

# Linking equivalent news across languages

Camelia Ignat and Ralf Steinberger  
*European Commission, Joint Research Centre*

## **Abstract**

We address the task of linking equivalent news items in real time across dozens of languages. The tool we developed will be part of a system processing a continuous multilingual high-volume news stream. For each language, the current system groups the related news articles into news clusters, recognises names of persons, organisations and locations and performs a content classification. Our new component identifies equivalent news clusters across languages, using language-independent features as weighted vectors, and calculates the news cluster similarity as their linear combination. In this paper, we describe the method and experimental results for the tuning and weighting of each of the feature vectors. By determining appropriate similarity thresholds, we manage to achieve good results ready to go live as part of the multilingual news processing system.

## **1. Introduction**

There are numerous services that aggregate and cluster monolingual news items, but there are few that link equivalent news across languages. The **purpose of our work** is to provide newsreaders with a qualitatively enhanced view of the newsscape by giving them an insight of what media in other countries say about the same news story. This functionality will allow them to be more widely informed by going past their dominant national viewpoint. As newsreaders will not necessarily understand the equivalent foreign language text, our platform offers English Machine Translation results for 17 languages and it displays extracted meta-information (news categories, named entities), giving the readers at least a glimpse of the complementary information in the foreign language text. These cross-lingual equivalence links furthermore open up the possibility to carry out an automated empirical study of differences in news reporting across countries and languages: Who started reporting first and where did a news story end earlier? What is the focus and context of the various national media outlets? More im-

portantly, what are their differences? Which stories get published internationally and which ones are predominantly local?

**What are equivalent news** across languages? Is it sufficient for both articles to be about the economy or about elections? These categories are too generic. Our view of equivalence is user-driven: If persons read the news, want to know what the media in other countries say about the same story or subject and follow our automatically generated link, they should recognise what they just read (e.g. ongoing Kenyan elections) while also finding complementary information. To give a concrete example: The news in country A reports about the expected outcome of the ongoing elections while the news in country B may focus on disagreements in the leading party just before the elections, or on suspected election fraud. While the grey areas are large, our human annotators had the task to judge whether they were satisfied with the proposed foreign language news cluster. In addition to ‘equivalent’ and ‘not equivalent’, they could also use the tag ‘related’, to leave room for individual interests and preferences. When tuning our cross-lingual linking algorithm, we most of all aimed at minimising the retrieval of news that were not equivalent (focus on *precision*). While gold-standard equivalent news were the aim, related stories were acceptable.

**Major challenges** for our task are the lack of multilingual training data, the continuous high-volume news flow combined with the requirement that readers can continuously see results, as well as the large number of different languages for which different amounts of structured meta-data are available. Our solution to the data bottleneck is to find invariant elements across languages that can serve as near-language-independent features. Largely, news answers the questions: “Who does What to Whom, Where and When?” The answers to these questions are Named Entities (persons, organisations, locations) and content categories. We assume the ‘When’ to be the day of reporting. We also make use of Machine Translation results (into English), when available. Our Named Entity Recognition (NER) and subject domain categorisation tools are not available for all languages and they have varying coverage for the many languages. Our tool thus has to be robust enough to deal with partial information.

Our approach was to test first which of these five features performs best if it were the *only* linking elements across languages, for three languages. While evaluating, our human annotators built up a gold-standard collection of bi-

lingual borderline news cluster pairs that we used later to optimise the parameter settings for each of them, and for the linear combination of the individual similarities. We hope to make this cluster equivalence set publicly available.

Next, we put our work into context by reporting about related work (Section 2). We then present the existing news analysis system (3) and detail our approach on cluster linking (4): the features with different weighting methods, the similarity calculation with the adjusting metrics, and the groups of clusters. The experiment section presents the dataset, the setup and results (5). The final section concludes and points to planned work (6).

## **2. Related work**

We present different approaches used for cross-lingual document similarity and cross-lingual linking of news described in the literature.

We distinguish the probabilistic approaches that describe the multilingual document collections as samples from generative probabilistic models, with variations on the assumptions on the model structure. Most of them are based on Gibbs sampling or variational inference, which are non-trivial to implement.

