NewsExplorer – Multilingual News Analysis with Cross-lingual Linking

Machine Learning for Multilingual Information Access
NIPS Workshop -- Whistler, Canada, 10 December 2006

Ralf Steinberger, Bruno Pouliquen, Camelia Ignat
European Commission – Joint Research Centre (JRC) – Italy

http://langtech.jrc.it/ http://press.jrc.it/NewsExplorer

Agenda

• Motivation for cross-lingual document linking
  • EU: need for multilinguality
  • Enhanced Information Extraction by combining information from texts in various languages
  • Demo of NewsExplorer (http://press.jrc.it/NewsExplorer)

• Related work
• Overview of our approach
• Description of the components, challenges
  • IE: Locations
  • IE: Person and Organisation names
  • Categorisation into Subject domains (Eurovoc classes)
  • Clustering of news
  • Linking clusters over time
  • Linking clusters across languages

• Future work and Conclusions
CLDS Methods / Related work

- Usage of Machine Translation \(\rightarrow\) monolingual document similarity (TDT-3, Leek et al. 1999)
- Usage of bilingual dictionaries \(\rightarrow\) monolingual document similarity (Wacltar 1999; Mathieu et al. 2004)
- Automatically produce bilingual lexical space for bilingual document representation and document similarity calculation, e.g.
  - bilingual Lexical Semantic Analysis (LSI, Landauer & Littman 1991)
  - Kernel Canonical Correlation Analysis (Vinokourov et al., 2002)

+ Achieved results are relatively good
- Bilingual approach is restricted to a few languages:

\[ \text{Language pairs} = \frac{N^2 - N}{2} \quad (N = \text{number of languages}) \]

- EU: 20 official languages \(\rightarrow\) 190 language pairs (380 language pair directions)!

\(\Rightarrow\) Attractiveness of highly multilingual interlingua approaches

---

Agenda

- Motivation for cross-lingual document linking
  - EU: need for multilinguality
  - Enhanced Information Extraction by combining information from texts in various languages
  - Demo of NewsExplorer (http://press.jrc.it/NewsExplorer)
- Related work
- Overview of our approach
- Description of the components, challenges
  - IE: Locations
  - IE: Person and Organisation names
  - Categorisation into Subject domains (Eurovoc classes)
  - Clustering of news
  - Linking clusters over time
  - Linking clusters across languages
- Future work and Conclusions
Approach to CLDS in a nutshell

- Map documents onto multilingual thesauri, nomenclatures, gazetteers, … (general, medical, technical, geographical, …)

- Identify and normalise language-independent text features
  - Numbers; dates; currency expressions; measurement expressions; …
    - le treize août 2007, am 13.8.2007, by 13/08/07 \(\rightarrow\) DATE_2007_08_13
  - Names of persons, organisations, …
    - Muqtada al-Sadr, Mouqtada Sadr, Mokdata Sadr, Muqtadā aṣ-Ṣadr, Муктада Садр
    \(\rightarrow\) ID=236

- Calculate the various vector similarities and add up
  - For details, see Steinberger et al. (2004, Informatica 28).

Information currently used in NewsExplorer

- Represent documents by vectors of ~language-independent features

  - S1: Multilingual thesaurus and classification system Eurovoc
  - S2: Locations (Latitude-Longitude information)
  - S3: Normalised and merged proper name variants (persons and organisations)
  - S4: Cognates and numbers

- \(\rightarrow\) CLDS (using cosine) based on these representations
  \[
  \text{CLDS} = \alpha \cdot S1 + \beta \cdot S2 + \gamma \cdot S3 + \delta \cdot S4
  \]
Agenda

