Language Independent Named Entity Classification by modified Transformation-based Learning and by Decision Tree Induction

William J. Black
Department of Computation, UMIST
P.O. Box 88, Sackville Street
Manchester M60 1QD, UK
wjb@co.umist.ac.uk

Argyrios Vasilakopoulos
Department of Computation, UMIST
P.O. Box 88, Sackville Street
Manchester M60 1QD, UK
A.Vassilakopoulos@postgrad.umist.ac.uk

Abstract

We describe our last results at the CoNLL2002 shared task of Named Entity Recognition and Classification using two approaches that we first applied to other NLL problems. We have been developing our own modified TBL learner initially to tackle the Part-of-Speech tagging problem, for integration in a hybrid NLL and rule-based system for information extraction (Ciravegna et al., 1999). After encouraging results in applying decision tree induction to the CoNLL2001 task of chunking, Jones (2002), where we attained an overall F-measure of 92.90 at this task, we have applied the same set-up to the NER task.

1 Introduction

Named Entity Classification (NEC) is the process of identifying and classifying names in text and is a crucial task for several natural language processing areas such as information retrieval, information extraction, machine translation and language understanding. From 1995 when the NE task was first introduced as part of the Message Understanding Conference (MUC-6) most systems that have attempted this task are based in lists of common names in order to provide some clues. These lists are, in most cases, very large and can provide an efficient way of dealing with NEC but it is still a nave method for recognizing names. About the size of the lists, Krupke and Hausman (1998) found that the good quality of a name list is quite more important than its total size, while Mikheev et al. (1999) concluded that small but well elaborated lists can be as effective as the larger ones. Except from name lists many of the NEC systems also use a number of other NLP tools such as hand-crafted rules, morphological disambiguators, chunkers and parsers taking advantage of what McDonald (1996) defines as internal and external evidence in the NEC. While many systems tend to recognize the names in text very efficiently - Zhou and Su (2002) report F-measures of 96.8% and 94.2% on the MUC6 and MUC7 datasets - most of them are designed for specific domains and specific languages.

Our aim has been to build a language independent system for dealing with the NEC task. We have attempted two different approaches using only the data provided for the CoNLL2002 shared task of NEC. The first approach is based on the Transformation- Based Learning method, a very efficient method which has been successfully attempted for other NLP tasks like PP-attachment disambiguation (Brill and Resnik, 1994), part-of-speech tagging (Brill, 1995), text chunking (Ramshaw and Marcus, 1995), dialog act tagging (Samuel et al., 1988) ellipsis resolution (Hardt, 1998) and spelling correction (Magnu and Brill, 1997). Our second approach is a simple decision tree induction scheme. In the next two sections we describe our approaches in more detail and in section 3 we present our results so far.

2 Modified TBL Approach

The learner we used in our first experiment differs from Brill's TBL Learner (Brill, 1995) in that it produces a set of transformation rules in a single pass without basing each learning cycle on a new initial state. We were motivated to do this after noticing how little the revision stage contributes to the final precision of the tagger, having found that improved unknown-word guessing contributes more than is lost by abandoning the multi-pass patch acquisition step.

The approach followed here consists of two main stages for both the learner and the tagger: (i) In the first stage we try just to *recognize* all

the named entities (NE's) in the training or test data by taking into account only the orthography feature of the NE's.
(ii) In the second stage we *classify* the NE's found during the stage (i) by using a corpus derived lexicon and contextual rules.

2.1 Learning

The learning process consists of the induction of two sets of rules and a lexicon. The first set of rules is induced in the first stage and results in a binary classification between the 'O' of the training corpus to 'NE' which generalizes over the remaining tags. This stage creates an initially annotated version of the text. An example of a rule at this phase of analysis is:

> *Change tag O to NE of the current word if the preceding and following words are both Capitalized and the current word is "de"*

which would result in the correction of *Rio/NE de/O Janeiro/NE* to *Rio/NE de/NE Janeiro/NE*. In the second stage we firstly create the lexicon from the training corpus in the way that Brill does in (Brill, 1995). The lexicon then is applied on the initially annotated text and the result is the *initial state* for the TBL algorithm. *The initial state* is afterwards compared to the original training corpus, according to a set of user defined templates and a second set of transformation rules is induced. The difference with Brill's original TBL is that in our case we keep all the rules which satisfy the accuracy and score thresholds, instead of keeping the best one and iterate the process until no more rule is found. We finally rank these rules. At this point it is essential to note that only the word sequences tagged as 'NE's at the first stage, are subject to stage (ii).

