Semi-supervised Semantic Role Labeling

Cynthia A. Thompson*
School of Computing, University of Utah
Salt Lake City, UT 84112
cindi@cs.utah.edu

Abstract

Most corpus-based approaches to language learning have focused on tasks for which a sufficient amount of human-labeled training data is available. However, it is difficult to produce such data, and models trained from such data tend to be brittle when applied to domains that vary, even in seemingly minor ways, from the training data. We claim that these difficulties can by overcome by applying *semi-supervised learning* techniques. Semi-supervised techniques learn from both labeled and "raw" data. In our case, the latter is raw text. Several researchers have used semi-supervised techniques for language learning (Nigam *et al.* 2000; Blum & Mitchell 1998; Joachims 1999; Riloff & Jones 1999), but we believe that this area is not yet well explored and definitely not well understood. Therefore, we present a challenge problem for semi-supervised learning: semantic role labeling and semantic relationship annotation. Semantic role labeling was introduced by Gildea & Jurafsky (2002), and we added semantic relationship annotation in Thompson, Levy, & Manning (2003). This problem is a difficult one for semi-supervised techniques, for three reasons. First, there are many possible classes (the role labels) for examples. Second, sequence learning is involved. Third, the learning scenario is plagued by sparse data problems. We describe the role labeling problem, our learning model and its extendibility to semi-supervised learning, and some preliminary experiments.

Semantic Role Labeling

Extracting semantic meaning from the surface form of sentences is important for many understanding tasks performed by human and machine alike. These include inference, antecedent resolution, word sense disambiguation, or even syntactic tasks such as prepositional phrase attachment. We focus here on a particular type of semantic analysis, that of determining the semantic roles and relationships at play in a sentence. For example, in the sentence

"The hammer broke the vase,"

we may want to find the participants, or roles, in the breaking action. In this case the roles are *instrument*, filled by "the hammer," and *patient*, filled by "the vase." We may also care about the over-arching relationship created by these roles and the breaking action, which by the theory we are using is the CAUSE_TO_FRAGMENT relationship. Having semantic roles allows one to recognize the semantic arguments of a situation, even when expressed in different syntactic configurations. For example, in

"I broke the vase with the hammer,"

the vase and hammer play the same roles as in the previous sentence.

We have developed a generative model for performing this type of semantic analysis. Generative models are probability models representing a joint distribution over a set of variables. The specific probability settings are called parameters. As pointed out by Jurafsky (2003) and others, probabilistic models capture cognitively plausible aspects of human language processing, generation, and learning. Our model takes as input the constituents of a sentence and a *predicator* word (or phrase) from that sentence. In the previous example *broke* is the predicator. The predicator takes semantic role arguments, instantiated by the constituents. We learn the parameters for this model from a body of examples provided by the FrameNet corpus (Baker, Fillmore, & Lowe 1998). The problem and some elements of our approach are similar to that of Gildea & Jurafsky (2002), but our work differs by use of a generative, not a discriminative (conditional), model. We also add the inference of the over-arching relationship between the roles, called the *frame*.

FrameNet

Our model was inspired by FrameNet, a large-scale, domain-independent computational lexicography project organized around the motivating principles of lexical semantics: that systematic correlations can be found between the meaning components of words, principally the semantic roles associated with events, and their combinatorial properties in syntax. This principle has been instantiated at various levels of granularity in different traditions of linguistic research; FrameNet researchers work at an intermediate level of granularity, termed the *frame*. Examples of frames

*Some of this work was supported by an ARDA Aquaint grant while the author was a Visiting Assistant Professor at Stanford University. Carolin Arnold at the University of Utah was crucial in running the EM experiments.

include MOTION_DIRECTIONAL, CONVERSATION, JUDGMENT, and TRANSPORTATION. Frames consist of multiple *lexical units*—items corresponding to a sense of a word. Examples for the MOTION_DIRECTIONAL frame are *drop* and *plummet*. Also associated with each frame is a set of semantic *roles*. Examples for the MOTION_DIRECTIONAL frame include the moving object, called the THEME; the ultimate destination, the GOAL; the SOURCE; and the PATH.

