Story tracking: linking similar news over time and across languages

Bruno Pouliquen & Ralf Steinberger
European Commission
Joint Research Centre
Via E. Fermi 2749, 21027 Ispra, Italy
Firstname.Lastname@jrc.it

Olivier Deguernel
Temis S.A.
Tour Gamma B, 193-197 rue de Bercy
75582 Paris Cedex, France
Olivier.Deguernel@temis.com

Abstract

The *Europe Media Monitor* system (EMM) gathers and aggregates an average of 50,000 newspaper articles per day in over 40 languages. To manage the *information overflow*, it was decided to group similar articles per day and per language into clusters and to link daily clusters over time into *stories*. A story automatically comes into existence when related groups of articles occur within a 7-day window. While cross-lingual links across 19 languages for individual news clusters have been displayed since 2004 as part of a freely accessible online application (http://press.jrc.it/NewsExplorer), the newest development is work on linking entire stories across languages. The evaluation of the monolingual aggregation of historical clusters into stories and of the linking of stories across languages yielded mostly satisfying results.

1 Introduction

Large amounts of information are published daily on news web portals around the world. Presenting the most important news on simple, newspaper-like pages is enough when the user wants to be informed about the latest news. However, such websites do not provide a long-term view on how any given story or event developed over time. Our objective is to provide users with a fully automatic tool that groups individual news articles every day into *clusters* of related news and to aggregate the daily clusters into stories, by linking them to the related ones identified in the previous weeks and months. In our jargon, *stories* are thus groups of articles talking about a similar event or theme *over time*. We work with the daily clusters computed by the NewsExplorer application (Pouliquen et al. 2004). For each daily cluster in currently nineteen languages, the similarity to all clusters produced during the previous seven days is computed and a link is established if the similarity is above a certain threshold. It is on the basis of these individual links that stories are built, i.e. longer chains of news clusters related over time. The current NewsExplorer application additionally identifies for all news clusters, whether there are related clusters in the other languages. These daily cross-lingual links are used to link the longer-lasting stories across languages.

After a review of related work (Section 12), we will present the *Europe Media Monitor* (EMM) system and its NewsExplorer application (section 3). We will then provide details on the process to build the multi-monolingual stories (Section 4) and on the more recent work on linking stories across languages (Section 5). Section 6 presents evaluation results both for the monolingual story compilation and for the establishment of cross-lingual links. Section 7 concludes and points to future work.

2 Related work

The presented work falls into the two fields of *Topic Detection and Tracking* and cross-lingual document similarity calculation.

2.1 Topic detection and tracking (TDT)

TDT was promoted and meticulously defined by the US-American DARPA programme (see Wayne 2000). An example explaining the TDT concept was that of the Oklahoma City bombing in 1995, where not only the bombing, but also the related memorial services, investigations, prosecution etc. were supposed to be captured.

Coling 2008: Proceedings of the workshop on Multi-source Multilingual Information Extraction and Summarization, pages 49–56
Manchester, August 2008

Human evaluators will often differ in their opinion whether a given document belongs to a topic or not, especially as 'topic' can be defined broadly (e.g. the Iraq war and the following period of insurgence) or more specifically. For instance, the capture and prosecution of Saddam Hussein, individual roadside bombings and air strikes, or the killing of Al Qaeda leader Abu Musab al-Zarqawi could either be seen as individual topics or as part of the Iraq war. This fuzziness regarding what is a 'topic' makes a formal evaluation rather difficult. Our system is more inclusive and will thus include all the mentioned sub-events into one topic (story). A separate clustering system was developed as part of the *EMM-NewsBrief* (http://press.jrc.it/NewsBrief/), which produces more short-lived and thus more specific historical cluster links.

2.2 Cross-lingual linking of documents

Since 2000, the TDT task was part of the TIDES programme (Translingual Information Detection, Extraction and Summarisation), which focused on cross-lingual information access. The goal of TIDES was to enable English-speaking users to access, correlate and interpret multilingual sources of real-time information and to share the essence of this information with collaborators. The purpose of our own work includes the topic detection and tracking as well as the cross-lingual aspect. Main differences between our own work and TIDES are that we need to monitor more languages, that we are interested in all cross-lingual links (as opposed to targeting only English), and that we use different methods to establish cross-lingual links (see Section 5).