2.2 Tagger

For tagging an unknown corpus, we firstly create the initial state of the corpus as described in the previous section, and we then apply all the ranked rules sequentially.

3 Decision Tree Induction

As with our modified TBL approach, we use no data outside of the training corpus for the decision tree experiment. We used an of-the-shelf system (Weka's J4.8 variant of C4.5 - Weka 3 (2001)) for this experiment. The training data is converted into a 49-attribute table, covering 12 attributes of the current token, its two predecessors and its successor. There are 46 nominal and 3 numeric features. These features are:

- The token itself if it is one of the 150 most frequent tokens, otherwise 'miskTok'.
- The orthography. Each token can only belong in one of the next 17 categories which are distinguished via regular expressions: { null, lowercase, capitalized, caphyphenated, lowerhyphenated, uppercase, multicap, upperdotted, initial, initialdot, punct, doublequote, apostrophe, number, numberrange, bracket, other } (Hopefully most of these are self-evident in meaning)
- 'True' if the token is a frequent word 'false' otherwise.
- 'True' if the suffix of the token is a frequent one 'false' otherwise.
- The most frequent category in the training data, or if not found in the training data 'O' for lowercase tokens and 'I-PER' for all the others.
- The total number of occurrences of the token in the training data.
- Six more features indicating 'True' or 'False' if the token appears as 'B-', 'I-', 'PER', 'LOC', 'ORG' and 'MISC' somewhere in the training data.

The decision tree induced from the training data by using these attributes is then used in order to predict the NE class of the unknown words of the test corpus. Finally, a filter is applied to the result, aiming at removing any discrepancies which refer to the patterns of the NE sequences. So, this filter corrects the following mistakes: a) it changes the following label sequence from "<B-X> <B-X> ..." to "<B-X> <I-X>..."and b) it changes the following label sequence from "<B-X> or <I-X> <I-Y> <I-X>..." to "<B-X> or <I-X> <I-X> <I-X>..." if <I-Y> is the classification label of a functional word. In the above, the X and Y variables can take one of the following values: LOC, PER, ORG, MISC.

4 Results

We have tested only our modified TBL approach with all the data available from the CoNLL2002. Especially, for the Dutch data we used only the NE tags of the tokens of the training corpus, as we did for the Spanish data. In the case of the decision tree induction, having spent most of our time on the data conversion and preparation, we have only conducted a run only for the Spanish data by using only the first 100,000 tokens of the training corpus. The overall results are as shown in the Tables 1, 2, 3 and 4.

Spanish train	Precision	Recall	$F_{\beta=1}$
LOC	87.47%	69.99%	77.76
MISC	80.39%	55.41%	65.60
ORG	74.51%	84.83%	79.34
PER	94.63%	93.18%	93.90
Overall	82.65%	79.02%	80.79

Table 1: NER task results from the modified TBL approach on the Spanish training data.

Dutch train	Precision	Recall	$F_{\beta=1}$
LOC	94.10%	84.99%	89.31
MISC	82.57%	72.04%	76.95
ORG	86.08%	34.94%	49.70
PER	93.98%	90.75%	92.34
Overall	90.44%	75.59%	82.35

Table 2: NER task results from the modified TBL approach on the Dutch training data.

5 Conclusion

This paper has presented two different approaches for solving the named entity classification task using supervised learning. Our target has been to use the least resources for creating our rules, for the modified Transformation-Based approach, and inducing our decision tree for the Decision Tree Induction approach. So far, it seems that the modified TBL approach gives better results on Spanish data than our second approach. After developing our modified TBL learner using the Spanish data and trained them using both the Spanish and the Dutch training corpora we observed the fact that our approach works better with the Spanish test texts. Our both systems seem to perform better on 'persons' regarding the recall, while the precision values look better for the 'locations' and the 'organizations'. Our best result is an F-measure value of 68.21. This is not high enough but the result is encouraging if we take into consideration that both our approaches are very simple and they do not make use of any extended resource except from the information contained in the training text.