- Motivation for cross-lingual document linking
  - EU: need for multilinguality
  - Enhanced Information Extraction by combining information from texts in various languages
  - Demo of NewsExplorer (http://press.jrc.it/NewsExplorer)
- Related work
- Overview of our approach
- **Description of the components, challenges**
  - IE: Locations
  - IE: Person and Organisation names
  - Categorisation into Subject domains (Eurovoc classes)
  - Clustering of news
  - Linking clusters over time
  - Linking clusters across languages
- Future work and Conclusions

Geo-coding multilingual text

- Place names cannot be recognised by looking for patterns in text (Gey 2000)
- Multilingual gazetteer needed
- Place name recognition via the **lookup** of text words in the gazetteer
Major challenges for geocoding (1)

1. Places homographic with common words

<table>
<thead>
<tr>
<th>English</th>
<th>French</th>
<th>German</th>
</tr>
</thead>
<tbody>
<tr>
<td>Place name</td>
<td>Country</td>
<td>Place name</td>
</tr>
<tr>
<td>And</td>
<td>Iran</td>
<td>De</td>
</tr>
<tr>
<td>To</td>
<td>Ghana</td>
<td>Du</td>
</tr>
<tr>
<td>Be</td>
<td>India</td>
<td>Lin</td>
</tr>
</tbody>
</table>

→ Geo-stopwords

2. Places homographic with person names

<table>
<thead>
<tr>
<th>Name</th>
<th>City: Country</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tony Blair</td>
<td>Tony: USA</td>
</tr>
<tr>
<td>Kofi Annan</td>
<td>Kofi: Mali</td>
</tr>
<tr>
<td>Javier Solana</td>
<td>Javier: Spain</td>
</tr>
</tbody>
</table>

→ Location only if not part of a person name

   e.g. ‘Kofi Annan’, ‘Annan’

Major challenges for geocoding (2)

3. Completeness of the gazetteer: exonyms, endonyms, …


4. Inflection

   • Romanian: **Parisului** (of Paris)
   • Estonian: **Londonit** (London), **New Yorgile** (New York)
   • Arabic: نوُسْتَرِيْبِّيْلَ (the Paris inhabitants)

   → Usage of context-sensitive suffix lists and replacement rules to generate all variants

   (Slovene example)

   Tony(a|o|u|om|em|mljuljem|ja)?\s+Blair(a|o|u|om|em|mljuljem|ja)
5. Homographic place names

→ Use size class information
→ Use country context
  (news source; unambiguous anchors)
→ Use kilometric distance
  E.g. “from Warsaw to Brest”
  Brest (France): 2000 km from Warsaw
  Brest (Belarus): 200 km from Warsaw

For details, see Pouliquen et al. (2006, LREC)

<table>
<thead>
<tr>
<th>Place name</th>
<th>Number of cities with this name</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aleksandrovka</td>
<td>244</td>
</tr>
<tr>
<td>Washington</td>
<td>32</td>
</tr>
<tr>
<td>London</td>
<td>18</td>
</tr>
<tr>
<td>Berlin</td>
<td>15</td>
</tr>
<tr>
<td>Paris</td>
<td>15</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Class</th>
<th>Explanation</th>
<th>Example</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>country name</td>
<td>Italy</td>
<td>80</td>
</tr>
<tr>
<td>1</td>
<td>capital</td>
<td>Rome</td>
<td>80</td>
</tr>
<tr>
<td>2</td>
<td>main city</td>
<td>Milan</td>
<td>80</td>
</tr>
<tr>
<td>3</td>
<td>province level</td>
<td>Varese</td>
<td>30</td>
</tr>
<tr>
<td>4</td>
<td>small city</td>
<td>Sesto Calende</td>
<td>20</td>
</tr>
<tr>
<td>5</td>
<td>village</td>
<td>Ispra</td>
<td>10</td>
</tr>
<tr>
<td>6</td>
<td>small settlement, hamlet</td>
<td>-</td>
<td>5</td>
</tr>
</tbody>
</table>

Place name recognition – Result

- List of place names found
- Frequency count per country (city, continent, region, …)
- Frequency can be normalised, using TF.IDF or similar
Agenda