There are matrix factorisation-based approaches that include non-negative matrix factorisation, Cross-lingual Latent Semantic Indexing (CL-LSI), Canonical Correlation Analysis (CCA), Oriented Principal Component Analysis, Cross-lingual Explicit Semantic Analysis or different combinations, like CCA and CL-LSI in Rupnik (2016).

Other approaches make use of language-independent representations, allowing applying monolingual similarity computing methods. These representations may use Machine Translation (Potthast et al. 2010, Seki, 2018), dictionaries or thesauri (Pouliquen, 2008; Rodriguez, 2008;), named entities - as person names, organisations, geographical places (Belyaeva, 2015; Pouliquen et al. 2008) - or cognates (Pouliquen, 2008).

### **Cross-lingual linking of news**

Although there are a number of services that aggregate news by identifying clusters of similar articles, very few services provide linking of news clusters over different languages.

Rupnik et al. (2017) describe different cross-lingual similarity measures trained on Wikipedia. Their system, *Event Registry*, uses *Canonical Correlation Analysis* (CCA) to detect related events in multilingual news. Belyaeva (2016) uses the same technique (CCA), combined with named entity vectors to improve the cross-lingual linking. Miranda (2018) describes a multilingual clustering method based on embedding vectors (of words) and timestamp features.

The *News Explorer* application of the *Europe Media Monitor* (Pouliquen et al. 2008, Steinberger & Pouliquen 2008) clusters news articles in 60 languages and determines which clusters in different languages report the same event. To achieve cluster linking, four different language-independent vector representations are used: named entities; locations mentioned in the clusters; a weighted list of Eurovoc subject domain descriptors. Similarity between clusters is computed using a linear combination of the cosine similarities of the three vectors. If the similarity is above the threshold, the clusters are linked. *MediaGist* (Steinberger 2016) is very similar to News Explorer. It has considerably lower coverage, but it includes a sentiment analysis and a summarisation module.

Our own method was inspired by News Explorer, but our situation is different: We work on live clusters (updated every 10 minutes), for which the cross-lingual links must be computed in real time. Our system uses more features (content categories and machine translation results). Finally, we establish the best-performant settings empirically, both for the individual features and for the overall linear combination.

### **3 Framework / Pipeline**

Our cross-lingual news-linking tool will be part of an existing system for news aggregation and analysis that processes hundreds of thousands of articles per day in dozens of languages (Anonymous Ref 1). The system continuously

monitors RSS feeds and web pages, detects and downloads new articles and sends them through a processing pipeline. At each step, the news article's RSS file gets enriched with additional information, part of which is used by our cluster-linking tool. At any moment, the existing news processing system gives a monolingual view of the latest news clustered into *stories*.

We now present those system modules that are relevant for our algorithm:

1. **Clustering of articles:** Every 10 minutes, the module performs an incremental topic-based clustering for the news that have arrived during the last 4 to 8 hours (depending on the volume of news per language). It uses a bottom-up hierarchical average-linking clustering algorithm.
2. **Entity extraction:** The entity extraction module uses pattern recognition to look up an automatically updated set of currently 500.000 known person and organisation names, plus their language variants. An entity guesser detects new entities or variants (Anonymous Ref 2). Monolingual and multilingual spelling variants have the same unique identifier.
3. **Geotagging:** The module (Anonymous Ref 3) uses a multilingual set of geo-data names enhanced with in-house data. The approximately 1.5 million place name variants mostly include national capitals, regional capitals and provincial capitals. Their coordinates (and a unique identifier) uniquely identify geolocations.
4. **Translation:** The in-house statistical machine translation module is based on Moses (Koehn et al. 2007), with models optimised for the news domain. Only titles and a short teaser text are translated from 17 languages into English. English thus is the pivot language as many language texts can be represented by their English vocabulary.
5. **Categorisation engine:** The module uses user-defined keywords and patterns (using Boolean combinations, proximity parameters and wildcards) to categorise into thousands of specific categories. However, categories can be overlapping and there is no ontology. Categories are the same across all system languages, but the categorisation coverage differs a lot between languages.
6. **Eurovoc:** For the cross-lingual linking task, we have developed a *new module* that additionally categorises clusters using the multilingual Eurovoc thesaurus because Pouliquen et al. (2003, 2008) showed the success of this resource for cross-lingual linking purposes. Using the

publicly available JEX tool (Steinberger et al. 2012) readily trained for 21 EU languages, it produces a list of weighted content classes for each cluster, each with its own unique identifier that is the same across all languages.