Spanish train	Precision	Recall	$F_{\beta=1}$
LOC	74.61%	76.82%	75.70
MISC	78.60%	56.96%	66.05
ORG	86.81%	70.39%	77.74
PER	80.57%	77.90%	79.21
Overall	81.03%	71.52%	75.98

Spanish dev.	Precision	Recall	$F_{\beta=1}$
LOC	58.48%	71.54%	64.35
MISC	52.95%	31.79%	39.73
ORG	74.20%	49.09%	59.09
PER	60.06%	76.11%	67.14
Overall	63.18%	58.21%	60.60

Spanish test	Precision	Recall	$F_{\beta=1}$
LOC	63.04%	60.04%	61.50
MISC	48.85%	31.36%	38.20
ORG	73.15%	58.04%	64.72
PER	60.11%	79.93%	68.62
overall	64.31%	59.87%	62.01

Table 3: NER task results from the Decision Tree Induction on Spanish data.

References

Weka 3. 2001. Machine learning software in java 2001. http://www.cs.waikato.ac.nz/ml/weka.

E. Brill and P. Resnik. 1994. A transformation-based approach to prepositional phrase attachment. In *Proceedings of COLING'94*. Kyoto, Japan.

E. Brill. 1995. Error-driven learning and natural language processing: A case study in part of speech tagging. *Computational Linguistics*, December 1995.

F. Ciravegna, A. Lavelli, M. Mana, L. Gilardoni, S. Mazza, M. Ferraro, J. Matiasek, W. J. Black, F. Rinaldi, and D. Mowatt. 1999. Facile: Classifying texts integrating pattern matching and information extraction. In *Proceedings of IJCAI99*. Stockholm, Sweden.

D. Hardt. 1998. Improving ellipsis resolution with transformation-based learning. AAAI Fall Symposium 1998.

D. Jones. 2002. *Machine Learning for Natural Language Analysis.*

G Krupke and K. Hausman. 1998. Isoquest inc: description of the netowl(tm) extractor system as used for muc-7. In *Message Understanding Conference Proceedings: MUC 7.*

L. Magnu and E. Brill. 1997. Automatic rule acquisition for spelling correction. In *Proceedings of The Fourteenth International Conference on Machine Learning ICML'97.* Morgan Kaufmann.

D McDonald. 1996. Internal and external evidence in the identification and semantic categorization of proper names. In *Corpus Processing for Lexical Acquisition*, pages 21–39. MIT Press, Cambridge, MA. ch. 2.

A. Mikheev, M. Moens, and C Grover. 1999. Named entity recognition without gazetteers. In *Proceedings of the ninth Conference of the European Chapter of the Association for Computational Linguistics*, pages 1–8.

L. A. Ramshaw and M. P. Marcus. 1995. Text chunking using transformation-based learning. In *Proceedings of the ACL Third Workshop on Very Large Corpora, June 1995.*

K. Samuel, S. Carberry, and K. Vijay-Shanker. 1988. Dialog act tagging with transformation-based learning. In *Proceedings of COLING/ACL'98.*

G. Zhou and J. Su. 2002. Named entity recognition using an hmm-based chunk tagger. http://citeseer.nj.nec.com/zhou02named.html.

Spanish dev.	Precision	Recall	$F_{\beta=1}$
LOC	67.54%	64.42%	65.94
MISC	63.16%	38.22%	47.62
ORG	73.29%	60.76%	66.44
PER	70.65%	90.10%	79.20
Overall	70.34%	66.20%	68.21

Spanish test	Precision	Recall	$F_{\beta=1}$
LOC	76.78%	54.22%	63.56
MISC	61.78%	30.13%	40.51
ORG	68.40%	71.57%	69.95
PER	66.74%	92.48%	77.53
overall	68.78%	66.24%	67.49

Dutch devel.	Precision	Recall	$F_{\beta=1}$
LOC	76.41%	40.26%	52.73
MISC	71.59%	39.58%	50.98
ORG	84.51%	17.98%	29.65
PER	50.13%	87.81%	63.82
Overall	58.88%	48.35%	53.10

Dutch test	Precision	Recall	$F_{\beta=1}$
LOC	76.02%	54.51%	63.49
MISC	70.61%	37.19%	48.72
ORG	76.79%	18.09%	29.28
PER	55.43%	87.60%	67.90
Overall	62.12%	51.69%	56.43

Table 4: NER task results from the modified TBL approach on Spanish and Dutch data sets.