- Motivation for cross-lingual document linking
  - EU: need for multilinguality
  - Enhanced Information Extraction by combining information from texts in various languages
  - Demo of NewsExplorer ([http://press.jrc.it/NewsExplorer](http://press.jrc.it/NewsExplorer))
- Related work
- Overview of our approach
- Description of the components, challenges
  - IE: Locations
  - **IE: Person and Organisation names**
  - Categorisation into Subject domains (Eurovoc classes)
  - Clustering of news
  - Linking clusters over time
  - Linking clusters across languages
- Future work and Conclusions

Multilingual name recognition and variant merging

<table>
<thead>
<tr>
<th>Language</th>
<th>Text</th>
</tr>
</thead>
<tbody>
<tr>
<td>en</td>
<td>death of former Prime Minister Rafik Hariri, blamed by many opposition</td>
</tr>
<tr>
<td>es</td>
<td>asesinato del exprimer ministro Rafic al-Hariri, que la oposición atribuyó</td>
</tr>
<tr>
<td>fr</td>
<td>l'assassinat de l'ex-dirigeant Rafic Hariri et le départ du chef de la diplom</td>
</tr>
<tr>
<td>nl</td>
<td>na de moord op oud-premier Rafiq al-Hariri gingen gisteren bijna een</td>
</tr>
<tr>
<td>de</td>
<td>libanesischen Regierungschef Rafik Hariri vor einem Monat wichtige B</td>
</tr>
<tr>
<td>sl</td>
<td>danjega libanonskega premiera Rafika Haririja. Libanonska opozicija si</td>
</tr>
<tr>
<td>et</td>
<td>möödumisele ekspeaminister Rafik al-Hariri surma põhjustanud pommip</td>
</tr>
<tr>
<td>ar</td>
<td>اغتيال رئيس الوزراء السابق رفيق الحريري بآيادي يهودية وما حدث سابقا</td>
</tr>
<tr>
<td>ru</td>
<td>Бывший премьер-министр Ливана Рафик Харири, который</td>
</tr>
</tbody>
</table>
Recognition of person names

- **Lookup of known names** from database
  - Currently over 560,000 names (excluding spelling variants)
  - ~1000 new names per day
  - Pre-generate morphological variants (Slovene example):
    \[ \text{Tony}(a|o|u|om|em|m|ljuljem|ja)\?s+\text{Blair}(a|o|u|om|em|m|ljuljem|ja) \]

- **Guessing names** using empirically-derived *lexical patterns*
  - Trigger word(s) + Name Surname
    - President, Minister, Head of State, Sir, American
    - “death of”, “[0-9]+-year-old”, …
    - Known first names (John, Jean, Hans, Giovanni, Johan, …)
    - …
    - Combinations: “56-year-old former prime minister Kurmanbek Bakiyev”

---

Person name recognition – dealing with inflection

- Recognition of person names, using regular expressions (Slovene example):
  - kandidat(a|u|om)?
  - legend(a|e|i|o)
  - milijarder(ja|ljuljem)?
  - predsednik(a|u|om|em)?
  - predsednina(a|e|i|o)
  - ministrnic(a|e|i|o)
  - sekretar(ja|ljuljem)?
  - diktator(ja|ljuljem)?
  - playboy(a|u|om|em)?

