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
• Formally defined in US-American DARPA program TDT (1997-2004).
  • Topic example: Oklahoma City bombing in 1995 incl. memorial services, investigations, prosecution, etc.

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

| Keyness Keyword | Keyness Keyword |
|-----------------|-----------------|
| 109.24          | jackson         |
| 41.54           | neverland       |
| 37.93           | santa           |
| 32.61           | molestation     |
| 24.51           | boy             |
| 24.43           | pop             |
| 20.68           | documentary     |
| 18.79           | accuser         |
| 13.59           | courthouse      |
| 11.12           | jury            |
| 10.06           | ranch           |
| 9.60            | california      |

- Vector of keywords and their keyness using log-likelihood test (Dunning 1993)
  - Enhanced with country information
    - Sum of references to a country, normalised using the log-likelihood test

### Monday, June 13, 2005

**Michael Jackson Jury Reaches Verdicts**

Jackson, 46, was accused of molesting the then-13-year-old boy and plying him with wine at the pop star’s Neverland ranch in 2003. Jackson had befriended the boy, a cancer survivor, and they appeared together when Jackson was interviewed for the documentary “Living With Michael Jackson.”

*ABCnews 19/06/2005 21:54*

Original cluster

### Multi-monolinguual news clustering – linking over time

- **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
Monolingual cluster linking + Evaluation

- Link to clusters of previous days if
  - Link within 7 days
  - Cosine cluster similarity > 0.5
- Evaluation results depending on similarity threshold

| Similarity threshold | Precision | Recall |
|----------------------|-----------|--------|
| 15%                  | 88%       | 100%   |
| 20%                  | 92%       | 98%    |
| 40%                  | 98%       | 86%    |
| 60%                  | 99%       | 78%    |
| 80%                  | 99%       | 67%    |

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] * sim(c,s))

Minimum cluster link strength: 0.15
Minimum story link strength: 0.5

If multiple stories match, assign to higher-scoring story
Linking clusters into stories – Evaluation

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

| Language | Number of stories | Correct components | All components | Precision |
|----------|-------------------|--------------------|----------------|-----------|
| German   | 93                | 249                | 265            | 0.94      |
| English  | 113               | 490                | 570            | 0.86      |
| Spanish  | 33                | 78                 | 91             | 0.86      |
| Italian  | 91                | 239                | 299            | 0.80      |
| All      | 330               | 1056               | 1225           | 0.86      |

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
      Language pairs = N * (N-1) / 2 (N = number of languages)
      19 NewsExplorer languages → 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

CLDS = α·S1 + β·S2 + γ·S3 + δ·S4

Sim1 (α = 40%): Multilingual Eurovoc subject domains

Sim2 (β = 30%): Geo-locations

Sim3 (γ = 20%): Names + variants

Sim4 (δ = 10%): Cognates and numbers (without country score)

Normalisation of entities and categories

Name variants

Geo-locations

Subject categories

Frequency list of numerical person identifiers

Frequency list of ISO codes:

Weighted list of Eurovoc codes:

ID: 81705
**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)

| Similarity threshold | EN → FR Precision | EN → FR Recall * | EN → IT Precision | EN → IT Recall * |
|----------------------|-------------------|------------------|------------------|------------------|
| 30%                  | 84%               | 99%              | 71%              | 97%              |
| 60%                  | 98%               | 46%              | 98%              | 42%              |

* 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:**
  \[
  \text{similarity}_2 = \text{similarity}_1 \times \frac{\text{number of links}}{\text{number of potential links}} \times \sqrt{\text{number of potential links}}
  \]

- 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

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

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%)