Active learning selection strategies for information extraction

Content Areas: Information Extraction, Active Learning

1 Introduction

Information Extraction (IE) is the process of identifying a set of pre-defined relevant items in text documents. For example, an IE system might convert free text resumes into a structured form. Numerous machine learning algorithms have been developed that promise to eliminate the need for hand-crafted extraction rules. Instead, users are asked to annotate a set of training documents selected from a large collection of unlabelled documents. From these annotated documents, an IE learning algorithm generalizes a set of rules that can be used to extract items from unseen documents.

It is infeasible for users to annotate large numbers of documents. IE researchers have therefore investigated Active Learning (AL) techniques to automatically identify documents for the user to annotate [Thompson *et al.*, 1999; Scheffer and Wrobel, 2001; Ciravegna *et al.*, 2002].

The essence of AL is a strategy for selecting the next document to be presented to the user for annotation. The selected documents should be those that will maximise the future performance of the learned extraction rules. Document selection algorithms attempt to find regions of the instance space that have not yet been sampled in order to select the most informative example for human annotation.

Several selection strategies have been studied in the more general context of machine learning. For example, confidence-based approaches select for annotation the unlabelled instance of which the learner is least confident. While such techniques are clearly applicable to IE, we focus on novel selection algorithms that exploit the fact that the training data in question is text.

2 Document selection strategies

We begin by introducing several novel AL document selection strategies for IE. Some of the strategies are applicable only in an IE or text classification context. While they are tailored for IE, they are generic in that they do not assume any specific IE algorithm. The first document is always selected randomly, and subsequent documents are selected as follows.

Compare Present for annotation the document that is textually least similar to the documents that have already been annotated. Similarity can be measured in various ways, such as raw term overlap, or using TFIDF weighting.

Dual First learn a set of rules to extract instances of some field x, and then invoke the learning algorithm again to learn rules to extract $\neg x$ (i.e. everything except instances of field x). We select the document with the largest overlap between x and $\neg x$.

Unusual Present for annotation the document with the most unusual proper nouns in it. Often the items we wish to extract are proper nouns such as names. These can be difficult to extract, because they are likely to be words that have not been seen before.

Committee Invoke two different IE learning algorithms (e.g. LP^2 [Ciravegna, 2001] and Rapier [Califf and Mooney, 1999]). The next selected document is the one on which the learned rule sets most disagree.

Bag Invoke the learning algorithm on different partitions of the available training data. As with **Committee**, the document that maximizes disagreement is selected.

Mine Following [Nahm and Mooney, 2000], learn a set of extraction rules, and then mine a set of association rules for predicting which extracted items frequently cooccur. Select for annotation the document whose extracted content most contradicts the association rules.

Each algorithm encodes different heuristics for identifying regions of the instance space that have not yet been sampled.

3 Preliminary results

We have evaluated two of our selection strategies with a subset of Freitag's well-known seminar announcements IE task. We report results for extracting the seminar speaker and location, which are known to be the hardest fields.

Fig. 1 shows the learning curve produced by two of our selection algorithms, as well as a baseline **Random** strategy that selects documents randomly. **Optimal** estimates an upper bound on performance by selecting the document that will result in the largest improvement. The horizontal axis shows the number of documents used for training, and the vertical axis shows the average F-measure over five trials. **Compare** measures text similarity between two documents as the number of words in their intersection divided by the total number of words; the similarity score for a potential training document is the sum of the similarity of that document and each document already selected.

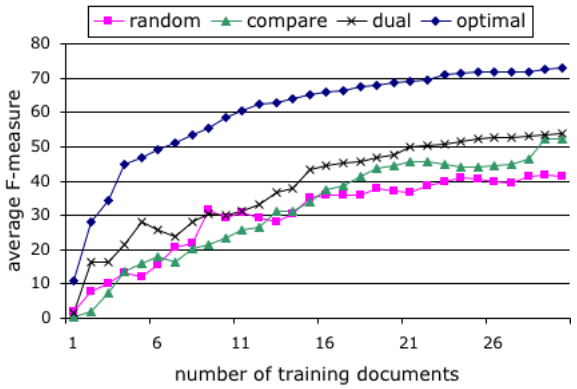

Figure 1: Learning curves for **Dual** and **Compare**.

Our preliminary results indicate that **Compare** shows a slight improvement over **Random**, and a large improvement by **Dual**, particular for very few documents. We anticipate that more sophisticated variants of **Dual** and **Compare** may lead to further improvements.