All TDT and TIDES participants used either Machine Translation (MT; e.g. Leek et al. 1999) or bilingual dictionaries (e.g. Wactlar 1999) for the cross-lingual tasks. Performance was always lower for cross-lingual topic tracking (Wayne 2000). An interesting insight was formulated in the "native language hypothesis" by Larkey et al (2004), which states that topic tracking works better in the original language than in (machine-)translated collections. Various participants stated that the usage of named entities helped (Wayne 2000). Taking these insights into account, we always work in the source language and make intensive use of named entities.

Outside TDT, an additional two approaches for linking related documents across languages have been proposed, both of which use bilingual vector space models: Landauer & Littman (1991) used bilingual *Lexical Semantic Analysis* and Vinokourov et al. (2002) used *Kernel Canonical Correlation Analysis*. These and the approaches using MT or bilingual dictionaries have in common that they require bilingual resources and are thus not easily scalable for many language pairs. For N languages, there are `N*(N-1)/2` language pairs (e.g. for 20 languages, there are 190 language pairs and 380 language pair directions). Due to the multilinguality requirement in the European Union (EU) context (there are 23 official EU languages as of 2007), Steinberger et al. (2004) proposed to produce an interlingual document (or document cluster) representation based on named entities (persons, organisations, disambiguated locations), units of measurement, multilingual specialist taxonomies (e.g. medicine), thesauri and other similar resources that may help produce a language-independent document representation. Similarly to Steinberger et al. (2004), the work described in the following sections equally goes beyond the language pair-specific approach, but it does not make use of the whole range of information types.

In Pouliquen et al. (2004), we showed how NewsExplorer links individual news clusters over time and across languages, but without aggregating the clusters into the more compact and high-level representations (which we call *stories*). This new level of abstraction was achieved by exploiting the monolingual and cross-lingual cluster links and by adding additional filtering heuristics to eliminate wrong story candidate clusters. As a result, long-term developments can now be visualised in timelines and users can explore the development of events over long time periods (see Section 4.2). Additionally, meta-information for each story can be compiled automatically, including article and cluster statistics as well as lists of named entities associated to a given story.

2.3 Commercial applications

Compared to commercial or other publicly accessible news analysis and navigation applications, the one presented here is unique in that it is the only one offering automatic linking of news items related either historically or across languages. The news aggregators *Google News* (http://news.google.com) and *Yahoo! News* (http://news.yahoo.com/), for instance, deliver daily news in multiple languages, but do not link the found articles over time or across languages. The monolingual English language applications *DayLife* (http://www.daylife.com/), *SiloBreaker* (http://www.silobreaker.com/), and *NewsVine*

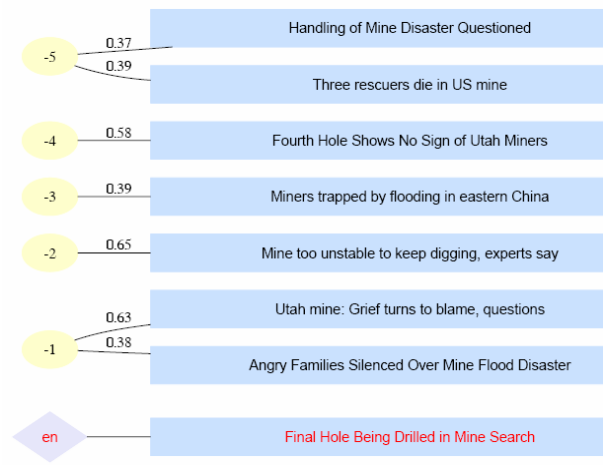

Figure 1. Example of historical links between clusters: The graph shows the cosine similarity between today's English language cluster (*Final hole being drilled …*) and seven clusters identified during five previous days. Only clusters with a similarity above 0.5 will be retained.

(http://www.newsvine.com/) do not link related news over time either. *NewsTin* (http://www.newstin.com) is the only one to offer more languages (ten) and to categorise news into a number of broad categories, but they, again, do not link related news over time or across languages.

