Special semi-supervised techniques for Natural Language Processing tasks

RICHÁRD FARKAS
University of Szeged
Department of Informatics
Árpad tér 2., 6720, Szeged
HUNGARY

GYÖRGY SZARVAS
University of Szeged
Department of Informatics
Árpad tér 2., 6720, Szeged
HUNGARY

JÁNOS CSIRIK
Hungarian Academy of Sciences
Research Group on Artificial Intelligence
Aradi vértanuk tere 1., 6720, Szeged
HUNGARY

Abstract: A labeled natural language corpus is often difficult, expensive or time-consuming to obtain as its construction requires expert human effort. On the other hand, unlabelled texts are available in abundance thanks to the World Wide Web. The importance of utilizing unlabeled data in machine learning systems is growing. Here, we investigate classic semi-supervised approaches and examine the potential advantages of applying special techniques for Natural Language Processing tasks.

Key–Words: Semi-supervised learning, Natural Language Processing, Named Entity

1 Introduction

Several Natural Language Processing (NLP) tasks like parsing and Named Entity Recognition are solved to a satisfactory accuracy for several languages and domains [1] in a supervised environment. However a labeled database is often difficult, expensive or time-consuming to obtain as it requires much expert human effort. On the other hand, unlabeled texts are available in abundance owing to the World Wide Web. Semi-supervised learning aims to build better classifiers by using large amounts of unlabeled data along with labeled data. This is of great interest in many machine learning algorithms both from theoretical and practical points of view. In this paper we attempt to answers to the following two questions:

Is it possible to achieve the same accuracy in NLP tasks with a smaller labeled corpus by utilizing unlabeled texts instead of training on a large labeled corpus? Which semi-supervised techniques are applicable and especially suitable for NLP tasks?

In Section 2 we introduce the main semi-supervised techniques described in the machine learning literature and discuss the potential advantages of applying special semi-supervised techniques in NLP. Experimental results by some of these techniques on the classical NER (identifying the classes organization, person, location and miscellaneous) for English and Hungarian will be presented in Section 3, followed in the last section by discussion and some concluding remarks.

2 Semi-supervised techniques

In this section we shall provide an overview on semi-supervised techniques along with a discussion on their speciality in Natural Language Processing.

2.1 General Semi-supervision in Machine Learning

In a semi-supervised learning environment a model is trained by using unlabeled data, together with labeled data. The goal is to utilize the unlabeled data during the training on labeled ones. We classify these kind of approaches into three categories, namely generative models, bootstrapping methods and low density separation. For a detailed description, see [2].

The first attempts were done by applying generative models (like HMMs)[3]. A generative model directly describes how the labels are probabilistically conditioned on the inputs (tokens). 'Directly' means here that the types of distributions are assumed, and their parameters are estimated from the data. Usually a mixture distribution is assumed and the great amount of unlabeled data helps to identify the mixture components [3]. However, there are several problems associated with using generative models. The most obvious one is that we have to know the types of the distributions, otherwise unlabeled data reduces the accuracy [4].

We call bootstrapping methods (self-training

and co-training) those type of methods where a train training dataset is expanded by automatically labeled (originally) raw data [5]. In self-training a classifier first learns on the labeled dataset and then classifies the unlabeled data. The most reliable examples are afterwards added to the training set and the procedure is repeated. In co-training two or more different classifiers are used for predicting unlabeled data, then they "teach" each other via the most reliable instances from the unlabeled pool. The two classifiers can be from a different algorithm class or they can be the same learning method trained on conditionally independent feature subsets.

The latest approaches of semi-supervised learning are based on the 'separate only on low density regions' principle (low density separation). These approaches also use the evaluation dataset as unlabeled data. Hence here the overall goal is not to build a general model which predicts well on previously unseen instances (inductive learning) but "just" give an as perfect as possible prediction on a specific evaluation set (transductive learning). Transductive Support Vector Machines (TSVM) [6] is the most known such method. In the optimization procedure, it looks for a labeling of the unlabeled data where the margin is maximized on both originally labeled and (currently labeled) unlabeled data. Clearly unlabeled data restrict the boundary to low density regions. Graph-based methods are a newer field of low density separation [7]. Here a graph is built where nodes are labeled and unlabeled inputs and the edges represent their similarity (usually just the nearest neighbors are connected). If two points are in the same cluster there exists a path between them that only goes through high density regions. Thus our aim here is to learn a function (find the clusters) which cuts on low similarity points.