4 Future work

We are currently evaluating our selection algorithms on a variety of IE tasks. To date, our work is mainly empirical, and our initial goal is to measure the relative strength of the various algorithms. We anticipate that our results will lead us to design additional strategies.

Our longer-term goal is to generalize our ideas beyond IE. As an initial formulation, we can treat each instance Y as comprising two feature sets: $Y = U + V$. U corresponds to the features used by the learning algorithm and, as usual in inductive learning scenarious, we assume that U is sufficient to learn the target concept T. V corresponds to additional features that the AL strategy can use for selection. For some tasks, it may be that V and U are sufficient for learning T, and our analysis reduces to a multi-view problem.

On the other hand, in IE and perhaps other settings, V may include features that compactly indicate the location of U in the instance space U, yet V alone is useless for actually learning T. For example, a typical U encoding for IE will specify the position, part-of-speech, etc. of every term in a document. It is prohibitively expensive to compare the detailed U encodings of all unlabelled documents. On the other hand, **Compare** uses a compact bag-of-words representation V which is efficiently computable and reasonably effective, yet is useless for learning extraction rules.

We are currently elaborating this theoretical treatment to answer questions such as: Are some heuristics provably more effective than others? Can we bound the utility of AL as a function of the efficiency of V compared to U? Are there general properties of learning tasks for which our approach is efffective?

5 Related work

There has been a large amount of work on adaptive information extracton, e.g. [Ciravegna, 2001; Califf and Mooney, 1999; Freitag and Kushmerick, 2000] and many others. These algorithms generally perform well, but all have the potential for further improvement through active learning techniques.

Multi-view learning [Blum and Mitchell, 1998] has received widespread attention. With this approach, predictions based on different logical views of the data can then be used to suggest which examples should next be added to the training set. Our **Dual** algorithm is in the mutli-view family.

There has been some work in the application of active learning to IE, but it often uses learning algorithm specific heuristics to choose the next document for annotation. For example, [Thompson *et al.*, 1999] take a confidence-based approach. They measure the certainty of a rule generated by Rapier based on its coverage of the training data. When choosing documents for annotation, they target rules with low certainty and attempt to find examples to confirm or reject this.

Another learning algorithm-specific approach is described in [Scheffer and Wrobel, 2001] for learning Hidden Markov Models from partially labelled data. They apply Active Learning to this problem by identifying "difficult" unlabelled tokens and asking the user to label them. Difficulty is estimated by the difference between the most likely and second most likely state of the HMM.

Both of these strategies use document selection techniques that are particular to the IE algorithm. A more general multi-view strategy is described by [Muslea *et al.*, 2000] and applied to a wrapper induction task. In their case the different views are created using forward and backward rules.

References

[Blum and Mitchell, 1998] A. Blum and T. Mitchell. Combining labeled and unlabeled data with co-training. In *Proc. Workshop Computational Learning Theory*, 1998.

[Califf and Mooney, 1999] M. Califf and R. Mooney. Relation learning of pattern-match rules for information extraction. In *Proc. 16th Nat. Conf. Artifical Intelligence*, 1999.

[Ciravegna *et al.*, 2002] F. Ciravegna, A. Dingli, D. Petrelli, and Y. Wilks. Timely and non-intrusive active document annotation via adaptive information extraction. In *Proc. Workshop Semantic Authoring Annotation and Knowledge Management (European Conf. Artificial Intelligence)*, 2002.

[Ciravegna, 2001] F. Ciravegna. Adaptive information extraction from text by rule induction and generalisation. In *Proc. 17th Int. Joint Conf. Artificial Intelligence*, 2001.

[Freitag and Kushmerick, 2000] D. Freitag and N. Kushmerick. Boosted wrapper induction. In *Proc. 17th Nat. Conf. Artificial Intelligence*, 2000.

[Muslea *et al.*, 2000] I. Muslea, S. Minton, and C. Knoblock. Selective sampling with reduntant views. In *Proc. 17th Nat. Conf. Artificial Intelligence*, 2000.

[Nahm and Mooney, 2000] U. Nahm and R. Mooney. A mutually beneficial integration of data mining and information extraction. In *Proc. 17th Nat. Conf. Artifical Intelligence*, 2000.

[Scheffer and Wrobel, 2001] T. Scheffer and S. Wrobel. Active learning of partially hidden Markov models. *Lecture Notes in Computer Science*, 2001.

[Thompson *et al.*, 1999] C. Thompson, M. Califf, and R. Mooney. Active learning for natural language processing and information extraction. In *Proc. 16th Int. Conf. Machine Learning*, 1999.