The next section shows how we exploit this news meta-data to link clusters across languages.

## 4. Cluster Linking

For the purposes of this work, we produce various independent near-language-independent feature vector representations per news cluster. For each representation, we produce separate cross-lingual similarity calculations, for each of the language pairs. After optimising each single-feature method, we combine the similarity calculations into a linear combined formula. The next sections describe the various vector representations of document clusters (4.1), the measure we use to calculate the similarity between them and some tweaks to improve the similarity calculations for low-dimensional vectors (4.2).

### 4.1 Cluster representation and term weighting

The standard vector space model (Slaton & Buckley, 1988) represents documents as vectors, where each term corresponds to a word or a phrase in a fixed vocabulary. Formally, document  $d$  is represented by a vector  $x \in R^n$ , where  $n$  corresponds to the size of the vocabulary, and vector elements  $x_k$  correspond to the number of times term  $k$  occurred in the document, also called term frequency ( $TF_k(d)$ ). We use the five near-language-independent features listed in Section (3) (numbers two to six) as terms. Each news cluster thus has five representations. The first vector originally consists of a frequency list of all person and organisation names mentioned in the news cluster, the second of a list of all locations, and so on. Regarding bullet number (4) in Section (3) (translation), the basic vector representation is a classic English word frequency list, which – for non-English languages – is based on the English translation of the original text. As for Eurovoc (bullet number (6)), we simply use the weighted list of content categories produced by the JEX tool. However, we changed two of the JEX default parameters to control the size of the JEX output: (a) *nbTermThr*, which limits the total number of Eu-

rovoc categories produced and (b) *evcThr*, which sets a threshold for the minimum relevance value of the document's Eurovoc categories.

For the features mentioned in bullet numbers (2) to (5) in Section (3), we did not use the pure frequency. Instead, we explored applying three different feature dimension-weighting mechanisms and selecting the one that performs best (See Section (5) for the experimental results). These are *normalised frequency*, *log-likelihood* and *TF-IDF*. We used seven months of news data to produce the reference lists, separately for each language.

We define **normalised frequency *freq*** (for a specific feature) as the number of term occurrences divided by the total number of terms in a cluster:

$$freq_k = TF_k / \sum_{i=0}^n TF_i$$

The log-likelihood weighting is based on **Dunning's statistical log-likelihood** test (or  $G^2$ ; Dunning 1993), which 'compares' the term frequency  $TF_k$  () in the cluster with the frequency in our seven-month reference corpus.

The commonly used **TF.IDF weighting** (*Term Frequency – Inverse Document Frequency*) is calculated as

$$TFIDF_k = TF_k \times IDF_k, \text{ where } IDF_k = \log(N/DF_k),$$

with  $N$  being the total number of documents in the corpus and  $DF_k$  the number of documents in the corpus that contain term  $k$ .

## 4.2 Similarity calculation – penalty metrics

A common way to compute similarity between vectors is cosine similarity. For each of the features separately, we calculate the similarity between two clusters as the cosine between the weighted vectors generated by the feature. Thus, for two documents represented by the feature vectors  $x$  and  $y$ , the similarity is:

$$sim(x, y) = \frac{\langle x, y \rangle}{\|x\| \|y\|}, \text{ where } \langle x, y \rangle \text{ is the inner product and } \|x\|, \|y\| \text{ are the Euclidian norms.}$$

For feature vectors with very few non-zero dimensions, the cosine formula produces similarity values that are inappropriately high. For instance, all

news clusters mentioning only one location (e.g. *Paris*) will have a similarity score of 1 for the geolocation feature. As news mentioning the same place do not necessarily address the same story, we use adjusting metrics, a **dimension penalty**. This penalty decreases the similarity value in case of low dimensionality (of non-zero values) of the feature vectors. We compute the similarity as:

$$Sim_{FEATURE}(x, y) = cosine(x, y) \times DimPenalty(x, y)$$

$DimPenalty(x, y) = \log_2(1 + \frac{|x \cap y|_{x_i, y_i \neq 0}}{3})$ , where  $|x \cap y|_{x_i, y_i \neq 0}$  corresponds to the number of non-zero common values between the feature vectors  $x$  and  $y$ .

