

NewsExplorer – Multilingual News Analysis with Cross-lingual Linking

Ralf Steinberger, Bruno Pouliquen, Camelia Ignat

European Commission – Joint Research Centre

Via E. Fermi 1, T.P. 267, 21020 Ispra (VA), Italy,

<http://langtech.jrc.it/>, <http://press.jrc.it/NewsExplorer/> – Firstname.Lastname@jrc.it

Motivation

The language barrier is a serious handicap in our globalised world as it limits the access to available information and slows the communication process. In Europe, where the European Union (EU) alone has 20 official languages, this is an even larger problem than in the USA.

Human translation clearly is not a solution for the vast amount of information available on the internet or in other settings. Machine Translation is only available for the major language pairs. Furthermore, it is by definition a bilingual application and is thus limited in its usefulness in highly multilingual settings such as the EU with its 190 language pairs. Approaches using multilingual vector space to represent documents or text segments, as proposed by Landauer & Littman (1989) using Latent Semantic Analysis (LSA) or Vinokourov et al. (2002) using Kernel Canonical Correlation Analysis (KCCA) are currently limited to bilingual representations, as well. Extending KCCA to more than two languages is the subject of current work carried out as part of the European research network PASCAL. All currently available solutions are bilingual.

We are presenting here an alternative, more interlingua-like working solution, to link news clusters for all language pair combinations in currently eight languages as part of the daily news analysis in NewsExplorer (<http://press.jrc.it/NewsExplorer>). The approach consists of producing a language-independent representation, based mostly on subject domains and normalised named entities, and to apply a similarity measure to this interlingua representation.

In NewsExplorer, all news articles of the same language are clustered once a day to group related articles from many different online sources. Each cluster of news articles, whose size can vary from two to two hundred articles per day, represents one story or topic and is henceforth treated as one meta-document. Where applicable, this cluster is automatically linked to the related clusters in the other languages.

Approach

Each monolingual NewsExplorer cluster is represented by four different vectors that serve as input to the cross-lingual similarity calculation, each with a different input weight, according to the formula:

$$(4A + 3B + 2C + 1D)/10$$

where A, B, C and D are the cosine similarity values

for the vectors based on the following representations (see Figure 1 for an example):

- (A) EUROVOC (Eurovoc, 1995) subject domain classes (40%);
- (B) Country score (30%), where each mention of the country or of a place in the country add to a counter, which is then normalised by the average country score of that country using the log-likelihood formula (Dunning 1993);
- (C) Person and organisation names (20%): this simply is a frequency list of the person and organisation identifiers, as found in the cluster;
- (D) List of monolingual words found in the cluster and their log-likelihood value (10%), where the cluster word frequency is compared to the average frequency in our NewsExplorer news corpus.

Each of the four ingredients thus has a different weight towards the overall cross-lingual cluster similarity. The relative weight has been derived by intuition. It is planned to identify more exact values using empirical methods.

Challenges

While the similarity calculation is straight-forward, the success of the method is dependant on the quality of the subject domain classification and on the identification, disambiguation and normalisation of the named entities: (A) The multi-label categorisation system using the wide-coverage EUROVOC thesaurus with its over 6000 hierarchically organised classes, had to be trained on highly unbalanced, heterogeneous, manually multi-label classified data (Pouliquen et al. 2003). (B) Geo-locations can be represented unambiguously by their latitude-longitude information, but to arrive there, disambiguating is necessary to avoid wrong hits for places like ‘And’ in Iran, ‘Kofi’

| | |
|--------------------|---------|
| Nicolas Sarkozy | (Eu,tl) |
| Nicolás Sarkozy | (es) |
| Nicholas Sarkozy | (de,it) |
| Nicolas Sar | (en,fr) |
| Nicolas - Sarkozy | (fr) |
| Nicolasa Sarkozyja | (sl) |
| Nicolás Sarkozi | (es) |
| Nicolas Sarzoky | (es) |
| Nicolas Sarcozy | (fr) |
| Nicolas Sartkozy | (fr) |

Table 1. Some of the name variants for French Interior Minister *Nicolas Sarkozy* found in multilingual news texts. All variants are linked to the same numerical name identifier through approximate matching techniques.

in Mali, ‘Annan’ in Scotland, and for the 32 homographic places world-wide called ‘Washington’ and the 15 places called ‘Berlin’ (Pouliquen et al. 2006). (C) For the mapping to be successful, the many different spelling variants found in real-life text for person names and other named entities need to be mapped to the same name identifier (Table 1; for some persons, the NewsExplorer database has automatically collected over one hundred spelling variants). For inflected languages, the base form of each name additionally needs to be identified automatically (e.g. Slovene ‘Tonyjem Blairom’ for ‘Tony Blair’) with methods that are fast, simple and cheap enough to be applied to many different languages (Pouliquen et al. 2005). Component (D) contributes by capturing cognates, names and other strings.

