Updating a Name Tagger Using Contemporary Unlabeled Data

Cristina Mota
L2F (INESC-ID) & IST & NYU
Rua Alves Redol 9
1000-029 Lisboa Portugal
cmota@ist.utl.pt

Ralph Grishman
New York University
Computer Science Department
New York NY 10003 USA
grishman@cs.nyu.edu

Abstract

For many NLP tasks, including named entity tagging, semi-supervised learning has been proposed as a reasonable alternative to methods that require annotating large amounts of training data. In this paper, we address the problem of analyzing new data given a semi-supervised NE tagger trained on data from an earlier time period. We will show that updating the unlabeled data is sufficient to maintain quality over time, and outperforms updating the labeled data. Furthermore, we will also show that augmenting the unlabeled data with older data in most cases does not result in better performance than simply using a smaller amount of current unlabeled data.

1 Introduction

Brill (2003) observed large gains in performance for different NLP tasks solely by increasing the size of unlabeled data, but stressed that for other NLP tasks, such as named entity recognition (NER), we still need to focus on developing tools that help to increase the size of annotated data.

This problem is particularly crucial when processing languages, such as Portuguese, for which the labeled data is scarce. For instance, in the first NER evaluation for Portuguese, HAREM (Santos and Cardoso, 2007), only two out of the nine participants presented systems based on machine learning, and they both argued they could have achieved significantly better results if they had larger training sets.

Semi-supervised methods are commonly chosen as an alternative to overcome the lack of annotated resources, because they present a good trade-off between amount of labeled data needed and performance achieved. Co-training is one of those methods, and has been extensively studied in NLP (Nigam and Ghani, 2000; Pierce and Cardie, 2001; Ng and Cardie, 2003; Mota and Grishman, 2008). In particular, we showed that the performance of a name tagger based on co-training decays as the time gap between training data (seeds and unlabeled data) and test data increases (Mota and Grishman, 2008). Compared to the original classifier of Collins and Singer (1999) that uses seven seeds, we used substantially larger seed sets (more than 1000), which raises the question of which of the parameters (seeds or unlabeled data) are causing the performance deterioration.

In the present study, we investigated two main questions, from the point of view of a developer who wants to analyze a new data set, given an NE tagger trained with older data. First, we studied whether it was better to update the seeds or the unlabeled data; then, we analyzed whether using a smaller amount of current unlabeled data could be better than increasing the amount of unlabeled data drawn from older sources. The experiments show that using contemporary unlabeled data is the best choice, outperforming most experiments with larger amounts of older unlabeled data and all experiments with contemporary seeds.

2 Contemporary labeled data in NLP

The speech community has been defending for some time now the idea of having similar temporal data for training and testing automatic speech recognition systems for broadcast news. Most works focus on improving out-of-vocabulary (OOV) rates, to which new names contribute significantly. For instance, Palmer and Ostendorf (2005) aiming at reducing the error rate due to OOV names propose to generate offline name lists from diverse sources, including temporally relevant news texts; Federico and Bertoldi (2004), and Martins et al. (2006) propose to daily adapt the statistical language model of a broadcast

Proceedings of the ACL-IJCNLP 2009 Conference Short Papers, pages 353–356,
Suntec, Singapore, 4 August 2009. ©2009 ACL and AFNLP

news transcription system, exploiting contemporary newswire texts available on the web; Auzanne et al. (2000) proposed a time-adaptive language model, studying its impact over a period of five months on the reduction of OOV rate, word error rate and retrieval accuracy on a spoken document retrieval system.

Concerning variations over longer periods of time, we observed that the performance of a semi-supervised name tagger decays over a period of eight years, which seems to be directly related with the fact that the texts used to train and test the tagger also show a tendency to become less similar over time (Mota and Grishman, 2008); Batista et al. (2008) also observed a decaying tendency in the performance of a system for recovering capitalization over a period of six years, proposing to retrain a MaxEnt model using additional contemporary written texts.