Another case where we found that the cosine formula produces an unjustifiably high similarity is when a pair of clusters presents high weights of one common term and larger numbers of low-weight uncommon terms. For instance, when using the named entity vector, the name *Donald Trump* can occur with a high frequency and score in two clusters, triggering a high cosine similarity value. Meantime, a large number of different non-common names may show that the clusters are not at all related. In order to downgrade the importance of the high-weight name *Donald Trump*, we use another adjustment metrics, which we call **Jaccard penalty**. It is based on the *Jaccard Index* (Jaccard, 1901), which considers the ratio between the number of common terms (intersection) and the total number of terms (union) of both vectors, as defined in the formula:

$$JaccPenalty(x, y) = \log_2(1 + JaccCoef(x, y))$$

where  $x, y$  are the feature vectors and  $JaccCoef(x, y) = \frac{\sum_i \min(x_i, y_i)}{\sum_j \max(x_j, y_j)}$ .

The corrected similarity then is:

$$Sim_{FEATURE}(x, y) = cosine(x, y) \times \log_2(1 + JaccCoef(x, y)).$$

We tested different ways of combining these two penalties: separately, combinations of two as the product, the minimum, or the maximum of the values:

$penalty \in \{JaccPenalty, DimPenalty, JaccPenalty \times DimPenalty, \min(JaccPenalty, DimPenalty), \max(JaccPenalty, DimPenalty)\}$ .



Using the penalty, we compute the vector similarity as

$$Sim_{FEATURE}(x, y) = cosine(x, y) \times penalty(x, y).$$

We ran hundreds of experiments to identify which weighting combination works best for each of the individual document vector representations, and to identify the best thresholds. After completing that step, we optimised the global similarity between two clusters by optimising the coefficients in the linear combination of the individual feature vector similarities:

$$sim = coef_{NE} \times Sim_{NE} + coef_{GL} \times Sim_{GL} + coef_{CAT} \times Sim_{CAT} + coef_T \times Sim_T + coef_{EVC} \times Sim_{EVC} + coef_{NG} \times Sim_{NG}.$$

Not all features are available for all languages. When one or more features are missing, the relative importance of the others will go up to ensure that there will always be a cross-lingual news similarity value. The sum of the coefficients will always be 1.

## 5. Experiments

In this section, we present and discuss the results of the experiments we carried out.

### 5.1 Dataset

We built a test data set as a subset of English, French and German news clusters produced on one day. Here is the language distribution:

*Table 1: Dataset - language distribution*

Language	English	French	German
Total clusters	2609	437	825
Test dataset clusters	635	234	244

We have annotated part of the possible links on the test dataset clusters with one of the three labels:

- SIMILAR: when the clusters in a pair report on the same event.

- RELATED: when the clusters in the pair are not exactly about the same event, but they are related. For instance, the causal, implication, inclusion relation.
- NOT RELATED: The clusters in the pair report different events.

We have annotated 1561 links (976 cross-lingual and 585 monolingual) with the distribution detailed in Tables 2 and 3. All experiments were performed on this cross-lingual data.

*Table 2: Cross-lingual annotated links*

<b>Language pair</b>	<b>SIMILAR</b>	<b>RELATED</b>	<b>NOT RELATED</b>	<b>TOTAL</b>
English-French	184	43	75	311
English-German	231	59	274	564
French-German	77	1	23	101
ALL	475	103	214	976

*Table 3 Monolingual annotated links*

<b>Language</b>	<b>SIMILAR</b>	<b>RELATED</b>	<b>NOT RELATED</b>	<b>TOTAL</b>
<i>English</i>	314	1	120	435
<i>French</i>	72	19	5	96
<i>German</i>	34	0	20	54
ALL	420	20	145	585

## 5.2 Experimental setup

We carried out experiments on cross-lingual linking that address the following issues:

1. Similarity tuning by feature:

By maximising the discriminative power of the features and the precision, we have selected the most performant weighting method and similarity formula by feature and we have detected the feature thresholds (monolingual and cross-lingual).

