

News Across Languages - Cross-Lingual Document Similarity and Event Tracking (Extended Abstract)

Jan Rupnik, Andrej Muhič, Gregor Leban, Blaž Fortuna, Marko Grobelnik

Artificial Intelligence Laboratory, Jožef Stefan Institute,

Jamova cesta 39, 1000 Ljubljana, Slovenia

{ jan.rupnik, andrej.muhic, gregor.leban, blaz.fortuna, marko.grobelnik }@ijs.si

Abstract

In today's world, we follow news which is distributed globally. Significant events are reported by different sources and in different languages. In this work, we address the problem of tracking of events in a large multilingual stream. Within a recently developed system Event Registry we examine two aspects of this problem: how to compare articles in different languages and how to link collections of articles in different languages which refer to the same event. Building on previous work, we show there are methods which scale well and can compute a meaningful similarity between articles from languages with little or no direct overlap in the training data. Using this capability, we then propose an approach to link clusters of articles across languages which represent the same event¹.

1 Introduction

Content on the Internet is becoming increasingly multilingual. A prime example is Wiki-pedia. In 2001, the majority of pages were written in English, while in 2015, the percentage of English articles has dropped to 14%. At the same time, online news has begun to dominate reporting of current events. In this paper we consider the intersection of these developments: how to track events which are reported about in multiple languages.

The term event is vague and ambiguous, but for the practical purposes, we define it as "any significant happening that is being reported about in the media." Examples of events would include shooting down of the Malaysia Airlines plane over Ukraine on July 18th, 2014 and HSBC's admittance of aiding their clients in tax evasion on February 9th, 2015. Events such as these are covered by many articles and the question is how to find all the articles in different languages that are describing a single event. Generally, events are more specific than general themes as the time component plays an important role – for example, the two wars in Iraq would be considered as separate events.

¹This paper is an extended abstract of an article in the Journal of Artificial Intelligence Research [Rupnik *et al.*, 2016]

As input, we consider a stream of articles in different languages. The task is divided into two parts: detecting events within each language and then linking events across languages. In this paper we address the second step.

We identify events with **clusters** of documents that report on them. Our approach to link clusters across languages combines two ingredients: a cross-lingual document similarity measure, which can be interpreted as a language independent topic model, and semantic annotation of documents, which enables an alternative way to comparing documents. Since this work represents a complicated pipeline, we concentrate on these two specific elements. Overall, the approach should be considered from a systems' perspective (considering the system as a whole) rather than considering these problems in isolation.

The first ingredient of our approach to link clusters across languages represents a continuation of previous work [Rupnik *et al.*, 2011a; 2012; 2011b; Muhič *et al.*, 2012] where we explored representations of documents which were valid over multiple languages. The representations could be interpreted as multilingual topics, which were then used as proxies to compute cross-lingual similarities between documents. To learn the representations, we use Wikipedia as a training corpus. Significantly, we do not only consider the major or *hub* languages such as English, German, French, etc. which have significant overlap in article coverage, but also smaller languages (in terms of number of Wikipedia articles) such as Slovenian and Hindi, which may have a negligible overlap in article coverage. We can then define a similarity between any two articles regardless of language, which allows us to cluster the articles according to topic. The underlying assumption is that articles describing the same event are similar and will therefore be put into the same cluster.

Using the similarity function, we present an algorithm for linking events/clusters across languages. We pose the task as a classification problem based on several sets of features. In addition to these features, cross-lingual similarity is also used to quickly identify a small list of potential linking candidates for each cluster. This greatly increases the scalability of the system.

We will first give an overview of the pipeline in Section 2, then present our approach to cross-lingual similarity computation in Section 3 and finally present our approach to linking clusters in Section 4.

This paper presents a summary of previous results (see [Rupnik *et al.*, 2016]) and omits certain aspects, such as related work and experimental section.

2 Pipeline

We base our techniques of cross-lingual event linking on an online system for detection of world events, called Event Registry [Leban *et al.*, 2014b; 2014a]. Event Registry is a repository of events, where events are automatically identified by analyzing news articles that are collected from numerous news outlets all over the world. We will now briefly describe the main components.