3 Name tagger overview

We assessed the name tagger described in Mota and Grishman (2008) to recognize names of people, organizations and locations. The tagger is based on the co-training NE classifier proposed by Collins and Singer (1999), and is comprised of several components organized sequentially (cf. Figure 1).

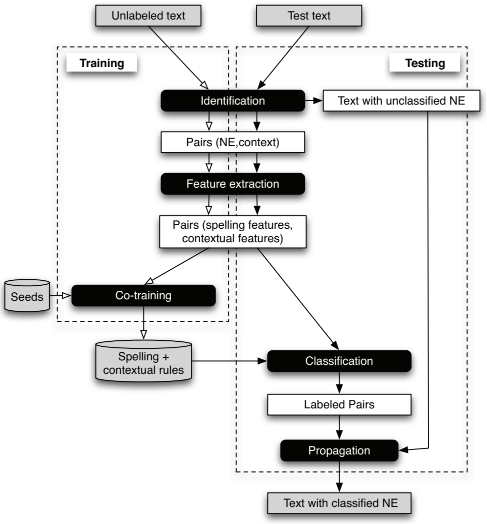

Figure 1: NE tagger architecture

4 Data sets

CETEMPúblico (Rocha and Santos, 2000) is a Portuguese journalistic corpus with 180 million words that spans eight years of news, from 1991 to 1998. The minimum size of epoch (time span of data set) available for analysis is a six-month period, corresponding either to the first half of the year or the second.

The data sets were created using the first 8256 extracts[1] within each six-month period of the politics section of the corpus: the first 192 are used to collect seeds, the next 208 extracts are used as test sets and the remaining 7856 are used to collect the unlabeled examples. The seeds correspond to the first 1150 names occurring in those extracts. From the list of unlabeled examples obtained after the NE identification stage, only the first 41226 examples of each epoch were used to bootstrap in the classification stage.

5 Experiments

We denote by S, U and T, respectively, the seed, unlabeled and test texts, and by (S_i, U_j, T_k) a training-test configuration, where $91a \leq i, j, k \leq 98b$, i.e., epochs i, j and k vary between the first half of 1991 (91a) and the second half of 1998 (98b). For instance, the training-test configuration ($S_{i=91a...98b}$, $U_{i=91a...98b}$, $T_{j=98b}$) represents the training-test configuration where the test set was drawn from epoch 98b, and the tagger was trained in turn with seeds and unlabeled data drawn from the same epoch i that varied from 91a to 98b.

5.1 Do we need contemporary labeled data?

In order to understand whether it is better to label examples falling within the epoch of the test set or to keep using old labeled data while bootstrapping with contemporary unlabeled data, we fixed the test set to be within the last epoch of the interval (98b), and performed backward experiments, i.e., we varied the epoch of either the seeds or the unlabeled data backwards. The choice of fixing the test within the last epoch of the interval is the one that most approximates a real situation where one has a tagger trained with old data and wants to process a more recent text.

Figure 2 shows the results for both experiments, where ($S_{j=98b}$, $U_{i=91a...98b}$, $T_{j=98b}$) represents the experiment where the test was within the same epoch as the seeds and the unlabeled data were drawn from a single, variable, epoch in turn, and ($S_{i=91a...98b}$, $U_{j=98b}$, $T_{j=98b}$) represents the experiment where the test was within the epoch of the

[1] Extracts are typically two paragraphs.

unlabeled data and the seeds were drawn in turn from each of the epochs; the graphic also shows the baseline backward training (varying the epoch of both the seeds and the unlabeled data together).

