Story tracking: linking similar news over time and across languages

CoLing’2008 Workshop MMIES:
Multi-source Multilingual Information Extraction and Visualisation
Manchester, UK, 23 August 2008

Bruno Pouliquen, Olivier Deguernel, Ralf Steinberger
& the JRC's Web Mining and Intelligence team

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

---

Agenda

- Introduction; motivation; news data
- Monolingual news clustering; linking clusters over time
- Building longer monolingual ‘stories’
- Linking daily clusters across languages
- Linking longer stories across languages
- Conclusion and future work
  • Topic example: Oklahoma City bombing in 1995 incl. memorial services, investigations, prosecution, etc.

• \textit{Topic} \neq \textit{category} (bombing)!

• To our knowledge, there is no other publicly or commercially available TDT system

---

Europe Media Monitor (EMM) – Public systems

http://press.jrc.it/overview.html
EMM news sources

- The EMM engine gathers approximately **50,000 news articles per day**
- In **over 40 languages**
- From **over 1,500 news portals** world-wide, plus 20 news wires

→ Need for aggregation, summarisation and visualisation
→ TDT allows users to follow a specific event or story over time

- NewsExplorer does this for 19 languages:
  - Cluster daily news in each language
  - Extract and display information on persons, organisations, locations
  - Link daily clusters over time and across languages
  - Aggregate into longer *stories*

Agenda

- Introduction; motivation; news data

- **Monolingual news clustering; linking clusters over time**
  - Building longer monolingual ‘stories’
  - Linking daily clusters across languages
  - Linking longer stories across languages

- Conclusion and future work
Clustering: Monolingual document representation

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

<table>
<thead>
<tr>
<th>Keyness Keyword</th>
<th>Keyness Keyword</th>
</tr>
</thead>
<tbody>
<tr>
<td>109.24</td>
<td>jackson</td>
</tr>
<tr>
<td>41.54</td>
<td>neverland</td>
</tr>
<tr>
<td>37.93</td>
<td>santa</td>
</tr>
<tr>
<td>32.61</td>
<td>molestation</td>
</tr>
<tr>
<td>24.51</td>
<td>boy</td>
</tr>
<tr>
<td>24.43</td>
<td>pop</td>
</tr>
<tr>
<td>20.68</td>
<td>documentary</td>
</tr>
<tr>
<td>18.79</td>
<td>accuser</td>
</tr>
<tr>
<td>13.59</td>
<td>courthouse</td>
</tr>
<tr>
<td>11.12</td>
<td>jury</td>
</tr>
<tr>
<td>10.08</td>
<td>ranch</td>
</tr>
<tr>
<td>9.60</td>
<td>california</td>
</tr>
</tbody>
</table>

- Enhanced with country information
  - Sum of references to a country, normalised using the log-likelihood test

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 minimum intra-node similarity of 45%

Use the title of the cluster’s medoid as the cluster title
Link to clusters of previous days if
  - Link within 7 days
  - Cosine cluster similarity > 0.5
• Link to clusters of previous days if
  • Link within 7 days
  • Cosine cluster similarity > 0.5
• 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

• Introduction; motivation; news data
• Monolingual news clustering; linking clusters over time
• Building longer monolingual ‘stories’
  • Linking daily clusters across languages
  • Linking longer stories across languages
• Conclusion and future work
Live demo: browsing a ‘story’

Linking daily clusters into longer ‘stories’

Score[s] += (1 - score[s] \cdot \text{sim}(c,s))
Linking clusters into stories – Evaluation

- Evaluate relatedness of clusters with story
- Test set: 330 randomly selected stories in four languages

<table>
<thead>
<tr>
<th>Language</th>
<th>Number of stories</th>
<th>Correct components</th>
<th>All components</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>German</td>
<td>93</td>
<td>249</td>
<td>265</td>
<td>0.94</td>
</tr>
<tr>
<td>English</td>
<td>113</td>
<td>490</td>
<td>570</td>
<td>0.86</td>
</tr>
<tr>
<td>Spanish</td>
<td>33</td>
<td>78</td>
<td>91</td>
<td>0.86</td>
</tr>
<tr>
<td>Italian</td>
<td>91</td>
<td>239</td>
<td>299</td>
<td>0.80</td>
</tr>
<tr>
<td>All</td>
<td>330</td>
<td>1056</td>
<td>1225</td>
<td>0.86</td>
</tr>
</tbody>
</table>

