Is this NE tagger getting old?

Language Resources and Evaluation Conference
Marrakech, Morocco - May 28th - 30th 2008

Cristina Mota and Ralph Grishman

IST & L2F INESC-ID (Portugal) & NYU (USA)
and
New York University (USA)

(Advisors: Ralph Grishman & Nuno Mamede)

This research was funded by *Fundação para a Ciência e a Tecnologia* (doctoral scholarship SFRH/BD/3237/2000)

Outline

1. Introduction
2. Corpus Analysis
3. NER Performance Analysis
4. Experiments
5. Final Remarks

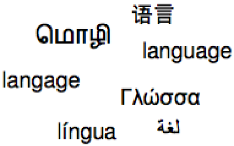
1. Introduction
   - Motivation
   - Approach
2. Corpus Analysis
3. NER Performance Analysis
4. Experiments
5. Final Remarks

What is NER?

Mary is studying in Rabat at Mohammed V University

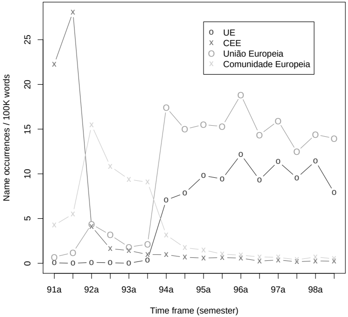
⇊ **NE Tagger** ⇊

Mary$_{PER}$ is studying in **Rabat**$_{LOC}$ at **Mohammed V University**$_{ORG}$



The Problem

O SÉCULO
O GENERAL ANTÓNIO DE SPÍNOLA PRESIDE À JUNTA DE SALVAÇÃO NACIONAL
Público



- Do texts vary over time in a way that affects NE recognition?
- Should NE taggers be also conceived time-aware?

Approach

Corpus Analysis

Measure corpus similarity based on

- Words

Compute name list overlaps

- By type
- By token

NER Performance Analysis

Assess performance by training and testing with different configurations (train,test)

- Increase time gap between training and test data

1. Introduction

2. Corpus Analysis
   - Corpus Similarity Algorithm (Kilgarriff, 2001)
   - Name List Overlaps

3. NER Performance Analysis

4. Experiments

5. Final Remarks

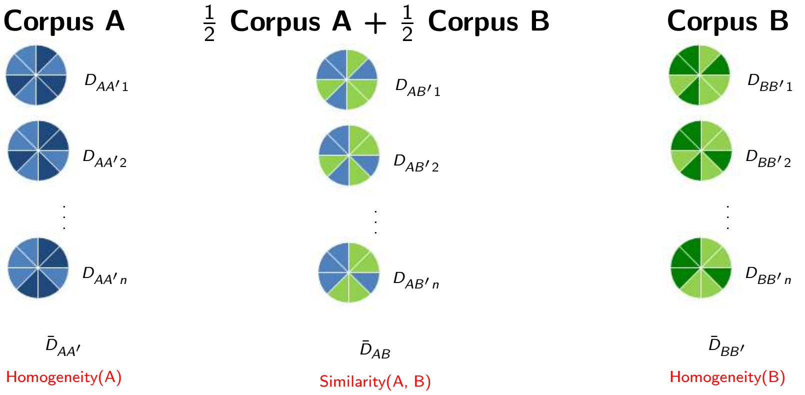
Corpus Similarity Algorithm (Kilgarriff, 2001)

Similarity(A,B):

- Split corpus A and B into k slices each
- Repeat m times:
    - Randomly allocate $\frac{k}{2}$ slices to A_i and $\frac{k}{2}$ to B_i
    - Construct word frequency lists for A_i and B_i
    - Compute CBDF between A and B for the n most frequent words of the joint corpus $(A_i + B_i)$
      [CBDF $= \chi^2$ by degrees of freedom]
- Output mean and standard deviation of CBDF of all experiments

Repeat using corpus A only: Similarity(A,A) → Homogeneity(A)
Repeat using corpus B only: Similarity(B,B) → Homogeneity(B)

Corpus Similarity Algorithm (Kilgarriff, 2001)



