameters
Diab	0.618	0.572	-
Sense M.	0.624	0.616	154,947
Concept M.	0.672	0.651	120,268

As can be seen from the following table, both our models clearly outperform Diab (2003), which is an improvement over Diab and Resnik (2002), in both accuracy and recall, while the Concept Model does significantly better than the Sense Model with fewer parameters. The comparison is restricted to the same subset of the test data. For our best results, the Sense Model has 20,361 senses, while the Concept Model has 20,361 English senses, 11,961 Spanish senses and 7,366 concepts. The Concept Model results are for the version that allows multiple senses for a Spanish word. Results for the

[2] Accuracy is the ratio of the number of correct predictions and the number of attempted predictions. Recall is the ratio of the number of correct predictions and the size of the test set.

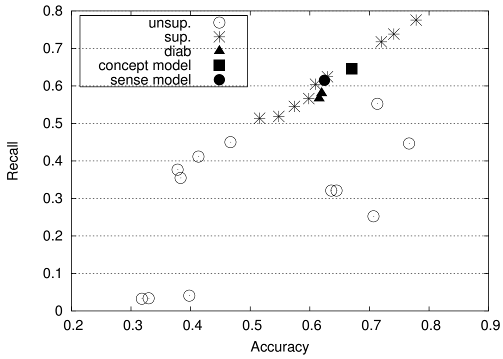

Figure 3: Comparison with Senseval2 Systems

single-sense model are similar.

In Figure 3, we compare the prediction accuracy and recall against those of the 21 Senseval-2 English All Words participants and that of Diab (2003), when restricted to the same set of noun instances from the gold standard. It can be seen that our models outperform all the unsupervised approaches in recall and many supervised ones as well. No unsupervised approach is better in both accuracy and recall. It needs to be kept in mind that we take into account *only bilingual data* for our predictions, and not monolingual features like context of the word as most other WSD approaches do.

5.2 Semantic Grouping of Spanish Senses

Table 2 shows some interesting examples of different Spanish senses for discovered concepts.[3] The context of most concepts, like the ones shown, can be easily understood. For example, the first concept is about government actions and the second deals with murder and accidental deaths. The penultimate concept is interesting because it deals with different kinds of *association* and involves three different senses containing the word *conexión*. The other words in two of these senses suggest that they are about *union* and *relation* respectively. The third probably involves the *link* sense of *connection*. Conciseness of the concepts depends on the similarity threshold that is selected. Some may bring together loosely-related topics, which can be separated by a higher threshold.

6 Model Analysis

In this section, we back up our experimental results with an in-depth analysis of the performance of our two models.

Our Sense Model was motivated by Diab and Resnik (2002) but the flavors of the two are quite different. The most important distinction is that the Sense Model is a probabilistic generative model for parallel corpora, where interaction between different words stemming from the same sense comes into play, even if the words are not related through translations, and this interdependence of the senses through common words plays a role in sense disambiguation.

We started off with our discussions on semantic ambiguity with the intuition that identification of semantic concepts in the corpus that relate multiple senses should help disambiguate senses. The Sense Model falls short of this target since it only brings together a single sense from each language. We will now revisit the motivating example from Section 2 and see how concepts help in disambiguation by grouping multiple related senses together.

For the Sense Model, $P(prevention|t_{e_2}) > P(prevention|t_{e_1})$ since it is the only word that t_{e_2} can generate. However, this difference is compensated for by the higher prior probability $P(t_{e_1})$, which is strengthened by both the translation pairs. Since the probability of joint occurrence is given by the product $P(t)P(w_e|t)P(w_s|t)$ for any sense t, the model does not develop a clear preference for any of the two senses.

The critical difference in the Concept Model can be appreciated directly from the corresponding joint probability $P(c)P(t_e|c)P(w_e|t_e)P(t_s|c)P(w_s|t_s)$, where c is the relevant concept in the model. The preference for a particular instantiation in the model is dependent not on the prior $P(t_e)$ over a sense, but on the sense conditional $P(t_e|c)$. In our example, since *<bar, obstrucción>* can be generated only through concept $c20$, $P(t_{e_1}|c20)$ is the only English sense conditional boosted by it. *<prevention, prevención>* is generated through a different concept $c6118$, where the higher conditional $P(prevention|t_{e_2})$ gradually strengthens one of the possible instantiations for it, and the other one becomes increasingly unlikely as the iterations progress. The inference is that only one sense of *prevention* is possible in the context of the parallel corpus. The key factor in this disambiguation was that two senses of *prevention* separated out in two different concepts.

