In this paper, we describe the two "Yahoo" named entity tagging (NER) systems. Both implement a Hidden Markov Model, based on a regularized perceptron classifier. The main feature of our approach is the choice of one model, in terms of features and classifier, for both the PoS and named entity tasks in the Evalita shared task.

Keywords: PoS tagging, HMMs, perceptrons.

1 Introduction

One of the main goals of our research is to investigate the application of natural language processing methods to information retrieval tasks which are challenging for traditional approaches (e.g., see [4]). As explained in the companion paper on our Evalita PoS system [3], since we expect to carry out several processing steps such as PoS tagging, named-entity detection, semantic role labeling, parsing, etc. on extremely large datasets, efficiency is a priority. To develop an efficient pipeline it is convenient to reduce feature extraction and use one single shared representation, i.e., based on the same features, for both PoS and NER. This way we perform feature extraction once, then PoS tagging, after which PoS-based features can be added to the initial data and passed on to the NER system. In the NER experiments reported we use the provided PoS tags, but the features extracted are the same as those utilized by the PoS system described in [3].

Our NER system was originally developed for English. However, the base model is essentially language-independent since no additional information is used beyond the provided training data, and no language-specific processing (e.g., lemmatization) is carried out\(^1\). Hence, in practice, we also evaluate the portability of a simple approach to new languages. Experimental results show that this approach is competitive and extremely efficient\(^2\).

2 HMM tagger and features

Our tagger (see [2] for more details), is an HMM trained with the sequence perceptron introduced in [1]. Label to label dependencies are limited to the previous tag (first order HMM). Models have a single adjustable parameter, the number of training iterations, chosen on the training data by cross-validation. Models are regularized by averaging [1], and by adding a constant feature to each token. We use the same exact features (1-4) described in [3], to this set we add the following PoS features for each token \(w_i\):

5. PoS: PoS tag of \(w_i\), \(w_{i-1}\) and \(w_{i+1}\);

3 Systems results and discussion

In addition to the base tagger we trained a second one in which we added extra training data gathered from the Italian Wikipedia. We compiled a list of names referring to streets, famous people, regions, lakes and rivers, institutions, most important towns, political parties, companies, universities and prominent organizations. Overall 3,184 entity names, which were PoS tagged using a tagger trained on the PoS tags of the NER training data, then added to the data with the corresponding label; e.g., ORG for universities and political parties, LOC for street and lake names, etc. Although the second model performed worse than the simpler base model in cross-validation, we

\(^1\)Our second, and less-successful, system utilizes some external information in the form of additional training data.

\(^2\)The tagger, called "sst-light", is available from http://sourceforge.net/projects/supersensetag/
Table 1: Results in terms of overall F-score, and F-score itemized per entity type, of the systems on the two tagsets. The table reports also the number of features for each model, the number of training iterations, the time it took to train and the speed in tagging the evaluation set which consists of 4,002 sentences (around 79 thousand tokens).

<table>
<thead>
<tr>
<th>System</th>
<th>T</th>
<th>F</th>
<th>Train</th>
<th>Test</th>
<th>Sent/s</th>
<th>F-ALL</th>
<th>F-GPE</th>
<th>F-LOC</th>
<th>F-ORG</th>
<th>F-PER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yahoo_Ciaramita_NER_s1</td>
<td>50</td>
<td>66K</td>
<td>39.40m</td>
<td>3.3s</td>
<td>1,200</td>
<td>68.99</td>
<td>75.38</td>
<td>52.83</td>
<td>49.08</td>
<td>78.89</td>
</tr>
<tr>
<td>Yahoo_Ciaramita_NER_s2</td>
<td>50</td>
<td>69K</td>
<td>44.33m</td>
<td>3.8s</td>
<td>1,050</td>
<td>68.15</td>
<td>75.08</td>
<td>52.31</td>
<td>46.85</td>
<td>78.36</td>
</tr>
</tbody>
</table>

Table 1 summarizes the systems results and additional information. We found the accuracy of the NER tagger unimpressive and somewhat puzzling. The same tagger, in a slightly more sophisticated implementation, evaluated on several tasks achieves state of the art accuracies; e.g., ACE (76% F-score, see [5]), BBN-WSJ Entity Corpus (87%), CoNLL (91%). Arguably, the ACE and BBN-WSJ tasks are more challenging; e.g., they involve more categories, respectively 46 and 105, and ACE also partially covers pronouns. The Evalita NER dataset adopts the ACE principles but there are some clear departures from the ACE standards which might explain this discrepancy. For example, here the name “Prodi” would not be annotated as an entity (PER) in a phrase like “Il governo di Prodi” (“Prodi’s government”), because it actually refers to the government (I-CAB guidelines). However, in the ACE 2007 data in phrases like “X’s administration” X is always tagged as person (Clinton, Bush, Arafat, etc.). Subtle distinctions such as above might be hard to apply consistently, even for humans; e.g., in the phrase “vuole affidare Prodi” (“wants to sink Prodi”), which probably refers to the government and not the person, Prodi is tagged as person.

REFERENCES


CONTACTS

**MASSIMILIANO CIARAMITA, JORDI ATSERIAS**

Yahoo! Research
Ocata 1, Barcelona
08003 Catalunya, Spain

Email: {massi | jordi}@yahoo-inc.com

**MASSIMILIANO CIARAMITA** is a researcher at Yahoo! Research Barcelona. He received degrees from the Università di Roma “La Sapienza” (BA), and Brown University (Sc.M, Ph.D.). He held a research fellowship from the Italian National Research Council (CNR) in Rome between 2004 and 2006. He has worked on several topics in computational linguistics such as knowledge acquisition, semantic tagging, information extraction and parsing, currently his research concerns the application of natural language processing and machine learning to a range of information retrieval problems.

**JORDI ATSERIAS** is a research engineer at Yahoo! Research Barcelona. He obtained the BS in Computer Science in 1994 at the Facultat d’Informàtica de Barcelona and his Ph.D. at the Euskal Herriko Unibertsitatea (University of the Basque country) in 2006. He was previously working at the TALP research group (Universitat Politecnica de Catalunya) and was involved in several European and Spanish projects related to NLP technologies. His research in NLP has focused mainly on parsing, word sense disambiguation and semantic role labeling.