Corpus A	$\frac{1}{2}$ **Corpus A** + $\frac{1}{2}$ **Corpus B**	**Corpus B**
$D_{AA'1}$	$D_{AB'1}$	$D_{BB'1}$
$D_{AA'2}$	$D_{AB'2}$	$D_{BB'2}$
$\vdots$	$\vdots$	$\vdots$
$D_{AA'n}$	$D_{AB'n}$	$D_{BB'n}$
$\bar{D}_{AA'}$	$\bar{D}_{AB}$	$\bar{D}_{BB'}$
Homogeneity(A)	Similarity(A, B)	Homogeneity(B)

Lower values of $\bar{D}$ $\Rightarrow$ higher homogeneity/similarity

Name List Overlaps

$$type\_overlap = \frac{|T_A \cap T_B|}{|T_A| + |T_B| - |T_A \cap T_B|} \quad (1)$$

$$token\_overlap = \frac{\sum_{i=1}^{N} min(f_A(i), f_b(i))}{\sum_{i=1}^{N} max(f_A(i), f_B(i))} \quad (2)$$

T_A = list of different names (name types) of text A
$f_A(i)$ = frequency of name i in text A

Name List Overlaps

A name list: Mary (3), Rabat (5), Mohammed V University (4)
B name list: John (1), Rabat (2), Mohammed V Universirty (6)

Type Overlap

$$\frac{|\{Rabat, MohammedVUniversity\}|}{|\{Mary, Rabat, MohammedVUniversity, John\}|} = 2/4$$

Token Overlap

$$\frac{min(3,0) + min(5,2) + min(4,6) + min(0,1)}{max(3,0) + max(5,2) + max(4,6) + max(0,1)} = 6/15$$

1 Introduction

2 Corpus Analysis

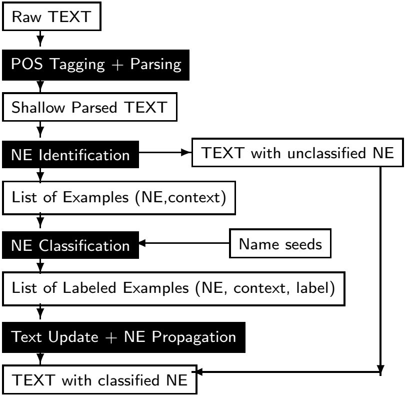
3 NER Performance Analysis

- NE Tagger Description (Collins & Singer, 1999)

4 Experiments

5 Final Remarks

NE Tagger Description (Collins & Singer, 1999)



Raw TEXT

POS Tagging + Parsing

Shallow Parsed TEXT

NE Identification

TEXT with unclassified NE

List of Examples (NE,context)

NE Classification

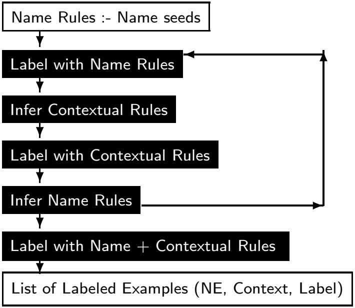
Name seeds

List of Labeled Examples (NE, context, label)

Text Update + NE Propagation

TEXT with classified NE

Classification in detail:



Name Rules :- Name seeds

Label with Name Rules

Infer Contextual Rules

Label with Contextual Rules

Infer Name Rules

Label with Name + Contextual Rules

List of Labeled Examples (NE, Context, Label)

1 Introduction

2 Corpus Analysis

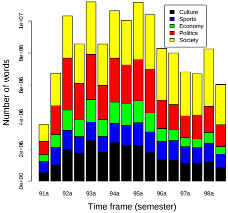
3 NER Performance Analysis

4 Experiments

- Experimental Setting
- F-Measure over Time
- Politics Dissimilarity over Time
- Politics Name List Overlap over Time
- F-Measure compared to Dissimilarity

5 Final Remarks

Experimental Setting



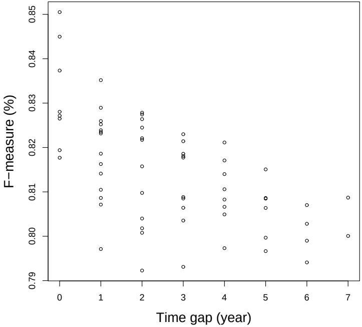
CETEMPublico (Santos & Rocha, 2001) is a Portuguese public journalistic corpus