Collection of meta-information about each story

Iran welcomes US downgrading of nuclear threat

- Stories consist of time-linked news clusters with overlapping keywords.
  - **Keywords:** Iran, United States, Iraq / UN Security Council, Islamic Republic / iranian, tehran, nuclear, sanctions, program, enrichment, uranium
  - **Importance:** 11586 articles in 596 clusters
  - **Start date:** Saturday, December 2, 2006  **End date:** Monday, July 21, 2008

- For each story, we collect:
  - Total number of articles and clusters
  - Start and end date
  - Title of the first cluster
  - Title of the biggest cluster
  - Most frequently mentioned person names ('related people')
  - Person names most highly associated with the story ('associated people')
  - Most frequently mentioned ‘other names’ (mostly organisations and events)
  - Most frequently referred to countries
  - List of keywords
Agenda

• Introduction; motivation; news data
• Monolingual news clustering; linking clusters over time
• Building longer monolingual ‘stories’
• Linking daily clusters across languages
• Linking longer stories across languages
• Conclusion and future work

Live demo: cross-lingual cluster linking
Approaches to cross-lingual document similarity calculation (1)

• How to find out whether two texts in different languages are related?

• Most common approach: use MT or bilingual dictionaries to translate into English, then use monolingual methods to calculate similarity.
  • Using MT (e.g. Leek et al. 1999 for Chinese-Mandarin to English);
  • Using bilingual dictionaries (e.g. Wactlar 1999 for Serbo-Croatian to English; Urizar & Loinaz for Basque, Spanish and English 2007)
  • Automatically produce bilingual word associations for bilingual document representation and document similarity calculation, e.g.
    • Bilingual Lexical Semantic Analysis (LSA) (Landauer & Littman 1991)
    • Kernel Canonical Correlation Analysis (KCCA) (Vinokourov et al., 2002)

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

\[
\text{Language pairs} = \frac{N \times (N-1)}{2} \quad \text{(N = number of languages)}
\]

19 NewsExplorer languages \rightarrow 171 language pairs (342 language pair directions)!

Approaches to cross-lingual document similarity calculation (2)

• Alternative: use entities as anchors:
  • Names of persons and organisations
  • Names of locations
  • Units of measurements:
    • Time
    • Speed
    • Temperature
    • Acceleration
  • Multilingual specialist dictionaries (Eurovoc for public administration, MeSH for medicine, etc.)
  • …

• Normalise these expressions

→ Use as kind of an interlingua; no language pair-specific resource needed
Language-independent features for multilingual document representation

No MT or bilingual dictionaries

19 languages

\[ \text{CLS} = \alpha \cdot S1 + \beta \cdot S2 + \gamma \cdot S3 + \delta \cdot S4 \]

- **Sim1 (\( \alpha = 40\% \))**: Multilingual Eurovoc subject domains
- **Sim2 (\( \beta = 30\% \))**: Geo-locations
- **Sim3 (\( \gamma = 20\% \))**: Names + variants
- **Sim4 (\( \delta = 10\% \))**: Cognates and numbers (without country score)

Normalisation of entities and categories

**Name variants**

Frequency list of numerical person identifiers

<table>
<thead>
<tr>
<th>ID: 81705</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ali Larijani (Eu, en)</td>
</tr>
<tr>
<td>Ali Larijani (de)</td>
</tr>
<tr>
<td>Ali Larijani (es)</td>
</tr>
<tr>
<td>Ali Larijani (si)</td>
</tr>
<tr>
<td>Ali Larijani (ar)</td>
</tr>
<tr>
<td>Ali Larijani (es)</td>
</tr>
<tr>
<td>Ali Larijani (en)</td>
</tr>
<tr>
<td>Ali Larijani (es)</td>
</tr>
<tr>
<td>Ali Larijani (it)</td>
</tr>
<tr>
<td>Ali Larijani (fr)</td>
</tr>
<tr>
<td>Ali Larijani (vi)</td>
</tr>
<tr>
<td>Ali Larijani (nl)</td>
</tr>
<tr>
<td>Ali Larijani (nl)</td>
</tr>
<tr>
<td>Ali Larijani (en)</td>
</tr>
<tr>
<td>Ali A. Larijani (en)</td>
</tr>
<tr>
<td>Ali Larijani (it)</td>
</tr>
<tr>
<td>Ali Larijani (es)</td>
</tr>
<tr>
<td>Ali Larijani (en)</td>
</tr>
</tbody>
</table>

**Geo-locations**

Frequency list of ISO codes:

- US
- FR
- DE
- ES
...