The collection of the news articles is performed using the Newsfeed service [Trampuš and Novak, 2012]. Collected articles are first semantically annotated by identifying mentions of relevant concepts – either entities or important keywords. The disambiguation and entity linking of the concepts is done using Wikipedia as the main knowledge base. The algorithm for semantic annotation uses machine learning to detect significant terms within unstructured text and link them to the appropriate Wikipedia articles. The details are reported by Milne and Witten 2008 and Zhang and Rettinger 2014a.

As the next step, an online clustering algorithm [Brank *et al.*, 2014] is applied to the articles in order to identify groups of articles that are discussing the same event. For each new article, the clustering algorithm determines if the article should be assigned to some existing cluster or into a new cluster.

Once the number of articles in a cluster reaches a threshold (which is a language dependent parameter), we assume that the articles in the cluster are describing an event. By analyzing the articles, we extract the main information about the event, such as the event location, date, most relevant entities and keywords, etc.

Since articles in a cluster are in a single language, we also want to identify any other existing clusters that report about the same event in other languages and join these clusters into the same event. This task is performed using a classification approach which is the second major contribution of this paper. It is described in detail in Section 4.

3 Cross-Lingual Document Similarity

Document similarity is an important component in techniques from text mining and natural language processing. Many techniques use the similarity as a black box, e.g., a kernel in Support Vector Machines. Comparison of documents (or other types of text snippets) in a monolingual setting is a well-studied problem in the field of information retrieval [Salton and Buckley, 1988].

Within each language we represent documents using the standard vector space model [Salton and Buckley, 1988] with *Term Frequency Inverse Document Frequency (TFIDF)* weights. The vector space models are obtained for each language separately with varying dimensionality.

We will start with introducing some notation.

3.1 Notation

The cross-lingual similarity models presented in this paper are based on comparable corpora. A *comparable corpus* is

a collection of documents in multiple languages, with alignment between documents that are of the same topic, or even a rough translation of each other. Wikipedia is an example of a comparable corpus, where a specific entry can be described in multiple languages (e.g., “Berlin” is currently described in 222 languages). News articles represent another example, where the same event can be described by newspapers in several languages.

More formally, a *multilingual document* $d = (u_1, \dots, u_m)$ is a tuple of m documents on the same topic (comparable), where u_i is the document written in language i . Note that an individual document u_i can be an empty document (missing resource) and each d must contain at least **two nonempty documents**. This means that in our analysis we discard strictly monolingual documents for which no cross-lingual information is available. A comparable corpus $D = d_1, \dots, d_s$ is a collection of s multilingual documents. By using the vector space model, we can represent D as a set of m matrices X_1, \dots, X_m , where $X_i \in \mathbb{R}^{n_i \times s}$ is the matrix corresponding to the language i and n_i is the vocabulary size of language i . Furthermore, let X_i^ℓ denote the ℓ -th column of matrix X_i and the matrices respect the document alignment - the vector X_i^ℓ corresponds to the TFIDF vector of the i -th component of multilingual document d_ℓ . We use N to denote the total row dimension of X , i.e., $N := \sum_{i=1}^m n_i$.

3.2 Hub Language Based Embedding

Building cross-lingual similarity models based on comparable corpora is challenging for two main reasons. The first problem is related to missing alignment data: when a number of languages is large, the dataset of documents that cover all languages is small (or may even be empty). Even if only two languages are considered, the set of aligned documents can be small (an extreme example is given by the Piedmontese and Hindi Wikipedias where no inter-language links are available), in which case none of the methods presented so far are applicable. The second challenge is scale - the data is high-dimensional (many languages with hundreds of thousands of features per language) and the number of multilingual documents may be large (over one million in case of Wikipedia).