… verskega voditelja Moktade al Sada je z notranjim …
= Muqtada al-Sadr (ID=236)
Learning name variants

• For all new names found: apply *approximate name matching*
  • Based on sets of letter bigrams and letter trigrams
  • Merge two names if cosine similarity is > 70%

• Collect variants automatically from Wikipedia

• *Cross-script* name matching
  (Cyrillic, Greek, Arabic + Farsi, Hindi)

  Тони Блеър  =>  Tony Blair
  Ιγνάτ Αλάουι  =>  Iyad Allaoui

• For details, see Pouliquen et al. (Journal Corela, 12/2005)

### Person name recognition – Result

• Frequency list of normalised person names found
  (numerical person ID)

#### Related People
- Alexander Litvinenko (17)
- Mario Scaramella (11)
- John Reid (8)
- Vladimir Putin (8)
- Andrei Lugovoi (8)
- Anna Politkovskaya (3)
- Leonid Nevzlin (3)
- Boris Berezovsky (3)
- Tony Blair (2)
Agenda

- Motivation for cross-lingual document linking
  - EU: need for multilinguality
  - Enhanced Information Extraction by combining information from texts in various languages
  - Demo of NewsExplorer (http://press.jrc.it/NewsExplorer)
- Related work
- Overview of our approach
- Description of the components, challenges
  - IE: Locations
  - IE: Person and Organisation names
  - **Categorisation into Subject domains (Eurovoc classes)**
    - Clustering of news
    - Linking clusters over time
    - Linking clusters across languages
- Future work and Conclusions

Eurovoc Thesaurus
http://europa.eu.int/celex/eurovoc

- **Over 6000 classes**
- Covering many different subject domains (wide coverage)
- **Multilingual** (over 20 languages, one-to-one translations)
- Developed by the European Parliament and others
- Actively used to **manually index and retrieve documents** in large collections (fine-grained classification and cataloguing system)
- Freely available for research purposes
- **Hierarchically organised** into up to 8 levels

Relations:
- Broader Terms
- Narrower Terms
- Related Terms
Eurovoc categorisation – Major challenges

- Eurovoc is a conceptual thesaurus

  E.g.
  - SPORT
  - PROTECTION OF MINORITIES
  - CONSTRUCTION AND TOWN PLANNING
  - RADIOACTIVE MATERIALS

  → categorisation vs. term extraction

- Large number of classes (~ 6000)
- Very unevenly distributed
- Various text types (heterogeneous training set)
- Multi-label categorisation (both for training and assignment)

Eurovoc categorisation – Approach

- Profile-based, category ranking task
  - Training: Identification of most significant words for each class
  - Assignment: combination of measures to calculate similarity between profiles and new document

- Empirical refinement of parameter settings
  - Training:
    - Stop words
    - Lemmatisation
    - Multi-word terms
    - Consider number of classes of each training document
    - Thresholds for training document length and number of training documents per class
    - Methods to determine significant words per document (log-likelihood vs. chi-square, etc.)
    - Choice of reference corpus
    - …
  - Assignment:
    - Selection and combination of similarity measures (cosine, okapi, …)
    - …

For details, see Pouliquen et al. (Eurolan 2003)
**Assignment Result (Example)**

**Title:** Legislative resolution embodying Parliament's opinion on the proposal for a Council Regulation amending Regulation No 2847/93 establishing a control system applicable to the common fisheries policy (COM(95)0256 - C4-0272/95 - 95/0146(CNS)) (Consultation procedure)

