Using Semantic Data to Improve Cross–Lingual Linking of Article Clusters

Evgenia Belyaeva[a,b,*], Aljaž Košmerlj[a], Andrej Muhič[a], Jan Rupnik[a], Flavio Fuart[a,*]

[a] *Jožef Stefan Institute, Jamova cesta 39, 1000 Ljubljana, Slovenia*
[b] *JSI International Postgraduate School, Jamova cesta 39, 1000 Ljubljana, Slovenia*

Abstract

This paper presents a system that uses semantic data to improve cross–lingual linking of news article clusters. Two approaches are compared. The first based on two different Canonical Correlation Analysis (CCA) feature vector definitions: MAX–CCA and SUM–CCA, whereas the second one has been developed using a better-performed CCA approach in combination with Entity vectors. The aim of the comparison was to determine whether taking into account the semantic aspect of news increases performance and improves linking. Evaluations of the aforementioned techniques on a news corpus, both against Google News and manual, revealed good performance of our system. The overall gain in precision and recall when using entity vectors was significant.

Keywords: semantic data, natural language processing, cross–linguality, canonical correlation analysis

1. Introduction

Automatic linking of clusters across languages is an important task in news monitoring applications. It is based on the core idea of article clusters in different languages sharing the same content to one another. Linking connects related clusters across all language pairs involved i.e. clusters dealing with a common subject matter involving the same named entities (actors, organizations and locations) and time–frame.

Traditionally, news monitoring systems were monolingual, but needs from the user community, technological and scientific progress lead to an increasing number of news monitoring services to add multilingual and cross-lingual features to existing monolingual components [1]. The main goal of automatic cross–lingual linking of clusters is to interconnect many related news items that are reported by numerous news outlets in different languages. It also enables knowledge from around the world to be monitored and aggregated [2]. Some research studies have been done to improve the quality of cross–lingual linking of clusters through such tasks as: classification [3], summarization [4], machine learning translations [5], similarity analysis [6], the use of dictionaries and information extraction [7].

Thus, in the news monitoring domain, the basic task is to create monolingual clusters and then, when language-pair resources become available, cross-lingual features are gradually introduced. This approach allows for a wide range of monolingual features, like clustering, classification, entity extraction and sentiment detection, to be introduced even when no resources for specific language pairs are available. This is probably the main reason why in the news monitoring domain cross-linking of monolingual clusters is preferred over full cross-lingual clustering. For example, the EMM system [1] gathers news in 20 languages with 190 implemented language pairs. Newsfeed [8] covers 35 languages, with 12 languages annotated with entities and 276 language pairs covered with article cross-link information. In this work we implemented cross-linking of clusters for three language pairs.

The approach described in this paper is a crucial component of an extension of the iDiversiNews application [9], which will allow users to explore related news stories in different languages and discover new aspects of a story. A similar approach is used in the Event Registry application [10] for detection of events from multilingual news article

*Corresponding authors
Email addresses: jenya.belyaeva@ijs.si (Evgenia Belyaeva), flavio.fuart@ijs.si (Flavio Fuart)

data.

In order to improve cross–lingual linking of clusters, we have proposed a new semantic approach for English, German and Spanish languages based on Canonical Correlation Analysis (CCA) technique in combination with named entity vectors. Instead of the typical approaches of machine translations and cross–lingual information retrieval, we take into consideration the semantic part of news i.e. named entities. The use of entities is explained by their amount of very vital information that is contributed to the news. They largely define the main topic of an article [11] and appear to play an important role when studying cross–linguality and multilingual information retrieval since research shows 30% of content–important words in a journalistic text are proper nouns [12]. We also show that the presented approach outperformed the results obtained by simply using two different CCA feature vectors.

To our knowledge, there is no existing database of curated and reliable data that can be used as the golden standard for testing cross–lingual linking of news clusters, thus we have performed manual evaluation, which we consider an important contribution, since human judgement is necessary to evaluate the performance of any system.