The low density separation methods have a good theoretical foundation, but at the moment they can handle just small datasets in practice. Even programs describing themselves as solution to large-scale problems cannot give results for a task of 20,000 samples with 120 features after running for a week[1]. There are several suggestions on how to scale up these methods, but databases containing hundreds of thousands of examples like in most of the NLP tasks seem feasible only in the future.

[1]Two packages were downloaded and tested: www.kyb.tuebingen.mpg.de/people/fabee/universvm.html and www.learning-from-data.com/te-ming/semil.htm

2.2 Semi-supervision in Natural Language Processing

The special nature of Natural Language Processing problems requires special semi-supervised techniques. The potential of this field has not yet been satisfactorily exploited. Two key points are discussed here:

Complex statistics can be gathered from unlabeled texts owing to the sequential structure of languages. Such statistics can be word and character bi-, trigrams, token or phrase frequencies and models of language in a wider sense (not just the usual $P(w_t|w_{t-1})$ distribution). This kind of information can be incorporated into the feature space for each machine learning process.

Another unique characteristic of NLP applications is that they can utilize the World Wide Web (WWW). The WWW can be viewed as an almost limitless collection of unlabeled data, but it cannot be handled by the classical semi-supervised (or unsupervised) techniques. It is feasible just via search engines (e.g. we cannot iterate through all of the occurrences of a word). There are two interesting problems here: first, appropriate queries must be sent to a search engine; second the response of the engine offers several opportunities (result frequencies, snippets etc.) in addition to simply "reading" the pages found. Although there are several papers that tell us how to use the WWW to solve simple natural language problems like [8], we think that it will be a rapidly emerging area and deeper analysis will be performed over the coming years.

3 Experiments

We studied the effects of self-training, co-training and several heuristics based on the WWW along with the impact of features gathered from unlabeled corpora. The datasets, our tools and the results obtained will now be described in detail.

3.1 Datasets and representation used

The identification and classification of Named Entities (NE) in plain text is of key importance in numerous natural language processing applications. For example, in Information Extraction systems NEs generally carry important information about the text itself, and thus are targets for extraction. We used English and Hungarian Named Entity reference corpora in this study.

Named Entity Recognition (NER) models in English were trained and tested on the CoNLL

2003 corpus [9], that consists of newswire articles provided by Reuters Inc. It has is approximately 200,000 tokens in size and contains texts from diverse domains ranging from sports news to politics and economics. The NE classes *organization*, *person*, *location* and *miscellaneous* are manually tagged in the corpus. We evaluated our methods on the development set of the contest, because the evaluation set differs in its characteristics from the train set. One of the aims of this contest was to discover the usefulness of unlabeled texts, but none of the participating systems made use of them in a sophisticated way. The database contains more than 18 million unlabeled tokens which were used in our experiments.

NER models on Hungarian texts were trained and tested on the SzegedNE corpus [10] which consists of short business news from 38 NewsML topics ranging from acquisitions to stock market changes or the opening of new industrial plants. The annotation of the corpus followed the CoNNL annotation. The size of this corpus is the same as that for the CoNNL corpus (200,000 tokens). Currently we do not have unlabeled texts from the same source as this corpus. Our investigations on other raw texts (from the domain economy) were unsuccessful (see Section 3.3). Hence we followed the transductive approach in Hungarian semi-supervised experiments, that is the evaluation dataset was used as unlabeled text.

We employed a rich feature set which describes the characteristics of each token along with its actual context (a moving window of size four). The same feature sets were used in the experiments on Hungarian and English. Our features fell into the following major categories:

Orthographical features: capitalization, word length, bit information about the word form (whether it contains a digit or not, has uppercase character inside the word, and so on), the most implicative character level bi/trigrams from the train texts for each NE class.

Phrasal information: chunk codes and forecasted class of several preceding words used by the classification approach (we used online evaluation).

Contextual information: sentence position, trigger words (the most frequent and unambiguous tokens in a window around the NEs) from the train text, and so on.

Dictionaries of first names, company types, sport teams, denominators of locations (mountains, city) and so on; we collected 12 English specific lists from the Internet and 4 additional ones for the Hungarian problem.

Frequency information: frequency of the token, the ratio of the token's capitalized and lowercase occurrences, the ratio of capitalized and sentence beginning frequencies of the token.

3.2 The impact of features derived from unlabeled corpora

The most successful sequence labeling method, the Conditional Random Fields (CRF) [11] was used for our investigations (implementation MALLET[2]). In the first experiments we examined the impact of training size in a clear supervised environment along with the features derived from unlabeled corpora.