The other significant difference between the models is in the constraints on the parameters and the effect that they have on sense disambiguation. In the Sense Model, $\sum_t P(t) = 1$, while in the Concept Model, $\sum_{t_e \in c} P(t_e|c) = 1$ *separately for each concept c*. Now for two relevant senses for an English word, a slight difference in their priors will tend to get ironed out when normalized over the en-

[3] Some English words are found to occur in the Spanish Senses. This is because the machine translation system used to create the Spanish document left certain words untranslated.

Table 2: Example Spanish Senses in a Concept. For each concept, each row is a separate sense. Dictionary senses of Spanish words are provided in English within parenthesis where necessary.

actos supremas decisión decisiones gobernando gobernante gubernamentales gobernación gobierno-proporciona prohibir prohibiendo prohibitivo prohibitiva gubernamental gobiernos	accidente accidentes muertes(*deaths*) casualty matar(*to kill*) matanzas(*slaughter*) muertes-le slaying derramamiento-de-sangre (*spilling-of-blood*) cachiporra(*bludgeon*) obligar(*force*) obligando(*forcing*) asesinato(*murder*) asesinatos

linterna-eléctrica linterna(*lantern*)
faros-automóvil(*headlight*)
linternas-portuarias(*harbor-light*)
antorcha(*torch*) antorchas antorchas-pino-nudo
	manía craze
	culto(*cult*) cultos proto-senility
	delirio delirium
	rabias(*fury*) rabia farfulla(*do hastily*)

oportunidad oportunidades ocasión ocasiones riesgo(*risk*) riesgos peligro(*danger*) destino sino(*fate*) fortuna suerte(*fate*) probabilidad probabilidades	diferenciación distinción distinciones especialización maestría (*mastery*) peculiaridades particularidades peculiaridades-inglesas especialidad especialidades
diablo(*devil*) diablos dickens heller lucifer satan satanás	modelo parangón ideal ideales santo(*saint*) santos san idol idols ídolo

deslumbra(*dazzle*)
cromo(*chromium*)
meteoro meteoros meteor meteoros-blue
meteorito meteoritos
pedregosos(*rocky*)
	dios god dioses
	divinidad divinity
	inmortal(*immortal*) inmortales
	teología teolog
	deidad deity deidades

variación variaciones discordancia desacuerdo(*discord*) discordancias desviación(*deviation*) desviaciones desviaciones-normales discrepancia discrepancias fugaces(*fleeting*) variación diferencia disensión	minutos minuto momento momentos un-momento minutos momentos momento segundos instante momento pestañeo(*blink*) guiña(*wink*) pestañean

adhesión adherencia ataduras(*tying*)
enlace(*connection*) ataduras
atadura ataduras
conexión conexiones
conexión une(*to unite*)
relación conexión
implicación (*complicity*) envolvimiento
	pasillo(*corridor*)
	aisle
	pasarela(*footbridge*)
	hall vestíbulos
	pasaje(*passage*)
	callejón(*alley*) callejas-ciegas (*blind alley*) callejones-ocultos

tire set of senses for the corpus. In contrast, if these two senses belong to the same concept in the Concept Model, the difference in the sense conditionals will be highlighted since the normalization occurs over a very small set of senses — the senses for only that concept, which in the best possible scenario will contain only the two contending senses, as in concept $c118$ of our example.