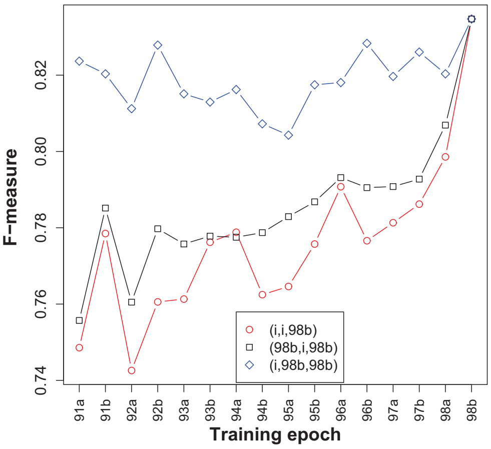

Figure 2: F-measure over time for test set $98b$ with configurations: $(S_{i=91a...98b}, U_{i=91a...98b}, T_{j=98b})$, $(S_{j=98b}, U_{i=91a...98b}, T_{j=98b})$, and $(S_{i=91a...98b}, U_{j=98b}, T_{j=98b})$

As can be seen, there is a small gain in performance by using seeds within the epoch of the test set, but the decay is still observable as we increase the time gap between the unlabeled data and the test set. On the contrary, if we use unlabeled data within the epoch of the test set, we hardly see a degradation trend as the time gap between the epochs of seeds and test set is increased.

An examination of the results shows that, for instance, *Sendero Luminoso* received the correct classification of organization when the tagger is trained with unlabeled data drawn from the same epoch, but is incorretly classified as person when trained with data that is not contemporary with the test set. Even though that name is not a seed in any of the cases, it occurs twice in good contexts for organization in unlabeled data contemporary with the test set (*líder do Sendero Luminoso*/*leader of the Shining Path* and *acções do Sendero Luminoso*/actions of the Shining Path), while it does not occur in the unlabeled data that is not contemporary. Given that both the name spelling and the context in the test set, *o messianismo do peruano Sendero Luminoso*/*the messianism of the Peruvian Shining Path*, are insufficient to assign a correct label, the occurrence of the name in the contemporary unlabeled data contributes to its correct classification in the test set.

5.2 Is more older unlabeled data better?

The second question we addressed was whether having more older unlabeled data could result in better performance than less data but within the epoch of the test set. In this case, we conducted two backward experiments, augmenting the unlabeled data backwards with older data than the test set (98b), starting in the previous epoch (98a): in the first experiment, the seeds were within the same epoch as the test set, and in the second experiment the seeds were within the same epoch as the unlabeled set being added. This corresponds to configurations $(S_{j=98b}, U'_{i=91a...98a}, T_{j=98b})$ and $(S_{i=91a...98a}, U'_{i=91a...98a}, T_{j=98b})$, respectively, where $U'_i = \bigcup_{k=i}^{98a} U_k$.

In Figure 3, we show the result of these configurations together with the result of the backward experiment corresponding to configuration $(S_{i=91a...98b}, U_{j=98b}, T_{j=98b})$, also represented in Figure 2. We note that, in the case of the former experiments, the size of the unlabeled examples is increasing in the direction 98a to 91a.

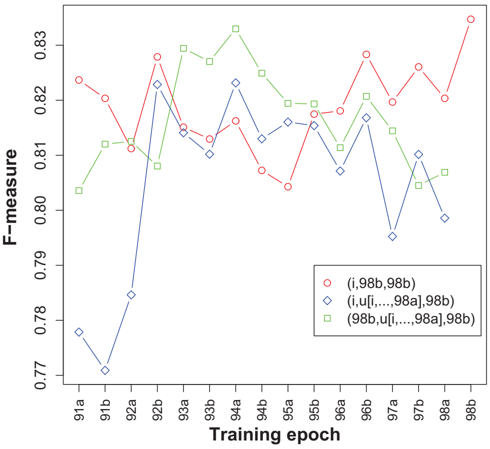

Figure 3: F-measure for test set 98b with configurations $(S_{i=91a...98b}, U_{j=98b}, T_{j=98b})$, $(S_{j=98b}, U'_{i=91a...98a}, T_{j=98b})$ and $(S_{i=91a...98a}, U'_{i=91a...98a}, T_{j=98b})$, where $U'_i = \bigcup_{k=i}^{98a} U_k$

As can be observed, increasing the size of the unlabeled data does not necessarily result in better performance: for both choices of seeds, performance sometimes improves, sometimes worsens, as the unlabeled data grows (following the curves from right to left).