3 Europe Media Monitor (EMM) & NewsExplorer

EMM has been gathering multilingual news articles from many different web portals since 2002. It's *NewsBrief* application has since displayed the world's most recent news items on its public web servers (http://emm.jrc.it/overview.html). Every day, and for each of 19 languages separately, EMM's *NewsExplorer* application groups related articles into *clusters*. Clusters are computed using a group average agglomerative bottom-up clustering algorithm (similar to Schultz & Liberman 1999). Each article is represented as a vector of keywords with the keywords being the words of the text (except stop words) and their weight being the log-likelihood value computed using word frequency lists based on several years of news. We additionally enrich the vector space representation of each cluster with country information (see Pouliquen et al., 2004), based on log-likelihood-weighted, automatically recognised and disambiguated location and country names (see Pouliquen et al. 2006).

Each computed daily cluster consists of its keywords (i.e. the average log-likelihood weight for each word) and the title of the cluster's medoid (i.e. the article closest to the centroid of the cluster). In addition we enrich the cluster with features that will be used in further processes. These include the cluster size, lists of persons, organisations, geo-locations and subject domain codes (see Section 5).

When comparing two clusters in the same language, the keywords offer a good representation (especially when the keywords are enriched with the country information). Section 5 will show that the additional ingredients are useful to compare two clusters in different languages.

4 Building stories enriched with meta-information

For each language separately and for each individual cluster of the day, we compute the *cosine* similarity with all clusters of the past 7 days (see Figure 1). Similarity is based on the keywords associated with each cluster. If the similarity between the keyword vectors of two clusters is above the empirically derived threshold of 0.5, clusters are linked. This optimised threshold was established by evaluating cluster linking in several languages (see Pouliquen et al. 2004). A cluster can be linked to several previous clusters, and it can even be linked to two different clusters of the same day.

4.1 Building stories by linking clusters over time

Stories are composed of several clusters. If a new cluster is similar to clusters that are part of a story, it is likely that this new cluster is a continuation of the existing story. For the purpose of building stories, individual and yet unlinked clusters of the previous seven days are treated like (single cluster) stories. If clusters have not been linked to within seven days, they remain individual clusters that are not part of a story. Building stories out of clusters is done using the following incremental algorithm (for a given day):

```
for each cluster c
 for each story s
  score[s]=0;
  for each cluster cp (linked to c)
   if (s: story containing cp) then
    score[s] += (1-score[s])*sim(cp,s);
   endif
  endfor
 endfor
 if (s: story having the maximum score)
 then
  add c to story s (with sim score[s])
 else // not similar to any story
  create new story containing only c
 endif
endfor
```

Lang	Biggest title	Keywords
En	US Airways won't pursue Delta forever	*United states* / **Doug Parker, Delta Airlines** / airways, offer, emerge, grinstein, bid, regulatory, creditors, bankruptcy, atlanta, increased
It	Stop al massacro di balene. Il mondo contro il Giappone	*Australia, N. Zealand, Japan*/ **Greenpeace International, John Howard**/ caccia, megattere, balene, sydney, acqua, mesi, antartico, salti
Es	Mayor operación contra la pornografía infantil en Internet en la historia de España	**Guardia Civil, Fernando Herrero Tejedor** / pornografía, imputados, mayor, cinco, delito, internet, registros, siete, informática, sci
De	Australian Open: "Tommynator" mit Gala-Vorstellung	*Russia, Australia, United states* / **Australian Open, Mischa Zverev** / satz, tennis, deutschen, bozoljac, erstrunden, melbourne, kohlschreiber, Donnerstag
Fr	Il faut aider l'Afrique à se mondialiser, dit Jacques Chirac	**Jacques Chirac, African Union** / afrique, sommet, continent, président, cannes, darfour, état, pays, conférence, chefs, omar

Table 1. Examples of stories, their biggest titles and their corresponding keywords. Countries are displayed in italic, person and organisation names in boldface.

with *sim(cp,s)* being the similarity of the cluster to the story (the first cluster of a story gets a *sim* of 1, the following depend on the *score* computed by the algorithm).

When deciding whether a new cluster should be part of an existing story, the challenge is to combine the similarities of the new cluster with each of the clusters in the story. As stories change over time and the purpose is to link the newest events to existing stories, the new cluster is only compared to the story's clusters of the last 7 days. A seven-day window is intuitive and automatically takes care of fluctuations regarding the number of articles during the week (weekends are quieter). In the algorithm to determine whether the new cluster is linked to the story, the similarity score is computed incrementally: The score is the similarity of the new cluster with the latest cluster of the story (typically yesterday's) plus the similarity of the new cluster with the story's cluster of the day before multiplied with a reducing factor (1-score$_{i-1}$), plus the similarity of the new cluster with the story's cluster of yet another day before multiplied with a reducing factor (1-score$_{i-2}$), etc. The reducing factor helps to keep the similarity score between the theoretical values 0 (unrelated) and 1 (highly related):

$$score_i = \begin{cases} 0 & (i = 0) \\ (1 - score_{i-1}) \cdot sim(c_i, s) & (0 < i < 7) \end{cases}$$