In addition to frame and role definitions, FrameNet has produced a large number of role-annotations for sentences that are drawn primarily from the British National Corpus. There are two releases of the corpus, FrameNet I and FrameNet II. The corpus identifies a a lexical unit of interest, which takes arguments, for each annotated sentence. We will call this word the *predicator*.[1] The words and phrases that participate in the predicator's meaning are labeled with their roles, and the entire sentence is labeled with the relevant frame. Finally, the corpus also includes syntactic category information for each role. We give some examples below, with the frame listed in braces at the beginning, the predicator in bold, and each relevant constituent labeled with its role and phrase type.

{MOTION_DIRECTIONAL} Mortars lob heavy shells high into the sky so that [$^{\text{NP}}_{\text{THEME}}$ they] **drop** [$^{\text{PP}}_{\text{PATH}}$ down] [$^{\text{PP}}_{\text{GOAL}}$ on the target] [$^{\text{PP}}_{\text{SOURCE}}$ from the sky].

{ARRIVING} He heard the sound of liquid slurping in a metal container as [$^{\text{NP}}_{\text{THEME}}$ Farrell] **approached** [$^{\text{NP}}_{\text{GOAL}}$ him] [$^{\text{PP}}_{\text{SOURCE}}$ from behind].

A Generative Model for Sentence-Role Labeling

Our goal is to identify frames and roles, given a natural language sentence and predicator. We developed a generative model that defines a joint probability distribution over predicators, frames, roles, and constituents. While the model is fully general in its ability to determine these variables, in this paper it is only tested on its ability to determine roles and frames when given *both* a list of constituents and a single predicator. The generative model, illustrated in Figure 1, functions as follows. First, a predicator, D, generates a frame, F. The frame generates a (linearized) role sequence, R_1 through R_n, which in turn generates each constituent of the sentence, C_1 through C_n. Note that, conditioned on a particular frame, the model is just a Hidden Markov Model. In that terminology, *transition* probabilities are those associated with roles following each other in the sequence, and *emission* probabilities are those associated with roles generating ("emitting") frames.

The FrameNet corpus contains annotations for all of the model components described above. To simplify the model, we chose to represent each constituent by its phrase type together with the head word of that constituent. So for the **approached** example given earlier, C_1 = Farrel/NP, C_2 = him/NP, and C_3 = from/PP. Most of the parameters for the model are estimated using a straightforward maximum likelihood estimate based on fully labeled training data. Emission probabilities are smoothed with phrase type information, due to the sparseness of head words.

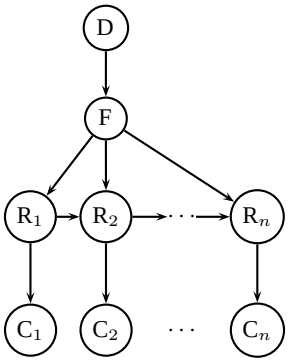

Figure 1: Role Tagger

To label a sequence of constituents and a predicator with a sequence of role labels and frame, we use a generalized Viterbi algorithm that calculates the most likely configuration of all the hidden variables. For further details, see Thompson, Levy, & Manning (2003).

Role Labeling Experiments

To test the above model, we trained it on annotated FrameNet I data, randomly dividing the data into a training set used to estimate the parameters and an unseen test set. We randomly split the sentences of each frame so that 70% were in the training set and 10% were in the test set. We report on three types of accuracy. First, role labeling accuracy is the percent of constituents correctly labeled. Second, full sentence accuracy is the percent of sentences for which all roles are correctly labeled. Finally, frame accuracy is the proportion of sentences for which the model chose the correct frame.

For a baseline comparison, we computed the accuracy of a zeroth-order Markov model that treats all transition probabilities between roles as uniform. We also computed the accuracy of choosing, for all constituents, the most common role given the predicator (BasePredicator), and the accuracy of first choosing a frame, and then choosing the most common role given the frame (BaseFrame); the frame is the most common for the known predicator ($\arg\max_F P(F|D)$).

Table 1 summarizes the results, providing accuracy on both the training (Trn) and test (Tst) sets.

Semi-Supervised Learning

Semi-supervised learning, as the name suggests, uses both learning under complete supervision (with fully labeled examples) and learning with little or no supervision (from unlabeled examples). Humans are able to learn from one or just

[1] What we call the *predicator* is called the *target* in the FrameNet theory, and what we are calling a *(semantic) role* is called in FrameNet a *frame element*, while what we call a *constituent* or *argument head*, (Gildea & Jurafsky 2002) call simply the *head*. We have found that most people find the FrameNet terminology rather confusing, and so have adopted alternative terms here.

