

# Using Temporal Language Models for Document Dating

Nattiya Kanhabua and Kjetil Nørvåg

Norwegian University of Science and Technology

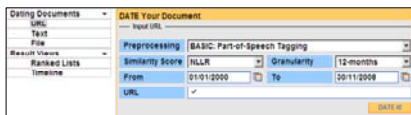
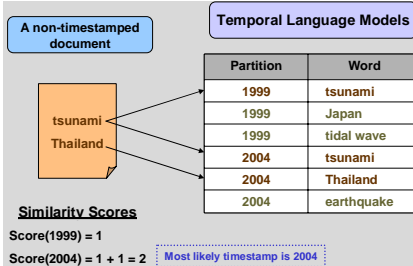
## Overview

**Problem statement:** Due to the decentralized nature and the lack of standards for date/time of web documents, it is difficult to find accurate and trustworthy timestamps

**Demo:** A tool for determining the timestamp of a non-timestamped document using temporal language models

### Document Dating Model

- Temporal Language Models proposed by de Jong et al. [1]
- Based on the statistic usage of words over time
- Compare a non-timestamped document with a reference corpus.
- A reference time partition mostly overlaps in term usage is the tentative timestamp



A live demo can be found at: <http://comidor02.idi.ntnu.no:8080/timedelivery/>

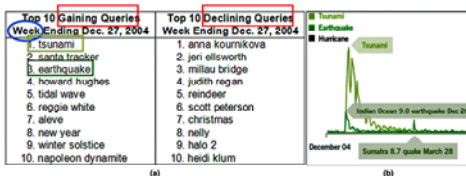
[1] F. de Jong, H. Rode, and D. Hiemstra. Temporal language models for the disclosure of historical text. In Proceedings of AHC/2005 (History and Computing), 2005.

## Search Statistics to Enhance Scores

**Motivation:** Search statistics provided by Google Zeitgeist (GZ) can be integrated as an additional score in order to increase the probability of a tentative time partition

### Approach:

- Analyze search statistics to increase the probability for a time partition containing top-ranked queries
- The higher probability the partition has, the more likely candidate it is



A linear combination of a GZ score to an original similarity score

### Parameters

- Part-of-speech tagging, collocation, WSD, concept extraction
- Similarity: Normalized log-likelihood ratio
- Input: <http://tsunami-thailand.blogspot.com>

### Results

- The correct time partition@7 (2004/12-2005/11)

| Ranked ID | Date Range        | Similarity Score | % Confidence |
|-----------|-------------------|------------------|--------------|
| 1         | 2004/12 - 2005/11 | 1.00             | 33.21        |
| 2         | 2002/12 - 2003/12 | 0.67             | 24.70        |
| 3         | 2003/12 - 2004/12 | 0.42             | 6.52         |
| 4         | 2001/12 - 2002/12 | 0.42             | 2.45         |
| 5         | 2000/12 - 2001/12 | 0.39             | 6.20         |
| 6         | 2000/01 - 2000/12 | 0.33             | 7.75         |
| 7         | 2007/11 - 2008/11 | 0.29             | 3.80         |
| 8         | 2005/11 - 2006/11 | 0.21             | 0.24         |
| 9         | 2006/11 - 2007/11 | 0.21             | 0.00         |

### Parameters

- Part-of-speech tagging, collocation, WSD, concept extraction
- Similarity: GZ
- Input: <http://tsunami-thailand.blogspot.com>

### Results

- The correct time partition@1 (2004/12-2005/11)

## Basic vs. Advanced Preprocessing

**Motivation:** Determining timestamp of a document from a direct comparison between extracted words and corpus partitions has limited accuracy

**Approach:** Integrate semantic-based techniques into document preprocessing

| Semantic-based Preprocessing | Description   |
|------------------------------|---|
| Part-of-speech tagging       | Select only interesting classes of words, e.g. nouns, verbs, and adjectives                                       |
| Collocation extraction       | Co-occurrence of different words can alter the meaning, e.g. "United States"                                      |
| Word sense disambiguation    | Identify the correct sense of a word from context, e.g. "bank"  |
| Concept extraction           | Compare concepts instead of original words, e.g. "tsunami" and "tidal wave" have the common concept of "disaster" |
| Word filtering               | Select the top-ranked words according to TF-IDF scores for a comparison   |