If the final score is above the threshold of 0.5, the cluster gets linked to the existing story. Otherwise it remains unlinked. The story building algorithm is language-independent and could thus be applied to all of the 19 NewsExplorer languages. Currently, it is run every day (in sequential order) in the following nine languages: Dutch, English, French, German, Italian, Portuguese, Slovene, Spanish and Swedish.

Out of the daily average of 970 new clusters (average computed for all nine languages over a period of one month), only 281 get linked to an existing story (29%) and 90 contribute to a new story (9%). The remaining 599 clusters (62%) remain unlinked singleton clusters. A small number of stories are very big and go on over a long time. This reflects big media issues such as the Iraq insurgence, the Iran-nuclear negotiations and the Israel-Palestine conflict. The latter is the currently longest story ever (see http://press.jrc.it/NewsExplorer/storyedition/en/RTERadio-5f47a76fe35215964cbab22dcbc88d7b.html).

4.2 Aggregating and displaying information about each story

For each story, daily updated information gets stored in the NewsExplorer knowledge base. This includes (a) the title of the first cluster of the story (i.e. the title of the medoid article of that first cluster); (b) the title of the biggest cluster of the story (i.e. the cluster with most articles); (c) the most frequently mentioned person names in the story (*related people*); (d) the person names most highly associated to the story (*associated people*, see below); (e) the most frequently mentioned other names in the story (mostly organisations, but also events such as *Olympics*, *World War II*, etc.); (f) the countries most frequently referred to in the story (either directly with the country name or indirectly, e.g. by referring to a city in that country); (g) a list of keywords describing the story (see below). This meta-information is exported every day into XML files for display on NewsExplorer. The public web pages display up to 13 keywords, including up to three country names and up to two person or organisation names (see Table 1). To

Figure 2. Examples of English language stories, as on the NewsExplorer main page (2.04. 2008).

see examples of all meta-information types for each story, see the NewsExplorer pages.

Stories are currently accessible through three different indexes (see Figure 2): the stories of the week, the stories of the month and the biggest stories (all displayed on the main page of NewsExplorer). The biggest stories are ordered by the number of clusters they contain without any consideration of the beginning date or the end date. The stories of the month present stories that started within the last 30 days, stories of the week those that started within the last seven days.

For each story, a time line graph (a flash application taking an XML export as input) is produced automatically, allowing users to see trends and to navigate and explore the story (Figure 3). While a story can have more than one cluster on a given day, the graph only displays the largest cluster for that day.

The story's keyword signature is computed using the keywords appearing in most of the constituent clusters. If any of the keywords represents a country, it will be displayed first. A filtering function eliminates keywords that are part of one of the selected entities. For instance, if a selected entity is *George W. Bush* and a selected country is *Iraq*, the keywords *Bush*, *George*, *Iraqi*, etc. will not be displayed.

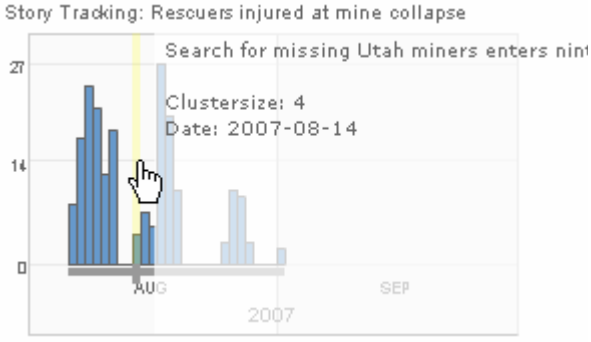

Figure 3. Sample of a short story timeline. When mousing over the graph, title, date and cluster size for that day are displayed. A simple click allows to jump to the relevant cluster, enabling users to explore the story. Available on page http://press.jrc.it/NewsExplorer/storyedition/en/guardian-ee9f870100be631c0147646d29222de9.html.