- Size: 180 million words
- Time span: 8 years
- Organization: randomly shuffled extracts [1 extract ≊ 2 paragraphs]
- Classification: 10 topics and 16 time frames (year + semester)
- Mark up: paragraphs, sentences, enumeration lists and authors

Experimental Setting

- **Topic:** politics
- **Time unit:** year
- **Text unit:** sentence
- **Size:** 10 slices x 60000 words per time frame
- **N most frequent words:** 2000 words
- **Names compared:** 82400 per time frame
- **Seeds (S):** different names in the first 2500 name instances [first 198 extracts per semester]
- **Test (T):** next 208 extracts per semester grouped by year
- **Unlabeled examples (U):** first 82456 names with context per year [following 7856 extracts]

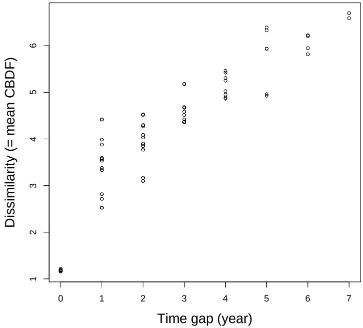
NER Performance: F-Measure over Time



- When the texts are from the same year (time gap = 0), the F-measure ranges approximately from 82% to 85%
- When the texts are 5 years apart the F-measure ranges from about 79% to 82%
- As the time gap between (S_k, U_k) and T_j increases, the F-measure shows a tendency to decay

Training-test configuration: (S_i, U_i, T_j), i=91..98, j=91..98 [64 tests]

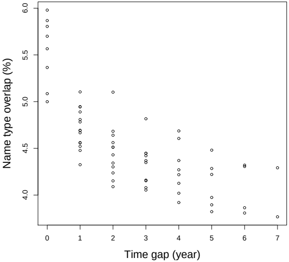
Politics Corpus Dissimilarity over time

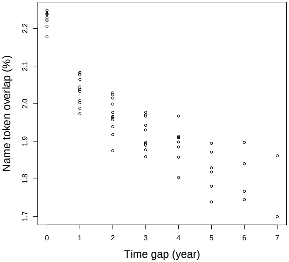


- The homogeneity for all the texts is very close to 1
- Increasing the time gap to one year, the dissimilarity ranges from 2.5 to 4.5
- At a distance of five years dissimilarity ranges from 4.7 to almost 6.5
- The dissimilarity shows a tendency to increase as the time gap increases

Corpus comparisons: (U_i, U_j), i=91..98, j=91..98 [64 comparisons; Higher values = Lower similarity]

Politics Name List Overlap over Time

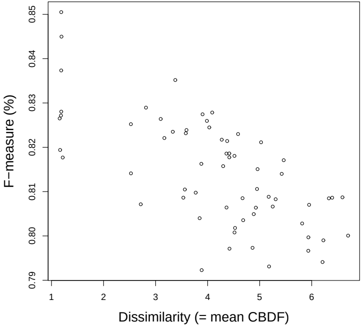




- Within the same time frame, the type overlap varies between 5% and 6%
- At a distance of 5 years it varies between 3.5% and 4.5%
- Within the same year, the name token overlap varied between 4.2% and 4.4%
- At distance of 5 years varied between 3.2% and 3.7%
- Overlap between name lists also decreases over time

Corpus comparisons: (U_i, T_j), i=91..98, j=91..98 [64 comparisons]

F-Measure compared to Dissimilarity



OBS: Higher values = Lower similarity

- There is an inverse association between dissimilarity and F-measure: for higher levels of dissimilarity (i.e, higher distance values) we obtain lower performance values

1 Introduction

2 Corpus Analysis

3 NER Performance Analysis

4 Experiments

5 Final Remarks
- Main Results
- Work in Progress

Main Results

Within a period of 8 years we observed that:

- Corpus similarity and name overlaps tend to decrease as the two corpora become more temporally distant
- The performance of a co-training based NE tagger trained and tested on those texts shows a decay as we increase the time gap between the training and the test data
- There is an association between the results of the corpus analysis and the tagger performance

Work in Progress

Other related issues we are currently investigating aiming at better named entity recognition

- Analyze the NE surrounding contexts to verify if they also tend to overlap less over time
- Investigate how we can avoid the performance decay
  - Do we need more data?
  - Do we need more labeled data within the same time frame?
  - Do we need more unlabeled data within the same time frame?