These kinds of feature are the frequency information and various dictionaries. The former one was gathered from corpora containing several billion of tokens (Gigaword and Szószablya). The Named Entity dictionaries were collected from the Web as well. These lists (for a certain category) can be gathered by automatic methods via search engines and simple frame-matching algorithms [8], but the elementary lists can be downloaded in a collected form and only their filtering and normalization have to be done. The lists used here are downloaded and cleared manually which required less than 1 person day.

The 4-4 curves of Figure 1 represents the results using the entire feature space (continuous), without frequency information (dotted), without dictionaries (dashed) and without either (long-dashed). The following tendency can be observed: the absence of dictionaries causes smaller loss in accuracy when the training set size is growing. The added value of dictionaries is important when only a small labeled database is present but this information can be gained from a great labeled dataset. The employment of frequency information eliminates 19% of the errors in average, the dictionaries 15% and their combined usage 28%.

Overall, Figure 1 has a logarithmic trend in performance as the corpus size increases. These results of the supervised model were used as the baseline to the semi-supervised investigations.

3.3 Results obtained by bootstrapping methods

We investigated self-training and co-training in parallel. In self-training CRF was applied and in co-training we extended the labeled training set of CRF with automatically labeled instances by our

[2]A. McCallum: A Machine Learning for Language Toolkit, url: http://mallet.cs.umass.edu

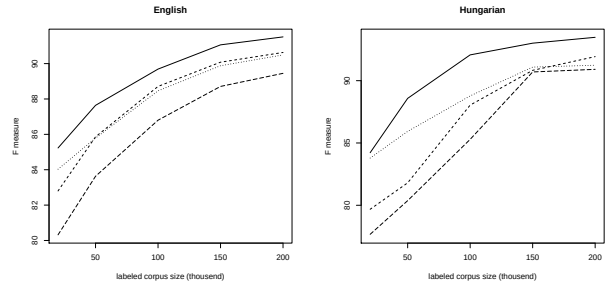

Figure 1: The effect of training corpus size on the supervised learning tasks

boosting and decision tree based NER model [1]. The CRF and the decision tree approaches have different theoretical bases. First, the decision tree forecasts for each token independently (the information about the surrounding words are incorporated into the feature space) while the CRF solves a sequence labeling problem. Second, CRF approximates the distribution conditioned on the joint feature space while the decision tree chooses a feature split at each step in a greedy way. This kind of diversity of the two models makes them a good candidate for co-training.

Figure 2 shows the results obtained by self-training (dotted line) and co-training (continuous line) with different sizes of labeled training data. The baselines (the zero point on the X axis) were the accuracies of the supervised CRF model. The accuracy of co-training could be increased when we do not use every predicted sentences but define a confidence threshold of the decision tree for choosing reliable sentences. Obviously the lower the threshold, the lower the ratio of sentences match the criteria, and thus a larger unlabeled initial database is required. Figure 3 shows two confidence level settings: the continuous line (the same as in Figure 2) represents a confidence threshold of 10^{-3}, while the dotted line represents a threshold of 10^{-10}.

The most important conclusion of these experiments is that the increasing trend remains stable even when we use large unlabeled datasets as well and we could achieve slightly better results by co-training with 100,000 labeled tokens (with an F measure of 91.28%) instead of using the supervised model with 200,000 labeled tokens (91.26%). In order to obtain this accuracy with 100,000 labeled tokens we gathered 23,507 reliable sentences (400,000 tokens) from a raw text of 3 million tokens.

Figures 2 and 3 show that, by using unlabeled text, the results of a supervised model can

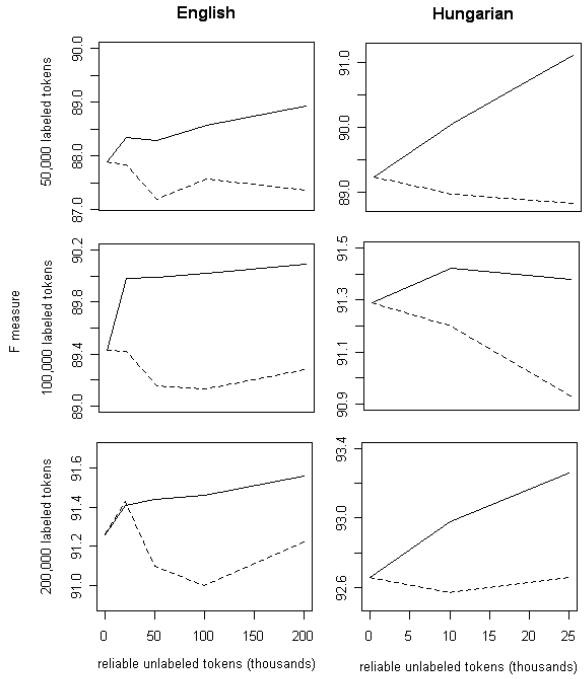

Figure 2: Self-training (dotted) and co-training (continuous) results on the NER tasks