As mentioned in the previous paragraph, a story's *related entities* are those that have been mentioned most frequently. This typically includes many media VIPs. *Associated entities* are names that appear in this particular story, but are *not* so frequently mentioned in news clusters outside this story, according to the following, TF.IDF-like formula:

$$related(S,e) = \sum_{c_i \in S} fr(c_i, e)$$

$$ass(S,e) = \frac{\sum_{c_i \in S} fr(c_i, e)}{\min(\log(fr(e)),1)} \cdot (1 + \log(C(S,e))$$

with *fr(e)* being the number of clusters the entity appears in (in a collection of three years of news) and *C(S,e)* being the number of clusters *in the story S* mentioning the entity. Inversely, the NewsExplorer person and organisation pages also display, for each entity, the biggest stories they are involved in.

5 Cross-lingual cluster and story linking

For each daily cluster in nine NewsExplorer languages, the similarity to clusters in the other 18 languages is computed. To achieve this, we produce three different language-independent vector representations for each cluster (for details, see Pouliquen et al. 2004): a weighted list of Eurovoc subject domain descriptors (*eurov*, available only for EU languages), a frequency list of person and organisation names (*ent*), and a weighted list of direct or indirect references to countries (*geo*). As a fourth ingredient, we also make use of language-dependent keyword lists because even monolingual keywords sometimes match

across languages due to cognate words (*cog*), etc. (e.g. *tsunami*, *airlines*, *Tibet* etc.). The overall similarity *clsim* for two clusters *c'* and *c''* in different languages is calculated using a linear combination of the four cosine similarities, using the values for α, β, γ & λ as 0.4, 0.3, 0.2 and 0.1, respectively (see Figure 4):

$$clsim(c', c'') = \alpha \cdot eurov(c', c'') + \beta.geo(c', c'') + \gamma.ent(c', c'') + \lambda.cog(c', c'')$$

5.1 Filtering and refining cross-lingual cluster links

The process described in the previous paragraphs produces some unwanted cross-lingual links. We also observed that not all cross-lingual links are transitive although they should be. We thus developed an additional filtering and link weighting algorithm to improve matters, whose basic idea is the following: When clusters are linked in more than two languages, our assumption is: If cluster A is linked to cluster B and cluster C, then cluster B should also be linked to cluster C. We furthermore assume that if cluster B is not linked to cluster C, then cluster B is less likely to be linked to cluster A. The new algorithm thus checks these 'inter-links' and calculates a new similarity value which combines the standard similarity (described in 5.0) with the number of inter-links. The formula punishes links to an isolated cluster (i.e. links to a target language cluster which itself is not linked to other linked languages) and raises the score for inter-linked clusters (i.e. links to a target language cluster which itself is linked to other linked languages). The new similarity score uses the formula:

$$clsim'(C', C'') = clsim(C', C'').\frac{Cl(C')}{\sqrt{El(C')}}$$

with *Cl*(*C*) being the number of computed cross-lingual links and *El*(C) being the number of expected cross-links (i.e. all cross-language links observed when looking at all languages). For instance, if a cluster is linked to three languages and these are linked to a further three, then *Cl(C')*=3 and *El(C')*=6.

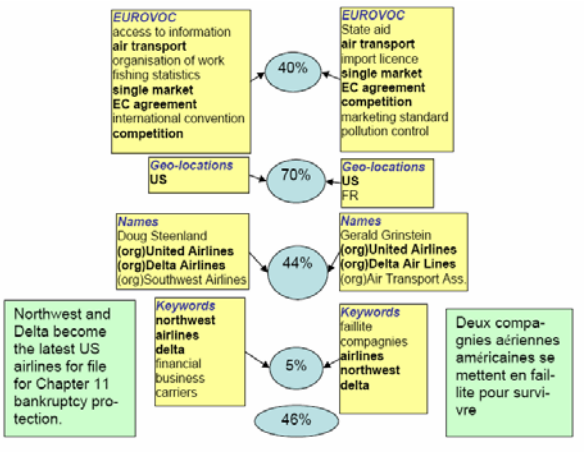

Figure 4. Example of the similarity calculation for an English and a French cluster. The overall similarity for these two clusters, based on the linear combination of four different vectors, is 0.46.

