Poly-co : an unsupervised co-reference detection system

Éric Charton, Michel Gagnon, Benoit Ozell
École Polytechnique de Montréal
2900 boulevard Edouard-Montpetit, Montreal, QC H3T 1J4, Canada.
{eric.charton, michel.gagnon, benoit.ozell}@polymtl.ca

Abstract

We describe our contribution to the Generation Challenge 2010 for the tasks of Named Entity Recognition and co-reference detection (GREC-NER). To extract the NE and the referring expressions, we employ a combination of a Part of Speech Tagger and the Conditional Random Fields (CRF) learning technique. We finally experiment an original algorithm to detect co-references. We conclude with discussion about our system performances.

1 Introduction

Three submission tracks are proposed in Generation Challenges 2010. **GREC-NEG**, where participating systems select a referring expression (RE) from a given list. **GREC-NER** where participating systems must recognize all mentions of people in a text and identify which mentions co-refer. And **GREC-Full**, end-to-end RE regeneration task; participating systems must identify all mentions of people and then aim to generate improved REs for the mentions. In this paper we present an unsupervised CRF based Named Entity Recognition (NER) system applied to the **GREC-NER** Task.

2 System description

The proposed system follows a pipelined architecture (each module processes the information provided by the previous one). First, a *Part of Speech* (POS) tagger is applied to the corpus. Then, the combination of words and POS tags are used by a CRF classifier to detect *Named Entities* (NE). Next, logical rules based on combination of POS tags, words and NE labels are used to detect pronouns related to *persons*. Finally, an algorithm identifies, among the *person* entities that have been detected, the ones that co-refer and cluster them. At the end, all collected information is aggregated in a XML file conform to **GREC-NER** specifications.

2.1 Part of speech

The part of speech labeling is done with the English version of Treetagger[1]. It is completed by a step where every *NAM* tag associated to a first nname is replaced by a *FNAME* tag, using a lexical resource of first names (see table 2, column *POS Tag*). The first name tag improves the NE detection model while it improves the estimation of conditional probabilities for words describing a person, encountered by a NER system.

Word from Corpus	POS Tag	NE Tag
Adrianne	FNAM	PERS
Calvo	NAM	PERS
enrolled	VVD	UNK
at	IN	UNK
Johnson	NAM	ORG
Wales	NAM	ORG
College	NAM	ORG

Table 2: Sample of word list with POS Tagging and NE tagging

2.2 Named entity and pronoun labeling

The Named Entity Recognition (NER) system is an implementation of the CRF based system (Béchet and Charton, 2010) that has been used in the French NER evaluation campaign ESTER 2 (Galliano et al., 2009)[2]. For the present task, training of the NER tool is fully unsupervised as it does not use the GREC training corpus. It is trained in English with an automatically NE annotated version of the Wikipedia Corpus (the full system configuration is described in (Charton and

[1] This work is granted by Unima Inc and Prompt Québec

[1] The Tree-tagger is a tool for annotating text with part-of-speech and lemma information. http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

[2] Referenced in this paper as *LIA*

Poly-co Score	B3			CEAF			MUC		
Set	Precision	Recall	FScore	Precision	Recall	FScore	Precision	Recall	FScore
Full set	91.48	85.89	**88,60**	85.40	85.40	**85.40**	92.15	86.95	**89.47**
Chef	91.12	87.84	**89.45**	86.53	86.53	**86.53**	91.86	88.55	**90.18**
Composers	92.01	87.14	**89.51**	86.87	86.87	**86.87**	92.11	87.02	**89.49**
Inventors	91.27	82.63	**86.74**	82.73	82.73	**82.73**	92.48	85.29	**88.74**

Table 1: System results obtained on dev-set

Torres-Moreno, 2010)). It is able to label PERS[3], ORG, LOC, TIME, DATE. We measured a specific precision of 0,93 on PERS NE detection applied to the English ACE[4] evaluation set.

Following the NE detection process, detection rules are used to label each personal pronoun with the PERS tag. Boolean *AND* rules are applied to triples {*word, POS tag, NE tag*}, where *word* = {*he, him, she, her ...*}, *POS tag=NN*, and *NE tag=UNK* . This rule structure is adopted to avoid the PERS labeling of pronouns included in an expression or in a previously tagged NE (i.e a music album or a movie title, using word *She*, and previously labeled with PROD NE tag). Finally, each PERS labeled entity is numbered by order of apparition and is associated with the sentences reference number where it appears (consecutive PERS labeled words, not separated by punctuation mark, receive the same index number).

2.3 Entities clustering by unstacking

In the final stage, our system determines which entities co-refer. First, a clustering process is achieved. The principle of the algorithm is as follows: entities characteristics (words, POS tags, sentence position) are indexed in a stack, ordered according to their chronological apparition in the text (the entity at the top of the stack is the first one that has been detected in the document). At the beginning of the process, the entity that is at the top of the stack is removed and constitutes the first item of a cluster. This entity is compared sequentially, by using similarity rules, with every other entities contained in the stack. When there is a match, entity is transfered to the currently instantiated cluster and removed from the stack. When the end of the stack is reached, remaining entities are reordered and the process iterates form the beginning. This operation is repeated until the stack is empty.

Comparison of entities in the stack is done in two ways according to the nature of the entity. We consider a candidate entity E_c from stack S. According to iteration k, the current cluster is C_k. Each element of the sequence E_c (i.e *Chester FNAME Carton NAM*) is compared to the sequences previously transfered in C_k during the exploration process of the stack. If $E_c \subseteq \bigcup C_k$, it is included in cluster C_k and removed from S. Finally inclusion of pronouns from S in E_c is done by resolving the anaphora, according to the Hobbs algorithm, as described in (Jurafsky et al., 2000)[5].

3 Results and conclusions

Table 1 shows our results on dev-set. We obtain good precision on the 3 subsets. Our system slightly underperforms the recall. This can be explained by a good performance in the NE detection process, but a difficulty in some cases for the clustering algorithm to group entities. We have observed in the Inventors dev-set some difficulties, due to strong variation of surface forms for specific entities. We plan to experiment the use of an external resource of surface forms for person names extracted from Wikipedia to improve our system in such specific case.

References

Frédéric Béchet and Eric Charton. 2010. Unsupervised knowledge acquisition for extracting named entities from speech. In *ICASSP 2010*, Dallas. ICASSP.

Eric Charton and J.M. Torres-Moreno. 2010. NLGbAse: a free linguistic resource for Natural Language Processing systems. In LREC 2010, editor, *English*, number 1, Matla. Proceedings of LREC 2010.

S. Galliano, G. Gravier, and L. Chaubard. 2009. The ESTER 2 Evaluation Campaign for the Rich Transcription of French Radio Broadcasts. In *International Speech Communication Association conference 2009*, pages 2583–2586. Interspeech 2010.

D. Jurafsky, J.H. Martin, A. Kehler, K. Vander Linden, and N. Ward. 2000. *Speech and language processing*. Prentice Hall New York.

[3] PERS tag is commonly used in NER Task to describe labels applied to people, ORG describe organisations, LOC is for places.

[4] ACE is the former NIST NER evaluation campaign.

[5] p704, 21.6