be improved significantly with every size of labeled training data both in English and Hungarian NER tasks. Along with these nice results we must mention some poor ones as well. Co-training was not robust when we evaluated it on the evaluation set of the CoNNL 2003 contest (instead of the development set). In the case of 100,000 labeled examples plus 200,000 raw "reliable" tokens, the supervised model achieved an F measure of 83.58%, while co-training with 10^{-3} confidence threshold gave just 83.21%. The model with a 10^{-10} confidence level improved the accuracy but not by a significant amount (F measure of 83.62%). We investigated bootstrapping methods with raw Hungarian economy texts obtained from another source than our corpus. We found that neither self-training nor co-training could achieve better results than the supervised CRF. Based on these experiments we came to the conclusion that the training set, the evaluation set and the unlabeled dataset as well should share very similar characteristics in a well functioning semi-supervised system.

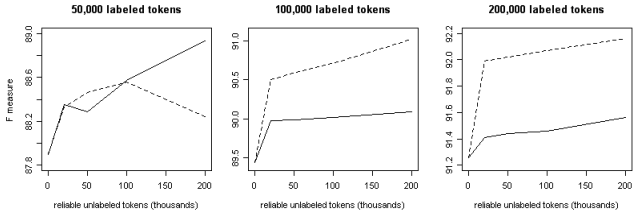

Figure 3: Co-training results with a confidence thresholds of 10^{-3} (continuous line) and 10^{-10} (dotted)

3.4 Utilizing the World Wide Web as an input for NLP tasks

There are several ways of gathering information for Natural Language Processing tasks from the World Wide Web (as an external knowledge source). In connection to NER, the published systems collect lists of Named Entities belonging to pre-specified classes [8] from the WWW. We introduced [12] three different approaches to fine-tuning the results of an NER system using the Google API and Wikipedia:

A significant part of system errors in NER taggers is caused by the erroneous identification of the beginning (or end) of a longer phrase. Token-level classifiers (like the one we applied here) are especially prone to this as they classify each token of a phrase separately. We considered a tagged entity as a candidate long-phrase NE if it was followed or preceded by a non-tagged uppercase word, or one/two stop words and an uppercase word. Our hypothesis for this first heuristic is that if the boundaries hence been marked correctly and the surrounding words are not part of the entity, then the number of web-search results for the longer query should be significantly lower (the NE is followed by the particular word in just certain contexts). But in the case of a dislocated phrase boundary, the number of search results for the extended form must be comparable to the results for the shorter phrase. This means that every time we found a tagged phrase that received more than 0.1% of web query hits in an extended form, we extended the phrase with its neighbouring word(s).

Our hypothesis for the second heuristics was that the most frequent role of a named entity can be statistically useful information. Thus we did the following: if the system was unable to decide the class label of a phrase (it could not find evidence in the context of the certain phrase) then we mined the most frequent usage of the corresponding NE using the WWW and took that as a prediction. Our approach works by invoking several special Google queries in order to find such noun phrases following or preceding the pattern that is a category name for a particular class (e.g. *NP such as NE*, *NP including NE*, *NE and other NP*). We used the lists of unambiguous NEs collected from the training data to acquire common NE category names from the WWW. Then using these category lists as a disambiguator (we assigned the class sharing the most words in common with those extracted for the given NE) when the NER system was unable to give a reliable prediction was beneficial to overall system performance. We used the simple way of interpreting the uncertainty of a decision, we measured the level of disagreement among individual models (committee-based learning).

In the majority of cases, consecutive Named Entities either follow each other with a separating punctuation mark (enumerations) or belong to different classes. In the first case, a non-labeled token separates the two phrases, while in the second case the different class labels identify the boundaries. Rarely do two or more NEs of the same type appear consecutively in a sentence. In such cases the phrasal boundaries must be marked with a tag ('B-' instead of the common 'I-' prefix). Such cases are rather problematic for a statistical model. We exploited the encyclopedic knowledge of Wikipedia to enable our system to distinguish between long phrases and consecutive entities. We queried the Wikipedia site for all entities that had two or more tokens. If we found an article sharing the same title as the whole query, or the majority of the occurrences of the phrase in the Google snippets occurred without punctuation marks inside, we treated the query phrase as a single entity. If a punctuation mark was inside the phrase in the majority of the cases, we separated the phrase at the position of the punctuation mark. This method allowed us to separate phrases like 'Golan Heights | Israel'.