5.2 Linking whole stories across languages

The stories contain clusters which are themselves linked to clusters in other languages (see 5.1). This information can be used to compute the similarity between two whole stories in different languages. The formula is quite simple:

$$Sclsim(S', S'') = \sum_{c'_i \in S', c''_j \in S''} clsim'(c'_i, c''_j)$$

with *S'* and *S''* being two stories in different languages, and *c'* and *c''* being constituent clusters. Cross-lingual cluster similarity values are only added if they are above the threshold of 0.15. Table 2 shows an English story and its links in seven languages.

As the evaluation results in Section 6 show, this formula produces reasonable results, but it has some limitations. Firstly, it relies exclusively on

Lang.	Biggest title	Nb. of clusters	Nb. of articles	Common clusters	Similarity
En	Rescuers injured at mine collapse	17	200	---	---
Pt	EUA: mineiros presos numa mina continuam incontactáveis	12	63	7	2.1363
Es	Colapsa mina en EE.UU.	5	24	3	0.9138
De	USA: Sechs Bergleute eingeschlossen	3	28	2	0.7672
Nl	Mijnwerkers vast na aardbeving in Utah	2	7	2	0.6082
Fr	Le sauvetage de mineurs dans l'Utah tourne au drame	3	16	2	0.5541
Nl	Reddingswerkers omgekomen in mijn Utah	2	12	2	0.4644
Sv	Mystisk "ubåt" undersöks i New York	4	16	2	0.3681

Table 2. Example of cross-lingual links between the English language *US mine collapse* story and stories in seven other languages. The Swedish story, which has the lowest similarity score, is actually unrelated.

daily cross-lingual links, whereas stories are not necessarily reported on the same day across languages. Secondly, we might be able to produce better results by making use of the available meta-information *at story level* described in Section 4.2. We are thus planning to refine this formula in future work.

6 Evaluation

Evaluating such a system is not straightforward as there is a lot of room for interpretation regarding the relatedness of clusters and stories. Cluster consistency evaluation and the monolingual and cross-lingual linking of individual clusters using a very similar approach has already been evaluated in Pouliquen et al. (2004).

In order to evaluate the precision for the story building in four languages, we have evaluated the relatedness of the individual components (the clusters) with the story itself. We compiled a list of 330 randomly selected stories (in the 4 languages English, German, Italian and Spanish) and asked an expert to judge if each of the clusters is linked to the main story. For each story, we thus have a ratio of 'correctly linked' clusters (see Table 3). The average ratio corresponds to the precision of the story tracking system. There clearly is room for improvement, but we found the results good enough to display the automatically identified stories as part of the live application.

We did make an attempt at evaluating also the recall for story building, but soon found out that the results would not make sense. The idea was to carry out a usage-oriented evaluation for the situation in which users are looking for any story of their choice using their own search words (e.g. *Oscar* and *nomination*, *Pavarotti* and *death*, etc.). It was found that relevant stories did indeed exist for almost every query. However, the results would entirely depend on the type of story the evaluator is looking for and on the evaluator's capacity to identify significant search words. We can thus not present results for the recall evaluation of the story tracking system.

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

Table 3. Evaluation of the monolingual linking of clusters into stories for four languages.

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

Table 4. Evaluation of cross-lingual story linking.

The purpose of a second test was to evaluate the accuracy of the cross-lingual story linking. For that purpose, we evaluated those 112 multilingual stories out of the 330 stories in the previous experiment that had cross-lingual links to any of the languages Dutch, English, French, German, Italian, Portuguese, Spanish or Swedish. Table 4 shows that only 59% of the automatically established cross-lingual story links were accurate, but that the situation improves when looking at stories consisting of more clusters, i.e. 5 or 10. This trend was confirmed by a separate study evaluating only the cross-lingual links for the 10 largest stories in the same four languages, into the same eight other languages: 87% of the cross-lingual links were correct. Note that – for these large stories – the cross-lingual links were 96.5% complete (270 out of 280 possible links were present). Further insights from this evaluation are that there are only two out of the 40 top stories that should be merged (there are two English top stories on Israel) and that there is one cluster in each of the four languages which should be split (all China-related news merges into one story). It is clear that more experiments are needed to improve the cross-lingual links for smaller stories. We have not evaluated the recall of the cross-lingual story linking as recall evaluation is very time-consuming and we first want to optimise the algorithm.

7 Conclusion and Future Work

The story tracking system has been running for two years. There is definitely space for improvement as unrelated clusters are sometimes part of a story, but informal positive user feedback makes us believe that users already find the current results useful. An analysis of the web logs shows that more than 400 separate visitors per day look at story-related information, split quite evenly across the different languages (Table 5).