## 2. Global similarity

We determine the threshold for global similarity and the feature coefficients in the linear combination, based on the feature thresholds and the discriminative power of each individual feature. Different coefficient settings were evaluated against the annotated data set. The threshold is selected by maximising the precision on the test dataset.

## 5.3 Results

We used the cross-lingual annotated data to determine the most performant weighting methods and similarity penalty by feature. We have carried out experiments for all the possible combinations of settings, for the three language pairs, with 10 different thresholds. For instance, the “Named Entities” (NE) feature was evaluated on 3 weighting methods (FREQ, LLH and TF.IDF), combined with 5 similarity metrics (cosine with no penalty, Jaccard penalty, dimension penalty, min(Jacc,Dim), max(Jacc,Dim)), performed for 3 language pairs and with 10 different thresholds (450 experiments). In each case, we calculate the number of true positive (the one annotated as Similar and Related), the number of false positive and the precision. The optimal setting was selected by maximising the precision. Table 4 describes the best setting by feature with the precision and recall calculated for the set of language pairs only on the cross-lingual annotated links.

Table 4 Similarity tuning by feature

Feature	Weighting	SimPenalty	Threshold	Prec	Recall
NE	TFIDF	minJaccDim	0.3	0.94	0.24
GL	LLH	maxJaccDim	0.4	0.88	0.17
CAT	TFIDF	minJaccDim	0.5	0.66	0.05
TRANSL	TFIDF	-noPenalty-	0.35	<b>0.99</b>	<b>0.39</b>
EVC	20-0.01	minJaccDim	0.3	0.52	0.13

The feature “Translation” (TRANSL) outcores the others, followed by “Named Entities” (NE). The precision of “Geolocations” (GL) is lower, as the

purpose of our cluster geotagging is to locate the news on a map, and not to give an exhaustive list of all the geographical places that occur in the cluster. The features “Categories” (CAT) and “Eurovoc” (EVC) have a very low discriminative power because they mostly give an indication on the general topics of the cluster.

We used the most performant individual feature settings to generate combinations of features. Table 5 presents experiments with the combined formula (giving the contribution of each feature in percentage) and their performance (in precision and recall) for threshold value 0.25.

Table 5 Combined feature evaluation on cross-lingual annotated data

#EXP	NE(%)	GL(%)	CAT(%)	TR(%)	EVC(%)	Prec	Recall
#1	13	7	2	73	5	<b>0.99</b>	<b>0.5</b>
#2	24	2	2	49	5	0.99	0.47
#3	27	18	3	45	6	0.99	0.46
#4	24	16	3	53	5	0.99	0.48
#5	21	18	3	59	6	0.99	0.49
#6	24	17	5	49	5	0.99	0.47
#7	22	15	4	54	4	0.99	0.49

The combined systems keep a high precision with a big increase in recall. They outperform each of the individual systems. The best performance is achieved when “Translation” (TR) has a coefficient greater than 50%. Wrong links are highly visible so we will not sacrifice the high precision for an improved recall. Performance across the language pairs was comparable.

## 6 Conclusion and future work

In our live news aggregation and analysis system ingesting hundreds of thousands of news articles per day in dozens of languages, we are currently adding a new functionality allowing newsreaders to see in real time equivalent news articles in other languages. As we do not have training data to learn the cross-linking news similarity from data, we present each news cluster by five different vectors that allow a cross-lingual comparison: (normalised) names of persons and organisations, locations, subject domains, Eurovoc categories and Machine Translation results into English. Four of the cluster representa-

tions are near-language-independent and one uses the pivot language English. This drastically reduces the number of cross-lingual similarity calculations for our highly multilingual dataset.