### Results of automatic evaluation across languages (F1 per document at rank=6)

<table>
<thead>
<tr>
<th>Descriptor ID</th>
<th>Descriptor text</th>
<th>Inverse Square Tilt</th>
<th>Cosine Tilt</th>
<th>Rank</th>
<th>Okapli</th>
<th>Rank</th>
<th>Okapli</th>
<th>Rank</th>
<th>Prec</th>
<th>Rec</th>
</tr>
</thead>
<tbody>
<tr>
<td>564104076000000</td>
<td>FISHING CONTROLS</td>
<td>0.0014233</td>
<td>0.260</td>
<td>1</td>
<td>85.359</td>
<td>1</td>
<td>1</td>
<td>100</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>564102000000000</td>
<td>FISHING GROUNDS</td>
<td>0.0024864</td>
<td>0.309</td>
<td>2</td>
<td>65.028</td>
<td>14</td>
<td>2</td>
<td>100</td>
<td>20</td>
<td></td>
</tr>
<tr>
<td>564104020000000</td>
<td>COMMON FISHERIES POLICY</td>
<td>0.0018023</td>
<td>0.280</td>
<td>3</td>
<td>62.910</td>
<td>20</td>
<td>3</td>
<td>100</td>
<td>30</td>
<td></td>
</tr>
<tr>
<td>564104010000000</td>
<td>FISHERY MANAGEMENT</td>
<td>0.0010708</td>
<td>0.279</td>
<td>4</td>
<td>79.352</td>
<td>6</td>
<td>4</td>
<td>100</td>
<td>40</td>
<td></td>
</tr>
<tr>
<td>564104070000000</td>
<td>FISHING REGULATIONS</td>
<td>0.0019793</td>
<td>0.279</td>
<td>5</td>
<td>79.982</td>
<td>6</td>
<td>5</td>
<td>100</td>
<td>50</td>
<td></td>
</tr>
<tr>
<td>564104070400000</td>
<td>FISHING PERMIT</td>
<td>0.0030651</td>
<td>0.251</td>
<td>6</td>
<td>71.577</td>
<td>6</td>
<td>6</td>
<td>100</td>
<td>60</td>
<td></td>
</tr>
<tr>
<td>564104010100000</td>
<td>CONSERVATION OF FISH STOCKS</td>
<td>0.0012918</td>
<td>0.253</td>
<td>7</td>
<td>83.932</td>
<td>7</td>
<td>7</td>
<td>85</td>
<td>60</td>
<td></td>
</tr>
<tr>
<td>564104060000000</td>
<td>FISHING AREA</td>
<td>0.0013247</td>
<td>0.252</td>
<td>8</td>
<td>84.178</td>
<td>8</td>
<td>8</td>
<td>75</td>
<td>60</td>
<td></td>
</tr>
<tr>
<td>562054010000000</td>
<td>CONSERVATION OF RESOURCES</td>
<td>0.0023420</td>
<td>0.251</td>
<td>9</td>
<td>65.311</td>
<td>9</td>
<td>9</td>
<td>66</td>
<td>60</td>
<td></td>
</tr>
<tr>
<td>564105000000000</td>
<td>FISHERY RESOURCES</td>
<td>0.0042283</td>
<td>0.323</td>
<td>10</td>
<td>75.046</td>
<td>10</td>
<td>10</td>
<td>68</td>
<td>60</td>
<td></td>
</tr>
<tr>
<td>564104000000000</td>
<td>CATCH OF FISH</td>
<td>0.0031310</td>
<td>0.213</td>
<td>11</td>
<td>67.587</td>
<td>11</td>
<td>11</td>
<td>54</td>
<td>60</td>
<td></td>
</tr>
<tr>
<td>564104000000000</td>
<td>FISHERIES POLICY</td>
<td>0.0023839</td>
<td>0.203</td>
<td>12</td>
<td>68.416</td>
<td>12</td>
<td>12</td>
<td>50</td>
<td>60</td>
<td></td>
</tr>
<tr>
<td>564104070500000</td>
<td>FISHING LICENCE</td>
<td>0.0027113</td>
<td>0.181</td>
<td>13</td>
<td>57.618</td>
<td>13</td>
<td>13</td>
<td>46</td>
<td>60</td>
<td></td>
</tr>
<tr>
<td>564106010000000</td>
<td>FISHING FLEET</td>
<td>0.0010047</td>
<td>0.179</td>
<td>14</td>
<td>63.323</td>
<td>14</td>
<td>14</td>
<td>42</td>
<td>60</td>
<td></td>
</tr>
<tr>
<td>564101000000000</td>
<td>FISHING INDUSTRY</td>
<td>0.0051253</td>
<td>0.176</td>
<td>15</td>
<td>39.228</td>
<td>15</td>
<td>15</td>
<td>40</td>
<td>60</td>
<td></td>
</tr>
<tr>
<td>564104020100000</td>
<td>EUROPEFISH</td>
<td>0.0073832</td>
<td>0.176</td>
<td>16</td>
<td>62.240</td>
<td>16</td>
<td>16</td>
<td>37</td>
<td>60</td>
<td></td>
</tr>
</tbody>
</table>