System	Trn Role	Tst Role	Trn Full	Tst Full	Tst Frame
FirstOrder	86.1%	79.3%	75.4%	65.3%	97.5%
ZeroOrder	–	60.0%	–	34.6%	96.5%
BasePredicator	39.9%	39.2%	10.5%	10.2%	N/A
BaseFrame	37.8%	37.6%	9.2%	9.5%	N/A

Table 1: FrameNet I Experimental Results. Key: Role=Role labeling accuracy, Full=full sentence accuracy, Frame=Frame choice accuracy. Trn=Training Set, Tst=Test Set.

a few labeled examples, so modeling this with automated techniques is desirable. Besides this, the expense of labeling data argues for the practical benefits of semi-supervised learning. Generative models are a natural choice for situations in which some variables have known values (are labeled) and others are unknown (unlabeled, or partially labeled). Nigam *et al.* (2000) demonstrated the use of Naive Bayes (the most simple of generative models) and the EM algorithm (Dempster, Laird, & Rubin 1977) in the context of semi-supervised learning for text classification. The model learned from texts, some of which were labeled with a category, and some of which were not. This work can be extended and applied to our generative model for semantic role labeling.

In preliminary experiments, we used EM with both labeled and unlabeled data to estimate the parameters of our model. We did not re-estimate the probabilities associated with predicators generating frames or frames generating roles, but only of roles transitioning to other roles and roles generating constituents. With very little tweaking of parameters, the first results of this experiment were negative. The model trained on both labeled and unlabeled data performed *worse* on unseen test data than the model trained only on labeled data. We decided to focus only on the frames occurring most frequently in the data, hypothesizing that semi-supervised learning would help most for common roles. So we determined the seven most common frames in Framenet II and divided up a subset of that data into a labeled and unlabeled set, evaluating the accuracy of trained models on the unlabeled set only. Figure 2 shows the resulting role labeling accuracy. On the first round of EM, the accuracy increases slightly, from 67% to 70.6%, but then decreases in subsequent EM iterations. While the accuracy never gets worse than the initial accuracy, the improvement over the initial accuracy is small.

We believe that the primary reason for the mixed performance of EM is that the current model does not capture all the factors governing the generation of sentences. As discussed in Nigam *et al.* (2000), a mismatch between the model and the actual generative process can lead to a situation in which the unlabeled data overwhelms the small number of labeled examples. We plan to weight the labeled data more strongly than the unlabeled data in future experiments to attempt to partially overcome this mismatch.

A second possible reason for the poor performance is that we are not adjusting the probabilities of predicators generating frames. Thus the model does not benefit from the new predicators in the unlabeled data. We used standard Baum Welch to estimate the Hidden Markov Model parameters, so we will have to add another layer of EM to this to estimate the other parameters.

In addition to this application of EM to our role labeling problem, there are other possible approaches to semi-supervised learning for this and other language learning tasks. For example, a variant of co-training (Blum & Mitchell 1998) could be attempted for this task. Co-training is based on the idea that some features that describe an example are redundant. Therefore, the features can be split into two (or more) sets each of which is sufficient for correct classification. Co-training does so, building two classifiers which provide labeled examples to each other from a large unlabeled pool of examples. Since our current feature set consists only of constituents, we would need to add more features for the second learner in a co-training scenario. The challenge would be to find features satisfying the independence and redundancy requirements of co-training. However, an alternative would be to adapt co-training by using multiple classifiers instead of multiple feature sets. So, for example, a discriminative classifier could provide a second source of labeled data, perhaps improving the performance of our generative model.

A different option for addressing the problem of limited amounts of training data is to turn to active learning, in which the learner and a teacher collaborate in labeling examples that are either chosen or constructed by the learner (Cohn, Atlas, & Ladner 1994). This type of situation certainly happens in human language learning, for example children asking for the names of objects and parents supplying them. From the computational perspective and in our situation, a system could choose sentences for annotation based on the certainty of their role and frame annotations, as indicated by the probability of the labeling. Using active learning would not be as desirable, from the point of view of labeling effort, as eliminating human interaction completely. However, this would likely lead to a more accurate model using fewer labeled examples than if a human labeled examples at random (Thompson, Califf, & Mooney 1999).