The story tracking algorithm is rather sensitive to the starting date for the process: Different starting dates may result in different stories and certain starting dates may result in having two separate parallel stories talking about very closely related subjects. Another issue is the seven-day window: We may want to extend the window as it happens occasionally that a story 'dies' because no related articles are published on the subject for a week, and that another story talking about the same subject starts 8 days later. Finally, our algorithm should try to cope with the fact that stories can split or merge (an issue not currently dealt with), but this is a non-trivial issue.

Regarding the cross-lingual linking, the current results are encouraging, but not sufficient. The accuracy needs to be improved before the results can go online. The most promising idea here is to make use of each story's meta-information (lists of related persons, organisations, countries and keywords at story level) and to allow a time delay in the publication of stories across languages. However, the application has high potential, as it will provide users with (graphically visualisable) information on how the media report events across languages and countries.

In a separate effort, a 'live' news clustering system has been developed within EMM, which groups the news as they come in during the day (see http://press.jrc.it/NewsBrief/). This process needs to be integrated with the daily and more long-term story tracking process so that users can explore the history and the background for current events.

Acknowledgements

We thank the *Web Mining and Intelligence* team and our team leader Erik van der Goot for the valuable news data and the robust web sites. A special thanks to Jenya Belyaeva for her evaluation.

Lang	Hits	Pct	Hits/ day	Visits	Visits /day	Pct
De	59993	14%	2143	1611	58	13%
En	164557	38%	5877	2273	81	19%
Es	49360	11%	1763	1431	51	12%
Fr	56023	13%	2001	1514	54	12%
It	29445	7%	1052	1425	51	12%
Nl	25175	6%	899	1242	44	10%
Pt	42933	10%	1533	2170	78	18%
Sv	7284	2%	260	575	21	5%
Total:	434770		15527	12241	437	

Table 5. Number of connections to *story-related* NewsExplorer web pages only, and distribution per language (period 1-28/06/2008). Only visits from different IP addresses were counted.

References

Landauer Thomas & Michael Littman (1991). A Statistical Method for Language-Independent Representation of the Topical Content of Text Segments. Proceedings of the 11th International Conference 'Expert Systems and Their Applications', vol. 8: pp. 77-85.

Larkey Leah, Fangfang Feng, Margaret Connell, Victor Lavrenko (2004). *Language-specific Models in Multilingual Topic Tracking*. Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 402-409.

Leek Tim, Hubert Jin, Sreenivasa Sista & Richard Schwartz (1999). The BBN Crosslingual Topic Detection and Tracking System. In 1999 TDT Evaluation System Summary Papers.

Pouliquen Bruno, Ralf Steinberger, Camelia Ignat, Emilia Käsper & Irina Temnikova (2004). *Multilingual and cross-lingual news topic tracking*. In: Proceedings of the 20th International Conference on Computational Linguistics, Vol. II, pp. 959-965.

Pouliquen Bruno, Marco Kimler, Ralf Steinberger, Camelia Ignat, Tamara Oellinger, Ken Blackler, Flavio Fuart, Wajdi Zaghouani, Anna Widiger, Ann-Charlotte Forslund & Clive Best (2006). Geocoding multilingual texts: Recognition, Disambiguation and Visualisation. Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'2006), pp. 53-58.

Schultz J. Michael & Mark Liberman (1999). Topic detection and Tracking using idf-weighted Cosine Coefficient. DARPA Broadcast News Workshop Proceedings.

Steinberger Ralf, Bruno Pouliquen & Camelia Ignat (2004). Providing cross-lingual information access with knowledge-poor methods. In: Andrej Brodnik, Matjaž Gams & Ian Munro (eds.): Informatica. An international Journal of Computing and Informatics. Vol. 28-4, pp. 415-423. Special Issue 'Information Society in 2004'.

Vinokourov Alexei, John Shawe-Taylor, Nello Cristianini (2002). *Inferring a semantic representation of text via cross-language correlation analysis*. Advances of Neural Information Processing Systems 15.

Wactlar Howard (1999). *New Directions in Video Information Extraction and Summarization*. Proceedings of the 10th DELOS Workshop.

Wayne Charles (2000). *Multilingual topic detection and tracking: Successful research enabled by corpora and evaluation*. Proceedings of 2nd International Conference on Language Resources and Evaluation.