This paper is organized as follows: Section 2 introduces the used data sources, algorithms and named entity recognition system. In section 3 we describe our corpus. Evaluation results are detailed in Section 4. Section 5 concludes the paper and announces future work.

2. Semantic analysis and clustering

The news articles we used were aggregated by the JSI Newsfeed[1] – a clean, continuous, real–time aggregated stream of semantically enriched articles from more than 1900 RSS–enabled sites across the world in all major languages with around 300 000 articles per day [8]. For our purposes the articles were processed by a linguistic and semantic analysis pipeline [2], which provides the semantic annotations used to support cross–linking.

The semantic annotation tool developed as a part of the XLike project consists of three main approaches for multi–lingual annotation: *named entity recognition* based on finding corresponding Wikipedia pages in the target language of previously detected named entities; *Wikipedia Miner Wikifier* – a simple approach based on Wikipedia articles that tries to detect similar phrases and links in any document of the same language as Wikipedia articles; *cross–lingual semantic analysis* that links no longer detected word phrases to corresponding Wikipedia pages, but articles by topic or concepts, as described in [13, 14]. The main aim of the cross–lingual semantic annotation is to find links between entities mentioned in articles in the source language with their corresponding entities in the target language [15, 16].

2.1. CCA–like approach for document comparison

There are many possible approaches to cross-lingual similarity computation in the literature: translation based [17], Google Translate [2], probabilistic topic models based[18, 19], approaches related to classification [20] and factorization based approaches[21, 22] which also cover our proposed approach. There are several aspects of the methods to consider when implementing a cross-lingual similarity component: scalability of training to large corpora, similarity computation speed and cost, as well as the ease of implementation and ease of use (the number of parameters to tune).

Two closely related approaches, [23] and [24], also rely on Wikipedia to compute document similarity. The first focuses on comparing short informal texts (tweets) is less applicable in our case, since we focus on news texts, which are longer and more formal. The second approach (CL-ESA) could be seen as a baseline, where no subspace computation is needed and each aligned pair represents a latent dimension. CL-ESA is easier to interpret, but performs worse than CL-LSI according to the authors, where our method can be seen as enhancing CL-LSI.

Using the vector space model we can represent documents written in a given language as vectors in a vector space[25], whose dimension is the number of terms (typically words, word n–grams, character n–grams, etc.). When dealing with more than one language, this results in several vector spaces with varying dimensionalities, members of which we need to compare.

The goal of the Canonical Correlation Analysis (CCA) based approach is to find a set of mappings from language–specific vector spaces into a common

[1] http://newsfeed.ijs.si/

[2] https://translate.google.com/

semantic vector space, in which standard machine learning tools apply. A good set of mappings should preserve the document similarities within each language, and work well across languages: a document and its translation should be similar when mapped to the common vector space. When given training data, CCA[26] is a method that can be applied to determining such linear mappings between two languages by finding aligned subspaces in each vector space that maximize a measure of linear dependence. Cross–Lingual Latent Semantic Indexing (CL–LSI)[22] is an alternative approach. Our approach is related to both.

The mappings can be learned if we have access to a large comparable corpus, in which the training data consists of tuples of documents in different languages that cover similar topics. For this purpose we used the Wikipedia[27] – a multilingual collection of interlinked articles, that are comparable but not necessarily translations of each other. Most of our analysis is based on low rank decompositions of cross–covariance matrices, their estimation based on the aligned corpus matrices.

This problem has some specific features that render standard approaches intractable: the numbers of dimensions and samples are high, which leads to overfitting issues with CCA. Another aspect, specific to the training dataset, is the fact that the majority of alignments includes English documents – this means that term cross–covariance matrices between English and other languages can be well estimated more reliably.