The embedding we consider is based on a generalization of Canonical Correlation Analysis (CCA) [Hotelling, 1935] to more than two views, introduced by Kettenring 1971, namely the Sum of Squared Correlations SSCOR, which we will state formally later in this section. Our approach exploits a certain characteristic of the data, namely the *hub language* characteristic (see below) in two ways: to reduce the dimensionality of the data and to simplify the optimization problem.

Hub Language Characteristic

In the case of Wikipedia, we observed that even though the training resources are scarce for certain language pairs, there often exists indirect training data. By considering a third language, which has training data with both languages in the pair, we can use the composition of learned maps as a proxy. We refer to this third language as a hub language.

A *hub language* is a language with a high proportion of non-empty documents in $D = \{d_1, \dots, d_\ell\}$. As we have mentioned, we only focus on multilingual documents that include

at least two languages. The prototypical example in the case of Wikipedia is English.

We use the following notation to define subsets of the multilingual comparable corpus: let $a(i, j)$ denote the index set of all multilingual documents with non-missing data for the i -th and j -th language:

$$a(i, j) = \{k \mid d_k = (u_1, \dots, u_m), u_i \neq \emptyset, u_j \neq \emptyset\},$$

and let $a(i)$ denote the index set of all multilingual documents with non missing data for the i -th language.

We now describe a two step approach to building a cross-lingual similarity matrix. The first part is related to Singular Value Decomposition (SVD), also known as Latent Semantic Indexing (LSI) [Deerwester *et al.*, 1990], and reduces the dimensionality of the data. The second step refines the linear mappings and optimizes the linear dependence between data.

Step 1: Hub Language Based Dimensionality Reduction

The first step in our method is to project X_1, \dots, X_m to lower-dimensional spaces without destroying the cross-lingual structure. Treating the nonzero columns of X_i as observation vectors sampled from an underlying distribution $\mathcal{X}_i \in V_i = \mathbb{R}^{n_i}$, we can analyze the empirical cross-covariance matrices:

$$C_{i,j} = \frac{1}{|a(i,j)| - 1} \sum_{\ell \in a(i,j)} (X_i^\ell - c_i) \cdot (X_j^\ell - c_j)^T,$$

where $c_i = \frac{1}{a_i} \sum_{\ell \in a(i)} X_i^\ell$. By finding low-rank approximations of $C_{i,j}$ we can identify the subspaces of V_i and V_j that are relevant for extracting linear patterns between \mathcal{X}_i and \mathcal{X}_j . Let X_1 represent the hub language corpus matrix. The SVD approach to finding the subspaces is to perform the singular value decomposition on the full $N \times N$ covariance matrix composed of blocks $C_{i,j}$. If $|a(i,j)|$ is small for many language pairs (as it is in the case of Wikipedia), then many empirical estimates $C_{i,j}$ are unreliable, which can result in overfitting. For this reason, we perform the truncated singular value decomposition on the matrix $C = [C_{1,2} \dots C_{1,m}] \approx USV^T$, where $U \in \mathbb{R}^{n_1 \times k}$, $S \in \mathbb{R}^{k \times k}$, $V \in \mathbb{R}^{(\sum_{i=2}^m n_i) \times k}$. We split the matrix V vertically in blocks with n_2, \dots, n_m rows: $V = [V_2^T \dots V_m^T]^T$. Note that columns of U are orthogonal but columns in each V_i are not (columns of V are orthogonal). Let $V_1 := U$. We proceed by reducing the dimensionality of each X_i by setting: $Y_i = V_i^T \cdot X_i$, where $Y_i \in \mathbb{R}^{k \times N}$. To summarize, the first step reduces the dimensionality of the data, but optimizes only the hub language related cross-covariance blocks.

Step 2: Simplifying and Solving SSCOR.

The second step involves solving a generalized version of canonical correlation analysis on the matrices Y_i in order to find the mappings P_i . The approach is based on the sum of squares of correlations formulation by Kettenring 1971, where we consider only correlations between pairs $(Y_1, Y_i), i > 1$ due to the hub language problem characteristic. We will present the original unconstrained optimization problem, then a constrained formulation based on the hub language problem characteristic. Then we will simplify the constraints and reformulate the problem as an eigenvalue problem by using Lagrange multipliers.

