Event Detection in Newspaper Texts

Nikola Ljubešić, senior research assistant
Department of Information Sciences
University of Zagreb

JOTA, 28 October 2010
Overview

1. What is event detection
2. Document clustering
3. The gold standard
4. Experimental setup
5. Results
6. Further steps
Event detection

- event - a particular thing that happens at a specific time and place (TDT, 2004)
- event detection - process of detecting an event description in a piece of information
- part of the topic detection and tracking problem set
- document : event == 1 : 1?
Classification problem

- events are categories - classification task

1. unknown classification schema - solvable only by unsupervised classification - clustering

2. unknown number of events - unknown number of classes - hierarchical clustering
Document clustering

document formalization

distance matrix

clustering
Gold standard

- 2,398 documents published on 17 Croatian news portals in three days
- two annotators, application developed for that purpose
- pooling - using a combination of all similarity metrics to obtain a candidate list
- built 1,214 and 955 clusters
Inter-annotator agreement

<table>
<thead>
<tr>
<th></th>
<th>Formula</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>kappa</td>
<td>$\kappa = \frac{2</td>
<td>A_1 \cap A_2</td>
</tr>
<tr>
<td>modified kappa</td>
<td>$\kappa_{mod} = \frac{</td>
<td>A_1 \cap A_2</td>
</tr>
</tbody>
</table>

- biggest story of May 3, 2009 - the Myanmar cyclone
- annotator 1 - one cluster with 52 documents
- annotator 2 - three clusters - the catastrophe, first rescue operations, Croatian Red cross reaction
Event cluster distribution

Number of clusters

Cumulative document function

Size of the cluster

Number of clusters

0 15 30 45 60

0 225 450 675 900
Workload-recall trade-off

workload document-loss function
Experimental setup

• 14 categorical variables with 2-6 levels - 2,073,600 experiments

• huge search space - independence assumption

• variable categories:
  • clustering algorithm
  • distance metrics
  • feature weight measures
  • feature selection and extraction methods
  • reference corpus significance
### Evaluation measures

<table>
<thead>
<tr>
<th>Measure</th>
<th>Formula</th>
</tr>
</thead>
<tbody>
<tr>
<td>Purity</td>
<td>$ purity(\Omega, C) = \frac{1}{N} \sum_k \max</td>
</tr>
<tr>
<td>Normalized mutual information</td>
<td>$ NMI(\Omega, C) = \frac{I(\Omega; C)}{[H(\Omega) + H(C)]^{0.5}} $</td>
</tr>
<tr>
<td>Rand index (accuracy)</td>
<td>$ RI = \frac{TP + TN}{TP + FP + TN + FN} $</td>
</tr>
<tr>
<td>Precision, recall</td>
<td>$ P = \frac{TP}{TP + FP}, \quad R = \frac{TP}{TP + FN} $</td>
</tr>
<tr>
<td>$ F_\beta $</td>
<td>$ F_\beta = \frac{(\beta^2 + 1)PR}{\beta^2P + R} $</td>
</tr>
</tbody>
</table>
Clustering

- partitional vs. hierarchical
- retrospective vs. on-line
- linkage criterion in hierarchical algorithms
  - maximum - complete-link
  - minimum - single-link
  - mean - average-link
# Clustering algorithms

<table>
<thead>
<tr>
<th>algorithm</th>
<th>linkage criterion</th>
<th>time complexity</th>
<th>on-line</th>
</tr>
</thead>
<tbody>
<tr>
<td>hierarchical agglomerative</td>
<td>complete</td>
<td>$O(n^2 \log n)$</td>
<td>no</td>
</tr>
<tr>
<td>hierarchical agglomerative</td>
<td>average</td>
<td>$O(n^2 \log n)$</td>
<td>no</td>
</tr>
<tr>
<td>single-pass</td>
<td>single</td>
<td>$O(n)$</td>
<td>yes</td>
</tr>
</tbody>
</table>
Distance metrics

- **Manhattan**
- **Jaccard**
- **Euclidean**
- **Dice**
- **cosine**
- **Jensen-Shannon**

**F0.5**
- Manhattan: 0.701
- Jaccard: 0.433
- Euclidean: 0.791
- Dice: 0.793
- cosine: 0.801
- Jensen-Shannon: 0.803

**Time**
- Manhattan: 310
- Jaccard: 327
- Euclidean: 143
- Dice: 668
- cosine: 607
- Jensen-Shannon: 550
Feature weight measures

- Probability: 0.647
- Conditional probability: 0.754
- PMI: 0.758
- TF-IDF: 0.813
- T-test: 0.783
Feature selection

• character case and punctuation obsolete
• information in title more relevant, optimal repetition rate is four
• function words (IDF) - minor decrease in model and memory complexity
• hapax legomena - decreases number of dimensions drastically, memory 5-10%
Feature extraction

• stemming, POS tagging, lemmatization (two stemmers, TnT, HML)
• multi-word expressions (chi-square)
• named entity recognition (person and business entities)
• no significant improvement
Heuristics

1. an event ranges on a one-day time span true in 83% of documents (non-singleton events)

2. one source reports only once about an event - true in 86% of documents (non-singleton events)

• implementing heuristics increases $F_{0.5}$, first heuristic simplifies calculation drastically
Reference corpus

![Graph showing the F0.5 metric for different numbers of documents, with two lines representing 'unknown ? max' and 'no unknowns'.]
Primorac saslušao studente (vijesti.hrt.hr)
Studenti nakon sastanka s Primorcem ipak ne odustaju od prosvjeda (index.hr)
Primorac pokušava izbjeći studentske prosvjede razgovorom s Rektorskim zborom (business.hr)
Primorac primio organizatore studentskog štrajka (javno.com)
Studenti ne odustaju od najavljenog prosvjeda (dnevnik.hr)

VIDEO: Istukla i opljačkala susjedu zbog ljubomore (javno.com)
U stanu ju udarila palicom po glavi i opljačkala (index.hr)
Prijateljicu nevjenčanog supruga pretukla palicom i opljačkala (vecernji.hr)
Opalila je palicom u stanu i opljačkala (index.hr)

Sindikalna košarica u travnju 0,17 posto skuplja nego u ožujku (vecernji.hr)
Sindikalna košarica u travnju 0,17 posto skuplja (tportal.hr)
Sindikalna košarica u travnju 0,17 posto skuplja nego u ožujku (poslovni.hr)
Životni troškovi četveročlane obitelji 6206 kuna (business.hr)
Further steps

- windowing technique $\iff$ decay function
- feature position - features found at the beginning (in the first sentence?) should be given more weight
- write a Java API (Apache license)
- events $\implies$ topics;
  event : document relationship
Thank you!