As can be seen from Table 1, the Concept Model not only outperforms the Sense Model, it does so with significantly fewer parameters. This may be counter-intuitive since Concept Model involves an extra concept variable. However, the dissociation of Spanish and English senses can significantly reduce the parameter space. Imagine two Spanish words that are associated with ten English senses and accordingly each of them has a probability for belonging to each of these ten senses. Aided with a concept variable, it is possible to model the same relationship by creating a separate Spanish sense that contains these two words and relating this Spanish sense with the ten English senses through a concept variable. Thus these words now need to belong to only one sense as opposed to ten. Of course, now there are new transition probabilities for each of the eleven senses from the new concept node. The exact reduction in the parameter space will depend on the frequent subsets discovered for the $sMap$s of the Spanish words. Longer and more frequent subsets will lead to larger reductions. It must also be borne in mind that this reduction comes with the independence assumptions made in the Concept Model.

7 Conclusions and Future Work

We have presented two novel probabilistic models for unsupervised word sense disambiguation using parallel corpora and have shown that both models outperform existing unsupervised approaches. In addition, we have shown that our second model, the Concept model, can be used to learn a sense inventory for the secondary language. An advantage of the probabilistic models is that they can easily incorporate additional information, such as context information. In future work, we plan to investigate the use of additional monolingual context. We would also like to perform additional validation of the learned secondary language sense inventory.

8 Acknowledgments

The authors would like to thank Mona Diab and Philip Resnik for many helpful discussions and insightful comments for improving the paper and also for making their data available for our experiments. This study was supported by NSF Grant 0308030.

References

E. Agirre, J. Atserias, L. Padr, and G. Rigau. 2000. Combining supervised and unsupervised lexical knowledge methods for word sense disambiguation. In *Computers and the Humanities, Special Double Issue on SensEval. Eds. Martha Palmer and Adam Kilgarriff. 34:1,2.*

Yoshua Bengio and Christopher Kermorvant. 2003. Extracting hidden sense probabilities from bitexts. Technical report, TR 1231, Departement d'informatique et recherche operationnelle, Universite de Montreal.

Peter F. Brown, Stephen Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1991. Word-sense disambiguation using statistical methods. In *Meeting of the Association for Computational Linguistics*, pages 264–270.

Rebecca Bruce and Janyce Wiebe. 1994. A new approach to sense identification. In *ARPA Workshop on Human Language Technology*.

Ido Dagan and Alon Itai. 1994. Word sense disambiguation using a second language monolingual corpus. *Computational Linguistics*, 20(4):563–596.

Ido Dagan. 1991. Lexical disambiguation: Sources of information and their statistical realization. In *Meeting of the Association for Computational Linguistics*, pages 341–342.

A.P. Dempster, N.M. Laird, and D.B. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, B 39:1–38.

Mona Diab and Philip Resnik. 2002. An unsupervised method for word sense tagging using parallel corpora. In *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL-02).*

Mona Diab. 2003. *Word Sense Disambiguation Within a Multilingual Framework*. Ph.D. thesis, University of Maryland, College Park.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

Nancy Ide and Jean Veronis. 1998. Word sense disambiguation: The state of the art. *Computational Linguistics*, 28(1):1–40.

Nancy Ide. 2000. Cross-lingual sense determination: Can it work? In *Computers and the Humanities: Special Issue on Senseval, 34:147-152.*

Adam Kilgarrif and Joseph Rosenzweig. 2000. Framework and results for english senseval. *Computers and the Humanities*, 34(1):15–48.

Dekang Lin. 2000. Word sense disambiguation with a similarity based smoothed library. In *Computers and the Humanities: Special Issue on Senseval, 34:147-152.*

K. C. Litkowski. 2000. Senseval: The cl research experience. In *Computers and the Humanities, 34(1-2), pp. 153-8.*

Philip Resnik and David Yarowsky. 1999. Distinguishing systems and distinguishing senses: new evaluation methods for word sense disambiguation. *Natural Language Engineering*, 5(2).

Philip Resnik. 1995. Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 448–453.

Philip Resnik. 1997. Selectional preference and sense disambiguation. In *Proceedings of ACL Siglex Workshop on Tagging Text with Lexical Semantics, Why, What and How?*, Washington, April 4-5.

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

David Yarowsky. 1993. One sense per collocation. In *Proceedings, ARPA Human Language Technology Workshop, Princeton.*

David Yarowsky. 1995. Unsupervised word sense disambiguation rivaling supervised methods. In *Meeting of the Association for Computational Linguistics*, pages 189–196.