---

**Results of automatic evaluation across languages**

- **Human evaluation**
  - English: 83%
  - Spanish: 80%

---

**Diagram:**

- **Without pre-processing**
- **With pre-processing (French=> only stop words)**

---

**MLIA'2006, Slide 23**
Eurovoc indexing – Result

• Ranked list of Eurovoc descriptor codes found for each document

<table>
<thead>
<tr>
<th>Descriptor ID</th>
<th>Cosine</th>
</tr>
</thead>
<tbody>
<tr>
<td>156410407060000000</td>
<td>0.360</td>
</tr>
<tr>
<td>156410402000000000</td>
<td>0.308</td>
</tr>
<tr>
<td>156410402000000000</td>
<td>0.280</td>
</tr>
<tr>
<td>156410401000000000</td>
<td>0.279</td>
</tr>
<tr>
<td>156410407000000000</td>
<td>0.270</td>
</tr>
<tr>
<td>156410407040000000</td>
<td>0.261</td>
</tr>
<tr>
<td>156410401010000000</td>
<td>0.253</td>
</tr>
<tr>
<td>156410406000000000</td>
<td>0.252</td>
</tr>
<tr>
<td>156200401000000000</td>
<td>0.251</td>
</tr>
<tr>
<td>156410500000000000</td>
<td>0.232</td>
</tr>
<tr>
<td>156410408000000000</td>
<td>0.219</td>
</tr>
<tr>
<td>156410400000000000</td>
<td>0.203</td>
</tr>
<tr>
<td>156410407050000000</td>
<td>0.181</td>
</tr>
<tr>
<td>156410400100000000</td>
<td>0.179</td>
</tr>
<tr>
<td>156410100000000000</td>
<td>0.176</td>
</tr>
<tr>
<td>156410402010000000</td>
<td>0.176</td>
</tr>
</tbody>
</table>

The ‘JRC-Acquis’ parallel corpus in 21 languages

• Freely available for research purposes on our web site: http://langtech.jrc.it/JRC-Acquis.html
• For details, see Steinberger et al. (2006, LREC)

• Average of 8.8 Million words per language
• Pair-wise alignment for all 210 language pairs!
• Average of 7600 documents per language
• Most documents have been Eurovoc-classified manually

用途包括:
1. 培训多语言主题领域分类器。
2. 创建多语言词汇空间（LSA, KCCA）
3. 培训自动系统用于统计机器翻译。
4. 制作多语言词汇或语义资源，如词典或概念网。
5. 培训和测试多语言信息提取软件。
6. 自动翻译一致性检查。
7. 测试和基准测试对齐软件（句子、单词等），涵盖更大范围的语言对。
8. 所有类型的多语言和跨语言研究。
Agenda

- Motivation for cross-lingual document linking
  - EU: need for multilinguality
  - Enhanced Information Extraction by combining information from texts in various languages
  - Demo of NewsExplorer (http://press.jrc.it/NewsExplorer)
- Related work
- Overview of our approach
- Description of the components, challenges
  - IE: Locations
  - IE: Person and Organisation names
  - Categorisation into Subject domains (Eurovoc classes)
  - Clustering of news
  - Linking clusters over time
  - Linking clusters across languages
- Future work and Conclusions

Monolingual document representation

- Vector of keywords and their keyness using log-likelihood test (Dunning 1993)