The original sum of squared correlations is formulated as an unconstrained problem:

$$\underset{w_i \in \mathbb{R}^k}{\text{maximize}} \sum_{i < j}^m \rho(w_i^T Y_i, w_j^T Y_j)^2.$$

We solve a similar problem by restricting $i = 1$ and omitting the optimization over non-hub language pairs. Let $D_{i,i} \in \mathbb{R}^{k \times k}$ denote the empirical covariance of \mathcal{Y}_i and $D_{i,j}$ denote the empirical cross-covariance computed based on \mathcal{Y}_i and \mathcal{Y}_j . We solve the following constrained (unit variance constraints) optimization problem:

$$\begin{aligned} \underset{w_i \in \mathbb{R}^k}{\text{maximize}} \quad & \sum_{i=2}^m (w_1^T D_{1,i} w_i)^2 \\ \text{subject to} \quad & w_i^T D_{i,i} w_i = 1, \quad \forall i = 1, \dots, m. \end{aligned} \quad (1)$$

The particular problem formulation coupled with the hub language assumption can be reduced to a low-dimensional eigenvalue problem using standard lagrangian multiplier techniques and algebraic manipulation. The solution is found for the hub language view, and other solutions for other languages are obtained by solving low-dimensional linear systems. For details as well as empirical results refer to [Rupnik *et al.*, 2016].

4 Cross-Lingual Event Linking

The problem of cross-lingual event linking is to match monolingual clusters of news articles that describe the same event across languages. For example, we want to match a cluster of Spanish news articles and a cluster of English news articles that both describe the same earthquake.

In order to identify clusters that are equivalent to cluster c_i , we have developed a two-stage algorithm. For a cluster c_i , we first efficiently identify a small set of candidate clusters and then find those clusters among the candidates, which are equivalent to c_i .

The details of the first step are described in Algorithm 1. The algorithm begins by individually inspecting each news article a_i in the cluster c_i . Using our proposed method to computing cross-lingual document similarity, it identifies the 10 (hand tuned) most similar news articles to a_i in each language $\ell \in L$. For each similar article a_j , we identify its corresponding cluster c_j and add it to the set of candidates. The set of candidate clusters obtained in this way is several orders of magnitude smaller than the number of all clusters, and at most linear with respect to the number of news articles in cluster c_i . In practice, clusters contain highly related articles and as such similar articles from other languages mostly fall in only a few candidate clusters.

The second stage of the algorithm determines which (if any) of the candidate clusters are equivalent to c_i . We treat this task as a supervised learning problem. For each candidate cluster $c_j \in \mathcal{C}$, we compute a vector of learning features that should be indicative of whether c_i and c_j are equivalent or not and apply a binary classification model that predicts if the clusters are equivalent or not. The classification algorithm that we used to train a model was a linear Support Vector Machine (SVM) method [Shawe-Taylor and Cristianini, 2004].

```

input: test cluster  $c_i$ , a set of clusters  $C_\ell$  for each
        language  $\ell \in L$ 
output: a set of clusters  $C$  that are potentially equivalent
        to  $c_i$ 
 $C \leftarrow \{\}$ ;
for article  $a_i \in c_i$  do
  for language  $\ell \in L$  do
    /* use hub CCA to find 10 most
       similar articles to article  $a_i$ 
       in language  $\ell$  */
     $SimilarArticles =$ 
       $getCCASimilarArticles(a_i, \ell)$ ;
    for article  $a_j \in SimilarArticles$  do
      /* find cluster  $c_j$  to which
         article  $a_j$  is assigned to
         */
       $c_j \leftarrow c$ , such that  $c \in C_\ell$  and  $a_j \in c$ ;
      /* add cluster  $c_j$  to the set
         of candidates  $C$  */
       $C \leftarrow C \cup \{c_j\}$ ;
    end
  end
end

