Journalistic Corpus Similarity over Time

American Association for Corpus Linguistics Conference
Provo, Utah - March 12th - 15th 2008

Cristina Mota

Instituto Superior Técnico & L2F INESC-ID (Portugal)
&
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. Methodology
3. Experiments
4. Final Remarks
1. Introduction
   - Motivation
   - Approach

2. Methodology

3. Experiments

4. Final Remarks
Mary is studying in the United States at Brigham Young University.

Mary_{PER} is studying in the United States_{LOC} at Brigham Young University_{ORG}. 
**Motivation**

- 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
- Upper case words
- Lower case words
Compute name and context overlaps
- By type
- By token

NER Performance Analysis
Assess performance by training and testing with different configurations (train,test)
- Time issue
- More data or better data?
- Labeled vs. unlabeled
1 Introduction

2 Methodology
   • Corpus Similarity Algorithm (Kilgarriff, 2001)
   • Application

3 Experiments

4 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 ($A_i + B_i$)
  - $[\text{CBDF} = \chi^2$ by degrees of freedom$]$
- Output mean and standard deviation of CBDF of all experiments

Repeat using corpus A only: Similarity(A,A) $\rightarrow$ Homogeneity(A)
Repeat using corpus B only: Similarity(B,B) $\rightarrow$ Homogeneity(B)
Application

**Corpus A**

\[ D_{AA'}^1 \]
\[ D_{AA'}^2 \]
\[ \vdots \]
\[ D_{AA'}^n \]

\[ \bar{D}_{AA'} \]

Homogeneity(A)

\[ \frac{1}{2} \text{ Corpus A } + \frac{1}{2} \text{ Corpus B } \]

\[ D_{AB'}^1 \]
\[ D_{AB'}^2 \]
\[ \vdots \]
\[ D_{AB'}^n \]

\[ \bar{D}_{AB} \]

Similarity(A, B)

**Corpus B**

\[ D_{BB'}^1 \]
\[ D_{BB'}^2 \]
\[ \vdots \]
\[ D_{BB'}^n \]

\[ \bar{D}_{BB'} \]

Homogeneity(B)

Lower values of \( \bar{D} \) ⇒ higher homogeneity/similarity
1 Introduction

2 Methodology

3 Experiments
   - Experimental Setting
   - Within-topic homogeneity
   - Within-topic similarity over time
   - Cross-topic similarity over time
   - Within-topic similarity using lower and upper case words

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

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

Experiment parameters:

- Topics: culture, sports, economy, politics and society
- Time unit: semester
- Text unit: sentence
- Size: 10 slices x 37600 words per topic and time frame (80% lower case and 8% upper case)
- N most frequent words: 2000 words (80% lower case and 8% upper case)
Within-topic homogeneity

Very low values of homogeneity for each semester within each topic
- Politics has the most homogeneous semesters
- Culture has the least homogeneous semesters
- Economy values spread widely

What happens if we begin comparing similarity between semesters?
Within-topic similarity decreases over time

The similarity decay is more significant for some topics
Cross-topic similarity doesn’t show a decreasing pattern

Within-topic similarity over time tends to cross-topic similarity
Within-topic similarity using lower and upper case words

- Similarity based on upper case words decreases more significantly
Within-topic similarity using lower and upper case words

Similarity profiles depend on types of words
→ Sports and Politics have similar profiles with lower case words
→ Politics varies much more than Sports based on upper case words
The within-topic similarity between texts decreases as we increase the time gap between them
  - This tendency doesn't seem to flatten in a short period of 8 years

The within-topic similarity decrease is more significant for some topics

In some cases the within-topic similarity values get close to cross-topic similarity values

The similarity based on capitalized words decreases in a more pronounced way
Final Remarks

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

- Does the decrease in similarity affect the performance of named entity recognition?
- How do other more structured lexical units (e.g., named entities and surrounding context) vary over time?