---

Michael Jackson Jury Reaches Verdicts

Monday, June 13, 2005

Jackson, 50, was accused of molesting the then-13-year-old boy and giving him VIP status at his ranch in 2003. Jackson had befriended the boy, a cancer survivor, and they appeared together when Jackson was interviewed for the documentary "Leaving Neverland."  

```
Keyness Keyword   Keyness Keyword
109.24 jackson    9.39   verdict
41.54 neverland   7.56   testimony
37.93 santa       6.50   maria
32.61 molestation 4.09   michael
24.51 boy         1.73   reached
24.43 pop         1.68   ap
20.68 documentary 1.05   appeared
18.79 accuser     0.53   child
13.59 courthouse  0.50   trial
11.12 jury        0.45   monday
10.08 ranch       0.26   children
9.60 california   0.09   family
```

`````````````````````````````````
Calculation of a text’s Country Score

- Aim: show to what extent a text talks about a certain country
- Sum of references to a country, normalised using the log-likelihood test
- Add country score vector to keyword vector

<table>
<thead>
<tr>
<th>Keyness</th>
<th>Keyword</th>
</tr>
</thead>
<tbody>
<tr>
<td>109.2478</td>
<td>jackson</td>
</tr>
<tr>
<td>41.5450</td>
<td>neverland</td>
</tr>
<tr>
<td>37.9347</td>
<td>santa</td>
</tr>
<tr>
<td>32.6105</td>
<td>molestation</td>
</tr>
<tr>
<td>24.5193</td>
<td>boy</td>
</tr>
<tr>
<td>24.4351</td>
<td>pop</td>
</tr>
<tr>
<td>20.6824</td>
<td>documentary</td>
</tr>
<tr>
<td>18.7973</td>
<td>accuser</td>
</tr>
<tr>
<td>13.5945</td>
<td>courthouse</td>
</tr>
<tr>
<td>11.1224</td>
<td>jury</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Keyness</th>
<th>Keyword</th>
</tr>
</thead>
<tbody>
<tr>
<td>7.5620</td>
<td>testimony</td>
</tr>
<tr>
<td>6.5014</td>
<td>maria</td>
</tr>
<tr>
<td>4.0957</td>
<td>michael</td>
</tr>
<tr>
<td>1.7368</td>
<td>reached</td>
</tr>
<tr>
<td>1.6857</td>
<td>ap</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Keyness</th>
<th>Keyword</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.5610</td>
<td><em>gb</em></td>
</tr>
<tr>
<td>1.5610</td>
<td><em>ii</em></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Keyness</th>
<th>Keyword</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.5610</td>
<td><em>br</em></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Keyness</th>
<th>Keyword</th>
</tr>
</thead>
<tbody>
<tr>
<td>10.0838</td>
<td>ranch</td>
</tr>
<tr>
<td>9.6021</td>
<td>california</td>
</tr>
<tr>
<td>9.3905</td>
<td>verdict</td>
</tr>
</tbody>
</table>

Monolingual news clustering

- Input: Vectors consisting of keywords and country score
- Similarity measure: cosine
- Method: Bottom-up group average unsupervised clustering
- Build the binary hierarchical clustering tree (dendrogram)
  - Retain only “big” nodes in the tree with a high cohesion (empirically refined minimum intra-node similarity: 45%)
- Use the title of the cluster’s medoid as the cluster title
- For details, see Pouliquen et al. (CoLing 2004)
Monolingual cluster linking - Evaluation

- Details: Pouliquen et al. (CoLing 2004)
- Evaluation results depending on similarity threshold

<table>
<thead>
<tr>
<th>Similarity threshold</th>
<th>“Precision”</th>
<th>“Recall”</th>
</tr>
</thead>
<tbody>
<tr>
<td>15%</td>
<td>88%</td>
<td>100%</td>
</tr>
<tr>
<td>20%</td>
<td>92%</td>
<td>98%</td>
</tr>
<tr>
<td>40%</td>
<td>98%</td>
<td>86%</td>
</tr>
<tr>
<td>60%</td>
<td>99%</td>
<td>78%</td>
</tr>
<tr>
<td>80%</td>
<td>99%</td>
<td>67%</td>
</tr>
</tbody>
</table>

Agenda

- Motivation for cross-lingual document linking