```

Algorithm 1: Algorithm for identifying candidate clusters C that are potentially equivalent to c_i

We use three groups of features to describe cluster pair (c_i, c_j) . The first group is based on **cross-lingual article links**, which are derived using cross-lingual similarity: each news article a_i is linked with its 10-nearest neighbors articles from all other languages (10 per each language). The group contains the following features:

- **linkCount** is the number of times any of the news articles from c_j is among 10-nearest neighbors for articles from c_i . In other words, it is the number of times an article from c_i has a very similar article (i.e., is among 10 most similar) in c_j .
- **avgSimScore** is the average similarity score of the links, as identified for **linkCount**, between the two clusters.

The second group are **concept-related features**. Articles that are imported into Event Registry are annotated by disambiguating mentioned *entities* and *keywords* to the corresponding Wikipedia pages [Zhang and Rettinger, 2014b]. Whenever Barack Obama is, for example, mentioned in the article, the article is annotated with a link to his Wikipedia page. In the same way, all mentions of entities (people, locations, organizations) and ordinary keywords (e.g., bank, tax, ebola, plane, company) are annotated. Although the Spanish article about Obama will be annotated with his Spanish version of the Wikipedia page, in many cases we can link the Wikipedia pages to their English versions. This can be done since Wikipedia itself provides information regarding which pages in different languages represent the same concept/entity. By analyzing all the articles in clusters c_i and c_j , we can identify the most relevant entities and keywords for each cluster. Additionally, we can also assign weights to the

concepts based on how frequently they occur in the articles in the cluster. From the list of relevant concepts and corresponding weights, we consider the following features:

- **entityCosSim** is the cosine similarity between vectors of entities from clusters c_i and c_j .
- **keywordCosSim** is the cosine similarity between vectors of keywords from clusters c_i and c_j .
- **entityJaccardSim** is Jaccard similarity coefficient [Levandowsky and Winter, 1971] between sets of entities from clusters c_i and c_j .
- **keywordJaccardSim** is Jaccard similarity coefficient between sets of keywords from clusters c_i and c_j .

The last group of features contains three **miscellaneous features** that seem discriminative but are unrelated to the previous two groups:

- **hasSameLocation** feature is a boolean variable that is true when the location of the event in both clusters is the same. The location of events is estimated by considering the locations mentioned in the articles that form a cluster and is provided by Event Registry.
- **timeDiff** is the absolute difference in hours between the two events. The publication time and date of the events is computed as the average publication time and date of all the articles and is provided by Event Registry.
- **sharedDates** is determined as the Jaccard similarity coefficient between sets of date mentions extracted from articles. We use extracted mentions of dates provided by Event Registry, which uses an extensive set of regular expressions to detect and normalize mentions of dates in different forms.

For empirical results on cluster linking, refer to [Rupnik *et al.*, 2016].

5 Conclusions

In this paper we have summarized our results [Rupnik *et al.*, 2016] on cross-lingual system for linking events in different languages focusing on two main aspects: building cross-lingual similarity functions and applying them to cluster linking.

Acknowledgments

The authors gratefully acknowledge that the funding for this work was provided by the projects X-LIKE (ICT-257790-STREP), MultilingualWeb (PSP-2009.5.2 Agr.# 250500), TransLectures (FP7-ICT-2011-7), PlanetData (ICT-257641-NoE), RENDER (ICT-257790-STREP), XLime (FP7-ICT-611346), and META-NET (ICT-249119-NoE). This paper re-uses and summarizes parts of the work [Rupnik *et al.*, 2016] and the authors acknowledge the Journal of Artificial Intelligence Research for allowing this presentation.

