Dissertation Summary: Learning to Extract Entities from Labeled and Unlabeled Text

Rosie Jones

May 5th, 2005

We describe and evaluate algorithms for learning to extract semantic classes from sentences in text documents, using the minimum of training information. The thesis of this research is that we can *efficiently* automate information extraction, that is, learn from tens of examples of labeled training data instead of requiring thousands, by exploiting redundancy and separability of the features *noun-phrases* and *contexts*. We exploit this redundancy and separability in two ways: (1) in the algorithms for learning semantic classes, and (2) in novel algorithms for active learning, leading to better extractors for a given amount of user labeling effort.

Our goal is to learn to extract noun-phrases of particular semantic types or classes from sentences in text documents. We perform in-depth experiments with identifying the three semantic classes `locations`, `people`, and `organizations` in sentences in text documents. We represent instances of semantic classes with an ordered pair, consisting of a noun-phrase and the local syntactic context. We identify these noun-phrases and contexts in documents using Sundance, a shallow parser. We take a small set of words which the user believes may be examples of the target class (which we will call *seeds*), and use them to perform *weak labeling* (providing noisy/partial labels to a relatively small number of the examples) on a collection of documents, by identifying any noun-phrase containing those words as a positive example. We weakly label all other examples as negative examples, and perform *semi-supervised learning* with this weakly labeled data.

We describe the algorithms *metabootstrapping*, *coEM* and *EM* and show how they are applicable to semi-supervised learning of this information extraction task. We find that *cotraining* does not work well using our representation on our task. We break the algorithms down into those that employ a separation of the feature sets, and those that combine the feature sets. We show that performance is affected by the set of seeds chosen, with more frequently occurring seeds leading to better extraction performance. However, correcting errors in the weak labeling performed with seeds does not lead to substantial performance improvement. We also show that we should use stopwords as part of the model, for best performance across classes.

Given that our goal is optimizing the trade-off between user training time and algorithm performance, it is important to see how we can improve results with active learning. We describe novel active learning algorithms which are algorithmically coupled with the separable feature sets. In addition, we describe a novel labeling technique, *single feature-set labeling.* In *single feature-set labeling*, the user sees only a partial example, for example the noun-phrase in isolation from the context. We then use the labeled noun-phrase to label all contexts it cooccurs with. This technique provides for economy in labeling, and we show that in some cases it is more effective than standard whole example labeling.

The novel active learning algorithms are *feature-set disagreement*, in which we select examples for labeling when the features disagree strongly on the target label (that is, when the noun-phrase and context disagree on the target label), and *context disagreement*, when we select noun-phrases for single feature-set labeling when their contexts disagree (for example, the noun-phrase occurs with several different contexts, which differ in their confidence of representing a positive example). We also compare against selecting the most frequent examples, which can be important when some classes are represented well by pronouns. Overall we show that judicious selection of examples for labeling can lead to greatly increased accuracy, without greatly increasing the burden on the user. In particular, we show that using the feature set redundancy allows selection of examples for labeling which are much more effective than examples chosen randomly, or without use of feature set redundancy. In addition, these results show that active learning can compensate for a bad choice of initial seeds and that the labeling effort is better spent *during* the active learning process rather than at the beginning.

Finally, we perform a deeper analysis of the results. We measure properties of the noun-phrase context connectivity graph, and show that it exhibits small-world graph structure rather than random graph structure. We analyze how this explains the failure of cotraining on our tasks, and how pronouns and certain very common nouns form the hubs of the connectivity graph. We measure the mutual information between noun-phrases and contexts in each class, in order to test our conditional independence hypothesis. We also perform Spearman rank correlation tests over multiple experiments, finding correlations between algorithm breakeven point and features including the number of contexts labeled by initial seeds and the percent of examples labeled positive in active learning. These features correlated with learning performance can help us pinpoint the important properties of active learning and bootstrapping algorithms for information extraction. Comparing these across classes also highlights the different desiderata for active learning algorithms for classes with sparse feature sets and extremely small priors.

Thesis Statement

The thesis of this research is that we can *efficiently* automate information extraction, that is, learn from tens of examples of labeled training data instead of requiring thousands, by exploiting redundancy and separability of the two features: (1) *noun-phrases* and (2) *contexts*. We exploit this redundancy and separability in two ways: (1) in the algorithms for learning semantic classes, and (2) in novel algorithms for active learning, leading to better extractors for a given amount of user labeling effort.

Contributions

The contributions of this research are

- In-depth experiments with bootstrapping algorithms across multiple semantic classes.
- Adapted existing semi-supervised learning algorithms for the task of information extraction.
- Novel active learning algorithms that take into account the feature set split into two sets.

- Analysis of the noun-phrase context co-occurrence graph to show that it exhibits small-world and power-law structure.

- Demonstration of the correlation between graph features and algorithm performance.

- Suggestions for seed selection for bootstrapping algorithms informed by the graph structure of the data.

- Suggestions for active learning for bootstrapping algorithms informed by the graph structure of the data.