Furthermore, the tagger trained with more unlabeled data in most cases did not outperform the tagger trained with less unlabeled data selected from the epoch of the test set.

6 Discussion and future directions

We conducted experiments varying the epoch of seeds and unlabeled data of a named entity tagger based on co-training. We observed that the performance decay resulting from increasing the time gap between training data (seeds and unlabeled examples) and the test set can be slightly attenuated by using the seeds contemporary with the test set. The gain is larger if one uses older seeds and contemporary unlabeled data, a strategy that, in most of the experiments, results in better performance than using increasing sizes of older unlabeled data.

These results suggest that we may not need to label new data nor train our tagger with increasing sizes of data, as long as we are able to train it with unlabeled data time compatible with the test set.

In the future, one issue that needs clarification is why bootstraping from contemporary labeled data had so little influence on the performance of co-training, and if other semi-supervised approches are also sensitive to this question.

Acknowledgment

The first author's research work was funded by Fundação para a Ciência e a Tecnologia through a doctoral scholarship (ref.: SFRH/BD/3237/2000).

References

Cédric Auzanne, John S. Garofolo, Jonathan G. Fiscus, and William M. Fisher. 2000. Automatic language model adaptation for spoken document retrieval. In *Proceedings of RIAO 2000 Conference on Content-Based Multimedia Information Access*.

Fernando Batista, Nuno Mamede, and Isabel Trancoso. 2008. Language dynamics and capitalization using maximum entropy. In *Proceedings of ACL-08: HLT, Short Papers*, pages 1–4, Columbus, Ohio, June. Association for Computational Linguistics.

Eric Brill. 2003. Processing natural language without natural language processing. In *CICLing*, pages 360–369.

Michael Collins and Yoram Singer. 1999. Unsupervised models for named entity classification. In *Proceedings of the Joint SIGDAT Conference on EMNLP*.

Marcello Federico and Nicola Bertoldi. 2004. Broadcast news lm adaptation over time. *Computer Speech & Language*, 18(4):417–435.

Ciro Martins, António Teixeira, and João Neto. 2006. Dynamic vocabulary adaptation for a daily and real-time broadcast news transcription system. In *IEEE/ACL Workshop on Spoken Language Technology*, Aruba.

Cristina Mota and Ralph Grishman. 2008. Is this NE tagger getting old? In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, may.

Vincente Ng and Claire Cardie. 2003. Weakly supervised natural language learning without redundant views. In *NAACL'03: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 94–101, Morristown, NJ, USA. ACL.

Kamal Nigam and Rayid Ghani. 2000. Analyzing the effectiveness and applicability of co-training. In *Proceedings of CIKM*, pages 86–93.

David D. Palmer and Mari Ostendorf. 2005. Improving out-of-vocabulary name resolution. *Computer Speech & Language*, 19(1):107–128.

David Pierce and Claire Cardie. 2001. Limitations of co-training for natural language learning from large datasets. In *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing (EMNLP-2001)*.

Paulo Rocha and Diana Santos. 2000. Cetempúblico: Um corpus de grandes dimensões de linguagem jornalística portuguesa. In Maria das Graças Volpe Nunes, editor, *Actas do V Encontro para o processamento computacional da língua portuguesa escrita e falada PROPOR 2000*, pages 131–140, Atibaia, São Paulo, Brasil.

Diana Santos and Nuno Cardoso, editors. 2007. *Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM, a primeira avaliação conjunta na área*. Linguateca, 12 de Novembro.