### Basic preprocessing

#### Parameters

- Part-of-speech tagging
- Similarity: Normalized log-likelihood ratio
- Input: <http://tsunami-thailand.blogspot.com>

#### Results

- Short processing time (6 s)
- Lower accuracy (correct partition@7)

| Ranked ID | Date Range        | Similarity Score | % Confidence |
|-----------|-------------------|------------------|--------------|
| 1         | 2000/01 - 2000/12 | 1.00             | 3.95         |
| 2         | 2000/12 - 2001/12 | 0.96             | 18.10        |
| 3         | 2003/12 - 2004/12 | 0.78             | 2.89         |
| 4         | 2007/11 - 2008/11 | 0.75             | 0.20         |
| 5         | 2000/12 - 2001/12 | 0.75             | 3.92         |
| 6         | 2001/12 - 2002/12 | 0.71             | 6.30         |
| 7         | 2004/12 - 2005/11 | 0.65             | 1.43         |
| 8         | 2005/11 - 2006/11 | 0.63             | 0.37         |
| 9         | 2006/11 - 2007/11 | 0.63             | 0.00         |

| Ranked ID | Date Range        | Similarity Score | % Confidence |
|-----------|-------------------|------------------|--------------|
| 1         | 2007/11 - 2008/11 | 1.00             | 0.42         |
| 2         | 2006/11 - 2007/11 | 1.00             | 0.57         |
| 3         | 2005/11 - 2006/11 | 0.99             | 4.89         |
| 4         | 2003/12 - 2004/12 | 0.94             | 0.74         |
| 5         | 2004/12 - 2005/11 | 0.93             | 7.77         |
| 6         | 2002/12 - 2003/12 | 0.88             | 8.22         |
| 7         | 2001/12 - 2002/12 | 0.77             | 6.56         |
| 8         | 2000/01 - 2000/12 | 0.71             | 9.03         |
| 9         | 2000/12 - 2001/12 | 0.62             | 0.00         |

### Advanced preprocessing

#### Parameters

- Part-of-speech tagging, collocation, WSD, concept extraction
- Similarity: Normalized log-likelihood ratio
- Input: <http://tsunami-thailand.blogspot.com>

#### Results

- Small additional processing time (14 s)
- Higher accuracy (correct partition@5)

## Temporal Entropy as a Trend

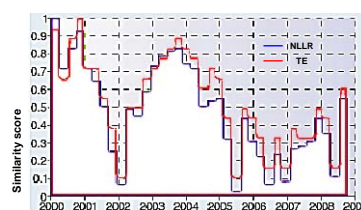
**Motivation:** Temporal entropy (TE), based on a term selection presented in Lochbaum and Streeter [2], weights terms differently depending on how well a term is suitable for separating time partitions and tells how important a term is in a specific partition

| Temporal Entropy  |
|---|
| A measure of temporal information which a word conveys.   |
| Captures the importance of a term in a <b>document collection</b> whereas TF-IDF weights a term in a particular document.   |
| Tells how good a term is in separating a <b>partition</b> from others.  |
| A term occurring in few <b>partitions</b> has higher temporal entropy compared to one appearing in many <b>partitions</b> . |
| The higher temporal entropy a term has, the better representative of a <b>partition</b> .                                   |

$$A \text{ probability of a partition } p \text{ containing a term } w_i = \frac{tf(w_i, p)}{\sum_{k=1}^p tf(w_i, p_k)}$$

$$TE(w_i) = 1 + \frac{1}{\log N_p} \sum_{p \in P} P(p|w_i) \times \log P(p|w_i)$$

$N_p$  is the total number of partitions in a corpus



### Parameters

- Part-of-speech tagging, collocation, WSD, concept extraction
- Input: US presidential election

**Normalized log-likelihood ratio (NLLR)**

- Lower scores for relevant periods (2000, 2004 and 2008)

### Temporal Entropy (TE)

- Higher scores for relevant periods (2000, 2004 and 2008)

[2] K. E. Lochbaum and L. A. Streeter. Comparing and combining the effectiveness of latent semantic indexing and the ordinary vector space model for information retrieval. Inf. Process. Manage., 25(6):665-676, 1989.