**Subject categories**

Weighted list of Eurovoc codes:

1. **5641040706000000** - Fishing Controls
2. **5641020000000000** - Fishing Grounds
3. **5641040200000000** - Common Fisheries Policy
4. **5641040100000000** - Fishery Management
5. **5641040700000000** - Fishing Regulations
6. **5641040704000000** - Fishing Permit
7. **5641040101000000** - Conservation of Fish Stocks
8. **5641040600000000** - Fishing Area
9. **5641040601000000** - Conservation of Resources

Cross-lingual cluster linking – evaluation

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

<table>
<thead>
<tr>
<th>Similarity threshold</th>
<th>EN → FR Precision</th>
<th>EN → FR Recall *</th>
<th>EN → IT Precision</th>
<th>EN → 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%

Assumption:
If EN is linked to FR, ES, IT, ...
FR should also be linked to ES, IT, ...
If not: lower link likelihood
Filter out bad links by exploiting all cross-lingual links

- Build a second similarity, based on the first. It uses the following input:
  - 1) the number of links between the set of clusters in the other languages;
  - 2) the strength (or similarity level) of these links;
  - 3) the number of potential links between the set of clusters in the other languages (which means all the links minus those between clusters in the same language)

- Empirical formula:
  - similarity_2 = similarity_1 * 
    (number_of_links / number_of_potential_links) * 
    square_root(number_of_potential_links)

- Result: elimination of some wrong links
- (No formal evaluation results available)

Agenda

- Introduction; motivation; news data
- Monolingual news clustering; linking clusters over time
- Building longer monolingual ‘stories’
- Linking daily clusters across languages
- Linking longer stories across languages
- Conclusion and future work
Cross-lingual linking of whole stories (ongoing work)

- Link whole stories by adding the daily cross-lingual similarity values (if $\geq 0.15$)
- Test set: those 112 of the 330 stories that have cross-lingual links into any language.
- Results improve with story size

<table>
<thead>
<tr>
<th>Type of story</th>
<th>Number of stories</th>
<th>Nb of correct cross-lingual links</th>
<th>Number of cross-lingual links</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>All stories</td>
<td>112</td>
<td>275</td>
<td>465</td>
<td>0.59</td>
</tr>
<tr>
<td>Stories containing at least 5 clusters</td>
<td>39</td>
<td>145</td>
<td>232</td>
<td>0.62</td>
</tr>
<tr>
<td>Stories containing at least 10 clusters</td>
<td>11</td>
<td>75</td>
<td>100</td>
<td>0.75</td>
</tr>
<tr>
<td>10 top stories in 4 languages</td>
<td>40</td>
<td>235</td>
<td>270</td>
<td>0.87</td>
</tr>
</tbody>
</table>

Agenda

- Introduction; motivation; news data
- Monolingual news clustering; linking clusters over time
- Building longer monolingual ‘stories’
- Linking daily clusters across languages
- Linking longer stories across languages
- Future work and conclusion
Future work

- Extend 7-day window to allow for low-reporting periods
- Deal with splitting or merging stories (difficult)
- Current story linking assumes same-day reporting
- Improve accuracy for cross-lingual story linking, e.g.
  - by additionally making use of accumulated meta-information
- Search interface specifically for stories
- Integration with more short-term live news clustering in EMM NewsBrief

Conclusion

- Nobody can read and digest all articles → aggregation is definitely helpful.
- To our knowledge, there are no commercial or other publicly accessible alternatives
- Logs show that ~ 400 users per day look at story-related information
- 86% Precision for monolingual story building → could be improved, but is already useful
- 59% Precision for cross-lingual story linking → this is too low to be used in the live system → can currently only be used for larger stories (P=87%)