The above issues are overcome by pre-mapping the data, using a method very similar to CL-LSI, and then refining the mappings in a lower dimensional setting. The refinement is based on analysing only cross–covariance matrices related to the English language, which enables us to solve effectively a generalisation of CCA to more than two languages (sum of squares of correlations, REF) as a lower–dimensional eigenvalue problem. For details refer to [28].

2.2. Clustering algorithm

Our clustering algorithm is an adaptation and extension of the Incremental Clustering of News Reports algorithm described by Azzopardi et al. [29]. The basic algorithm performs well when evaluated against the Google News corpus, thus deemed to be tailored to cluster high–volume time–dependent text streams, which is suitable for our needs.

Azzopardi et al. create a BoW TF–IDF vector for each news article and calculate cosine similarity against centroids of already existing clusters. The new article is added to the first cluster that matches above a predefined similarity threshold (typically between 0.3 and 0.4). If there is no match, then a new cluster is formed.

We have tested and adapted the basic algorithm in order to process reliably the Newsfeed data flow, which is considerably higher compared to test performed with the original algorithm. Additionally, it is necessary to take into account that the news stream may contain material of a lower quality, obtained from a moderated list of sources. From our point of view Azzopardi et al. fail to provide a satisfactory explanation of cluster ageing, so we have determined an optimal time interval for keeping the clusters alive.

Even with no extension, an evaluation would be essential, because the original implementation was tested against the English news corpus.

Our extension of the original algorithm is explained through the list below, where each feature can be tuned with a set of parameters:

similarity thresholds The last level defines the similarity threshold for clusters, i.e. what is the minimum similarity for an item to be added to a cluster. Other levels are used for a rough classification of clusters for performance reasons. Typically we use a hierarchy of two thresholds: 0.005 and 0.4.

stop words threshold Word document–frequency tables were created for long–term time periods. This threshold defines the percentage of top words used as stop words. Typically 0,0002.

word limit Only the initial N words of a document are used by the algorithm. This parameter defines the number of words used. The parameter can also be set to *all*.

smallest cluster size Smaller clusters are discarded.

cluster expiry in hours If there are no new news items in the defined period then the cluster is deemed as final, i.e. no more updates are possible.

initial buffer The first clustering run is performed after enough items have been collected.

2.3. Cross–linking clusters

Consequently, in order to identify clusters describing the same news item in different languages, two CCA feature vector definitions, MAX-CCA and SUM-CCA, are used. In both cases, for each cluster a set of feature vectors is created, one per language. Vector components are newsfeed article unique identifiers from CCA fields of all news items in the cluster.

For example, suppose an English cluster contains news items E1 and E2 with their corresponding CCA similarities for Spanish articles (S1, S2, S3) as listed in Table 1. The corresponding feature vector is given in the third column. Due to the large number of articles feature vectors are very large and sparse.

Article ID	*CCA – spa*	*MAX-CCA*	*SUM-CCA*
E1	S1, 0.2	S1, 0.2	S1, 0.2
	S2, 0.8	S2, 0.8	S2, 1.0=0.8+0.2
E2	S2, 0.2	S3, 0.5	S3, 0.5
	S3, 0.5		

Table 1: Example: news items and cluster CCA-MAX and CCA-SUM feature vectors

Additionally, for each cluster a feature vector of unique identifiers (i.e. English Wikipedia links) of annotated entities is constructed. It consists of entities that appear in at least 20% of the articles. The values in the feature vectors are percentages. For example, if Robin Williams appears in 30% of articles, the vector component for the unique identifier http://en.wikipedia.org/wiki/Robin_Williams would be 30.

It is important to note that our system cross–links clusters over a time window of 24h. Cosine similarity is used to compare their CCA vectors for a specific language pair and Entity vectors. We manually evaluate all links that have at least one of the three vectors over the threshold of 0.2 and discard all other pairs. Keeping a large number of candidates for manual evaluation allows us to get a rough estimate for recall as well.