We carried out hundreds of experiments to determine the best parameter settings for each vector representation, as well as the relative contribution of each of these vectors for a combined cross-lingual news cluster similarity. An important feature of the method is its robustness: When one or more of the individual contributing vectors is not available for a certain language pair, the relative contribution of the others will rise. By selecting a safe threshold, we avoid false positives while identifying up to hundreds of equivalent news clusters per language pair. The new method is currently being implemented in the live system. We intend to explore whether we can use the newly created human annotation data to learn the cross-lingual similarity calculations, but we will need to avoid at any cost the similarity calculation for all possible language pair combinations. We would like to make our annotations available publicly and we are currently checking out possible legal constraints.

We plan to apply the same similarity measures monolingually, for each of the languages involved. Our aim is to provide readers with an entirely new reading experience by additionally showing them past news related to the news story they are currently looking at.

## References

Belyaeva Evgenia, Aljaž Košmerlj, Andrej Muhič, Jan Rupnik, Flavio Fuart (2015) - Using semantic data to improve cross-lingual linking of article clusters. *Web Semantics: Science, Services and Agents on the World Wide Web* 35:64-70.

De Smet Wim & Marie-Francine Moens (2009). Cross-language linking of news stories on the web using interlingual topic modelling. *SWSM*, Hong Kong, China.

Dunning Ted (1993). *Accurate Methods for the Statistics of Surprise and Coincidence* Archived 2011-12-15 at the Wayback Machine. *Computational Linguistics*, Volume 19, issue 1 (March 1993).

Jaccard, P. (1901) Distribution de la flore alpine dans le bassin des Dranses et dans quelques régions voisines. Bulletin de la Société Vaudoise des Sciences Naturelles 37, 241-272.

Koehn Philip, Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., & Herbst, E. (2007). Moses: Open Source Toolkit for Statistical Machine Translation. In Proceedings of the 45th ACL, on Interactive Poster and Demonstration Sessions, ACL '07, pp. 177–180 Stroudsburg, PA, USA.

Miranda, S., Znotins, A., Cohen, S.B. & Barzdins, G. (2018). Multilingual Clustering of Streaming News. Proceedings of EMNLP.

Potthast Martin, Alberto Barrón-Cedeño, Benno Stein & Paolo Rosso (2010). Cross-language plagiarism detection. Language Resources and Evaluation 45(1):45-62.

Pouliquen Bruno, Olivier Deguernel, & Ralf Steinberger (2008) – Story tracking: linking similar news over time and across languages. In: Proceedings of the Coling'2008 workshop: Multi-source, multilingual information extraction and summarization, Manchester, August 2008.

Pouliquen Bruno, Ralf Steinberger & Camelia Ignat (2003) – Automatic Annotation of Multilingual Text Collections with a Conceptual Thesaurus In: Proceedings of the EUROLAN Workshop *Ontologies and Information Extraction* at the Summer school The Sematic Web and Language Technology – Its Potential and Practicalities. Bucharest, Romania.


Rupnik Jan, Andrej Muhič, Gregor Leban, Blaž Fortuna and Marko Grobelnik (2017) – News across languages: Cross-lingual document similarity and event tracking, Proceedings of the 26th International Joint Conference on artificial Intelligence (IJCAI 2017), 5050-5054.

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. In *Information processing and management*, pp. 513-523.

Seki, K. (2018). Exploring Neural Translation Models for Cross-Lingual Text Similarity. CIKM.

Steinberger Josef (2016). MediaGist: A cross-lingual analyser of aggregated news and commentaries. Proceedings of the 54th ACL – System Demonstrations, pp 145-150. Berlin, Germany, August 2016.

Steinberger Ralf & Bruno Pouliquen (2008). NewsExplorer - combining various text analysis tools to allow multilingual news linking and exploration. Lecture notes for the SORIA Summer School Cursos de Tecnologias Linguisticas.

Steinberger Ralf, Mohamed Ebrahim & Marco Turchi (2012).  JRC EuroVoc Indexer JEX - A freely available multi-label categorisation tool . Proceedings of LREC, pp. 798-805, Istanbul, 21-27 May 2012. Available online at <https://ec.europa.eu/jrc/en/language-technologies/jrc-eurovoc-indexer>.