The empirical results of [12] on the techniques introduced above confirm the usefulness of the external information gathered from the Web. We came up against two problems when adapting our approach to the Hungarian task. First, we could not use each query of the most frequent role heuristics translated from English as the substantive verb in the third person singular is not present in Hungarian. We had to look for new query expressions and found one that was helpful: *NE egyike NP* (*NE is one of NP*). Second,

the Hungarian web (we used the *site:.hu* expression in our queries) seems to be too small to get really useful responses. On average about 70% of our queries got zero results from the Google API and the size of Hungarian Wikipedia is only about 3.5% of the English one. This fact suggests that the above mentioned WWW-based methods probably cannot provide satisfactory results for less common languages like Hungarian.

4 Discussion and Conclusions

The aim of this paper was to reveal to the utilization potential of unlabeled texts in Natural Language Processing tasks. We found experimental evidence for the usefulness of raw texts on English and Hungarian NER tasks. Due to learner diversity of co-training, unlabeled data improved supervised models with every size of labeled dataset. What is more we achieved the same level of accuracy with 100,000 labeled examples and raw texts as that from using supervised learning with 200,000 labeled tokens. However we discussed that employing standard semi-supervised techniques for NLP tasks is still unfeasible (low density separation based approaches) or requires a very careful unlabeled data selection (very similar domain/structure), hence we argue to discover special semi-supervised techniques for NLP.

We investigated the effect of features derived from unlabeled corpora. These features brought an average error reduction of 28% and can be obtained in couple of hours of human work. We think that the use of the World Wide Web as an unlimited unlabeled corpus for semi-supervised learning will be done in a more sophisticated way in the near future. Our heuristics are based on the assumption that, even though the World Wide Web contains much useless and faulty information, for our simple features the correct usage of language dominates over misspellings and other sorts of noise. Our experiments confirmed this hypothesis: we described three heuristics based on the Google API response (hit counts and snippets) and on the Wikipedia encyclopedia. Using them as a supplementary heuristics a state-of-the-art multi-lingual NER system was further improved.

In the future we would like to develop solutions which employ the WWW more effectively and find novel semi-supervised procedures that are especially suitable for other natural language processing tasks.

References:

[1] Gy. Szarvas, R. Farkas, and A. Kocsor. A multilingual named entity recognition system using boosting and c4.5 decision tree learning algorithms. *DS2006, LNAI*, 4265:267–278, 2006.

[2] Xiaojin Zhu. Semi-supervised learning literature survey. Technical Report 1530, Computer Sciences, University of Wisconsin-Madison, 2005.

[3] Ch. D. Manning and H. Schütze. *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge, 1999.

[4] Fabio Gagliardi Cozman, Ira Cohen, and Marcelo Cesar Cirelo. Semi-supervised learning of mixture models. In *ICML*, pages 99–106, 2003.

[5] Tom Mitchell. The role of unlabeled data in supervised learning. In *Proceedings of the Sixth International Colloquium on Cognitive Science*, 1999. (invited paper).

[6] Vladimir N. Vapnik. *Statistical Learning Theory*. Wiley, New York, NY, USA, September 1998.

[7] O. Chapelle and A. Zien. Semi-supervised classification by low density separation. *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics*, pages 57–64, 2005.

[8] O. Etzioni, M. J. Cafarella, D. Downey, A. Popescu, T. Shaked, S. Soderland, D. S. Weld, and A. Yates. Unsupervised named-entity extraction from the web: An experimental study. *Artificial Intelligence*, 165(1):91–134, 2005.

[9] E. F. Tjong, K. Sang, and F. De Meulder. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proc. of CoNLL-03, 7th Conf. on Computational Natural Language Learning*, pages 142–147, Edmonton, Canada, 2003.

[10] Gy. Szarvas, R. Farkas, L. Felföldi, A. Kocsor, and J. Csirik. A highly accurate named entity corpus for Hungarian. In *Proc. of LREC-06, 5th Int. Conf. on Language Resources and Evaluation*, Genoa, Italy, 2006.

[11] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proc. of ICML-01, 18th Int. Conf. on Machine Learning*, pages 282–289, Williamstown, USA, 2001.

[12] R. Farkas, Gy. Szarvas, and R. Ormándi. Improving a state-of-the-art named entity recognition system using the world wide web. *ICDM2007, LNCS*, 4597:163–172, 2007.