The presentation at the workshop will give an overview of the cross-lingual linking approach, as well as of the proposed solutions to the individual challenges mentioned here.

Acknowledgements

We are grateful to the JRC’s Web Technology team for providing us with the *Europe Media Monitor* news data and for making our analysis results available to the wider public on a highly visited web site (over 600,000 hits per day in October 2006).

References

Dunning T. (1993). *Accurate Methods for the Statistics of Surprise and Coincidence*. Computational Linguistics, Volume 19, number 1, pp. 61-74. 1993.

Eurovoc (1995). *Thesaurus Eurovoc - Volume 2: Subject-Oriented Version. Ed. 3/English Language*.
 Landauer Thomas & Michael Littman (1991). *A Statistical Method for Language-Independent Representation of the Topical Content of Text Segments*. 11th International Conference *Expert Systems and Their Applications*, vol. 8: 77-85. Avignon, France.
 Pouliquen Bruno, Ralf Steinberger & Camelia Ignat (2003). *Automatic Annotation of Multilingual Text Collections with a Conceptual Thesaurus*. Workshop *Ontologies and Information Extraction* at EUROLAN'2003.
 Pouliquen Bruno, Ralf Steinberger, Camelia Ignat, Irina Temnikova, Anna Widiger, Wajdi Zaghouani & Jan Žižka (2005). *Multilingual person name recognition and transliteration*. Journal CORELA. Numéros spéciaux, *Le traitement lexicographique des noms propres*.
 Pouliquen Bruno, Marco Kimler, Ralf Steinberger, Camelia Ignat, Tamara Oellinger, Ken Blackler, Flavio Fuat, Wajdi Zaghouani, Anna Widiger, Ann-Charlotte Forslund, Clive Best (2006). *Geocoding multilingual texts: Recognition, Disambiguation and Visualisation*. LREC'2006.
 Steinberger Ralf, Pouliquen Bruno & Camelia Ignat (2004). *Providing cross-lingual information access with knowledge-poor methods*. *Informatica* 28-4.
 Vinokourov, A., Shawe-Taylor, J., Cristianini, N. (2002). *Inferring a semantic representation of text via cross-language correlation analysis*. Advances of Neural Information Processing Systems 15, 2002.

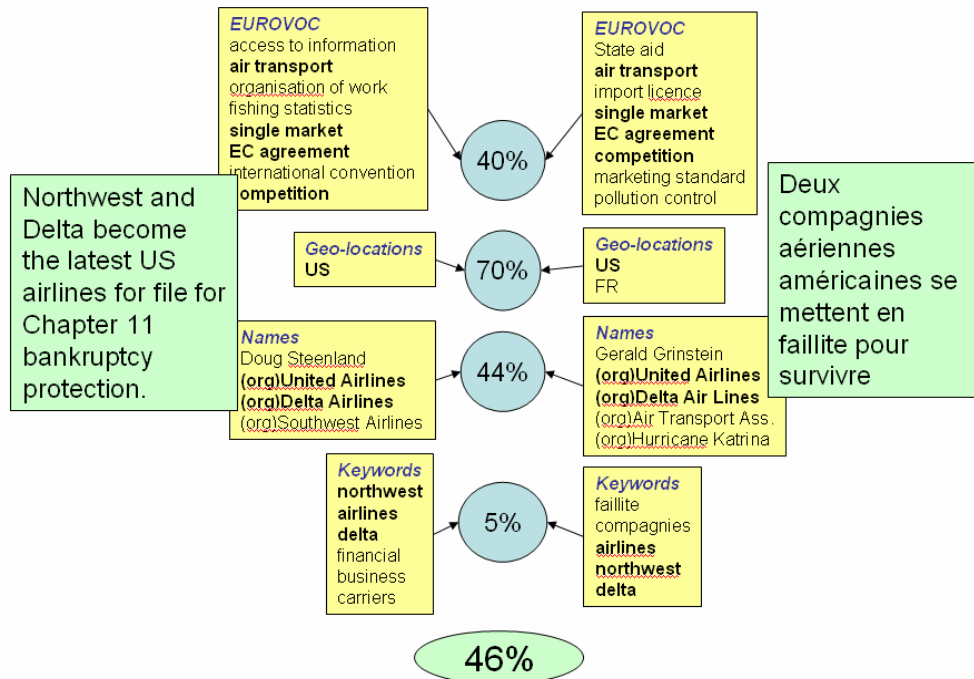


Figure 1. Combination of four ingredients to calculate the document similarity between an English and a French document.