References

[Brank *et al.*, 2014] Janez Brank, Gregor Leban, and Marko Grobelnik. A high-performance multithreaded approach

- for clustering a stream of documents. In *Proceedings of the 17th International Multiconference Information Society 2014, Volume E, Ljubljana, Slovenia*, pages 5–8, 2014.
- [Deerwester *et al.*, 1990] S. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407, 1990.
- [Hotelling, 1935] Harold Hotelling. The most predictable criterion. *Journal of educational Psychology*, 26(2):139, 1935.
- [Kettenring, 1971] Jon R. Kettenring. Canonical analysis of several sets of variables. *Biometrika*, 58:433–45, 1971.
- [Leban *et al.*, 2014a] Gregor Leban, Blaz Fortuna, Janez Brank, and Marko Grobelnik. Cross-lingual detection of world events from news articles. In *Proceedings of the 13th International Semantic Web Conference*, pages 21–24, Riva del Garda - Trentino, Italy, 2014.
- [Leban *et al.*, 2014b] Gregor Leban, Blaz Fortuna, Janez Brank, and Marko Grobelnik. Event registry: Learning about world events from news. In *Proceedings of the Companion Publication of the 23rd International Conference on World Wide Web Companion*, WWW Companion '14, pages 107–110, Seoul, Republic of Korea, 2014. International World Wide Web Conferences Steering Committee.
- [Levandowsky and Winter, 1971] Michael Levandowsky and David Winter. Distance between sets. *Nature*, 234(5323):34–35, 11 1971.
- [Milne and Witten, 2008] David Milne and Ian H. Witten. Learning to link with wikipedia. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management, CIKM '08*, pages 509–518, New York, NY, USA, 2008. ACM.
- [Muhič *et al.*, 2012] Andrej Muhič, Jan Rupnik, and Primož Škraba. Cross-lingual document similarity. In *Information Technology Interfaces (ITI), Proceedings of the ITI 2012 34th International Conference on*, pages 387–392, Cavtat / Dubrovnik, Croatia, 2012. IEEE.
- [Rupnik *et al.*, 2011a] Jan Rupnik, Andrej Muhic, and Primoz Skraba. Low-rank approximations for large, multilingual data. *Low Rank Approximation and Sparse Representation, Neural Information Processing Systems 2011 Workshop*, 2011.
- [Rupnik *et al.*, 2011b] Jan Rupnik, Andrej Muhic, and Primoz Skraba. Spanning spaces: Learning cross-lingual similarities. *Beyond Mahalanobis: Supervised Large-Scale Learning of Similarity, Neural Information Processing Systems 2011 Workshop*, 2011.
- [Rupnik *et al.*, 2012] Jan Rupnik, Andrej Muhic, and Primoz Skraba. Multilingual document retrieval through hub languages. In *Proceedings of the 15th Multiconference on Information Society 2012 (IS-2012)*, pages 201–204, Ljubljana, Slovenia, 2012.
- [Rupnik *et al.*, 2016] Jan Rupnik, Andrej Muhič, Gregor Leban, Primož Škraba, Blaž Fortuna, and Marko Grobelnik. News Across Languages - Cross-Lingual Document Similarity and Event Tracking. *Journal of Artificial Intelligence Research, JAIR*, 55:283 – 316, 2016.
- [Salton and Buckley, 1988] Gerard Salton and Christopher Buckley. Term-weighting approaches in automatic text retrieval. volume 24, pages 513–523. Elsevier, 1988.
- [Shawe-Taylor and Cristianini, 2004] John Shawe-Taylor and Nello Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.
- [Trampuš and Novak, 2012] Mitja Trampuš and Blaž Novak. The internals of an aggregated web news feed. In *Proceedings of 15th Multiconference on Information Society 2012 (IS-2012)*, pages 221–224, Ljubljana, Slovenia, 2012.
- [Zhang and Rettinger, 2014a] Lei Zhang and Achim Rettinger. Semantic annotation, analysis and comparison: A multilingual and cross-lingual text analytics toolkit. In *Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2014)*, pages 13–16, Gothenburg, Sweden, April 2014. Association for Computational Linguistics.
- [Zhang and Rettinger, 2014b] Lei Zhang and Achim Rettinger. X-lisa: Cross-lingual semantic annotation. *Proceedings of the Very Large Data Bases (VLDB) Endowment*, 7(13):1693–1696, August 2014.