Conclusions and Future Work

Our next steps are to perform further experimentation with semi-supervised learning as outlined in the previous section, to show that such learning is broadly beneficial. Then, we will move on to applying the technique to new domains. At that point, evaluation of the learned model will be an issue since currently FrameNet "ground truth" annotations are only available for a limited domain. Human evaluation of the results will be needed, and the results should have the

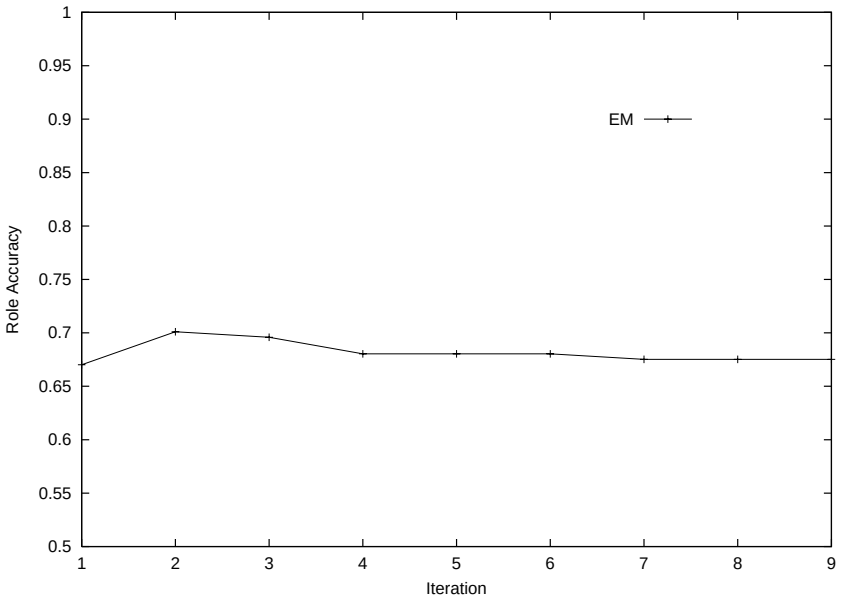

Figure 2: Role Labeling Accuracy with EM.

added benefit of shedding further light on the generality of the FrameNet theory.

We have introduced a new problem for semi-supervised learning, that of making the best use of raw data to supplement scarce labeled data for learning to produce semantic roles and relationships. This is an important problem due to the difficulty of hand-labeling, and an approachable problem due to the large amounts of available unlabeled language data. It is also needed to demonstrate the feasibility and usability of role labeling for a variety of domains. In a more speculative vein, this work may even serve to connect with work in grounded language learning, where a teacher is present but explicit labels typically are not. Finally, role labeling presents many challenges for semi-supervised learning, which we believe will further illuminate the space of problems for which it is possible to learn from small amounts of labeled data.

References

Baker, C. F.; Fillmore, C. J.; and Lowe, J. B. 1998. The Berkeley FrameNet project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and COLING-98*, 86–90.

Blum, A., and Mitchell, T. 1998. Combining labeled and unlabeled data with co-training. In *Proceedings of the 11th Annual Conference on Computational Learning Theory*, 92–100.

Cohn, D.; Atlas, L.; and Ladner, R. 1994. Improving generalization with active learning. *Machine Learning* 15(2):201–221.

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

Gildea, D., and Jurafsky, D. 2002. Automatic labeling of semantic roles. *Computational Linguistics* 28:245–288.

Joachims, T. 1999. Transductive inference for text classification using support vector machines. In *Proceedings of the Sixteenth International Conference on Machine Learning*, 200–209.

Jurafsky, D. 2003. Probabilistic modeling in psycholinguistics: Linguistic comprehension and production. In Bod, R.; Hay, J.; and Jannedy, S., eds., *Probabilistic Linguistics*. MIT Press.

Nigam, K.; McCallum, A.; Thrun, S.; and Mitchell, T. 2000. Text classification from labeled and unlabeled documents using EM. *Machine Learning* 39(2/3):103–134.

Riloff, E., and Jones, R. 1999. Learning dictionaries for information extraction by multi-level bootstrapping. In *Proceedings of the Sixteenth National Conference on Artificial Intelligence*, 474–479.

Thompson, C. A.; Califf, M. E.; and Mooney, R. J. 1999. Active learning for natural language parsing and information extraction. In *Proceedings of the Sixteenth International Conference on Machine Learning*, 406–414.

Thompson, C. A.; Levy, R.; and Manning, C. 2003. A generative model for semantic role labeling. In *Proceedings of the Fourteenth European Conference on Machine Learning*, 397–408.