  - EU: need for multilinguality
  - Enhanced Information Extraction by combining information from texts in various languages
  - Demo of NewsExplorer (http://press.jrc.it/NewsExplorer)
- Related work
- Overview of our approach
- Description of the components, challenges
  - IE: Locations
  - IE: Person and Organisation names
  - Categorisation into Subject domains (Eurovoc classes)
  - Clustering of news
  - Linking clusters over time
  - Linking clusters across languages
- Future work and Conclusions
Cross-lingual cluster linking – combination of 4 ingredients

\[ \text{CLDS} = \alpha \cdot S_1 + \beta \cdot S_2 + \gamma \cdot S_3 + \delta \cdot S_4 \]

- Ranked list of Eurovoc classes (40%)
- Country score (30%)
- Names + frequency (20%)
- Monolingual cluster representation \textit{without country score} (10%)

Cross-lingual cluster linking – evaluation

- Evaluation results depending on similarity threshold
- Ingredients: 40/30/30 (names not yet considered)
- Evaluation for EN \rightarrow FR and EN \rightarrow IT (136 EN clusters)

<table>
<thead>
<tr>
<th>Similarity threshold</th>
<th>EN \rightarrow FR Precision</th>
<th>EN \rightarrow FR Recall *</th>
<th>EN \rightarrow IT Precision</th>
<th>EN \rightarrow IT Recall *</th>
</tr>
</thead>
<tbody>
<tr>
<td>30%</td>
<td>84%</td>
<td>99%</td>
<td>71%</td>
<td>97%</td>
</tr>
<tr>
<td>60%</td>
<td>98%</td>
<td>46%</td>
<td>98%</td>
<td>42%</td>
</tr>
</tbody>
</table>

* Recall at 15% similarity threshold = 100%

For details, see Pouliquen et al. (CoLing 2004)
Filter out bad links by exploiting all cross-lingual links

Assumption:
If EN is linked to FR, ES, IT, …
FR should also be linked to ES, IT, …
If not: lower link likelihood

Agenda

• Motivation for cross-lingual document linking
  • EU: need for multilinguality
  • Enhanced Information Extraction by combining information from texts in various languages
  • Demo of NewsExplorer (http://press.jrc.it/NewsExplorer)

• Related work

• Overview of our approach

• Description of the components, challenges
  • IE: Locations
  • IE: Person and Organisation names
  • Categorisation into Subject domains (Eurovoc classes)
  • Clustering of news
  • Linking clusters over time
  • Linking clusters across languages

• Future work and Conclusions
Planned work (1)

1. Add more languages (analysis and cross-lingual linking)

2. Add more information facets for cross-lingual cluster linking
   - Dates
   - Professions
   - Expressions of measurement
   - Vehicles
   - ...

3. Empirically optimise weighting of ingredients
   \[ \text{CLDS} = \alpha \cdot S_1 + \beta \cdot S_2 + \gamma \cdot S_3 + \delta \cdot S_4 + \ldots \]

Planned work (2)

4. Link longer-running stories to other languages (currently, only individual clusters are linked)

5. Categorise persons (governmental, military, religious, civil, …)

6. Categorise relations between persons
   - Family relation
   - Contact/Meeting information
   - ORG-Membership relation
   - ...
   - …
Conclusion

• Cross-lingual linking of documents/clusters via language-independent representations is feasible.

• More information can be extracted when using texts written in several languages.

• Simple means can take you a long way
  • JRC effort to add a new language is 1 - 6 months
    • Simple lexical patterns
    • Simple morphological suffix treatment
  • Clustering and categorisation
  • Language-independent statistics and heuristics.

• Monolingual analysis is sufficient; no language pair-specific information is needed.