It would have been possible to combine the entity information with the output of the CCA-like similarity measure into a single metric to directly cluster the articles across all languages. However, clustering articles in their respective languages first allows us to detect the level of reporting for each language separately, which is important as the volume of reporting across languages is unbalanced as shown in Table 2. This is especially important for "small volume" languages we process with our system, like Slovenian or Catalan.

3. Corpus and news items statistics

The main goal was to determine whether named entities had any influence on the overall quality of cross–lingual cluster linking results. In order to measure this effect, we first describe our data and calculate automatic and manual evaluation of clusters to ensure that the clusters and its parameters used are good enough to be used further for our main evaluation i.e. the manual evaluation of cross–lingual links – the main contribution of this paper.

3.1. News items

To explore the cross–lingual clusters, using several techniques, English, German and Spanish news items published over the period from 19/05/2014 to 01/06/2014 have been processed. In order to evaluate the performance of our clustering algorithm against the commercial, but publicly available news navigation and analysis application Google News stories, we have focused on the period between 28/05/2014 and 30/05/2014, since there existed a satisfactory number of Google News references for the above–mentioned period. We selected Wednesday, 28/05/2014 for the manual evaluation of clusters and cross–links. The numbers of news items used are listed in Table 2.

timespan	*dataset*	*deu*	*eng*	*spa*	*total*
entire period	Newsfeed	70 245	275 349	49 497	395 091
	Google	8 648	44 028	6 961	59 637
28/05/2014	Newsfeed	11 477	56 350	9 213	77 040
	Google	1 919	9 696	1 538	13 153

Table 2: Numbers of news items used in the evaluation.

3.2. Evaluation against Google News stories

For our evaluation against Google News stories, we explored the whole period from 19/05/2014 to 01/06/2014, despite the fact that Google data was available for a shorter time range. We performed 51 clustering runs with different sets of parameters. For each run and language, standard metrics are used: Speed of execution, Recall, precision, and F1 measure, along with some additional indicators.

3.3. Manual Evaluation

For the second evaluation executed manually, we have selected empirically the most suitable set of parameters for clustering per language. Parameters were set as listed in Table 3.

parameter	*deu*	*eng*	*spa*
thresholds	(0, 0.4)	(0, 0.4)	(0, 0.4)
Smallest cluster size	3	5	3
Cluster expiry (h)	24	6	24
initial buffer size (h)	1 000	1 000	1 000

Table 3: Clustering parameters used for manual evaluation

After the clustering was performed, all clusters that started on 28/05/2014 were provided to an expert, who evaluated the quality of each cluster separately, giving the number of correct/wrong articles in the cluster. This allows the overall clustering precision to be calculated by weighting the precision for each cluster by its size, using the following equations:

$$P_c = \frac{TP}{TP + FP} \qquad P_{overall} = \frac{\sum_{c \in C} |c| \times P_c}{\sum_{c \in C} |c|}$$

3.4. Evaluation of cross–links

In the next evaluation step, we checked the multilingual news story links in the following two phases. First, we analysed which feature vector between MAX–CCA and SUM–CCA performs better. We compared the results and used the better performed approach for our next step in the evaluation, where we aim at checking whether taking into account the semantic aspect of news would increase the overall performance of the system.

The expert was given a list of inter–cluster links. We provided the following data for each cluster: language, title, link to the cluster web page and its size. The evaluator then scored each relation with the following scores: 1 – correct link, 2 – partially correct link and 3 – wrong link. Partially correct cross–linking implies links, in one or another language, that may mention the main topic of the story, but have an additional information or a different angle of the coverage.

We define two modes of evaluation: strict and weak, which also reflects possible use–case scenarios of the system. Strict evaluation means that a link is correct only when annotated with 1. This would reflect a use case, in which the story topic, information provided and angle of coverage match in both languages. Weak evaluation means that a link is correct when marked with 1 or 2. In that case, it is enough if the clusters are about the same topic. Any additional information or different angles of coverage give us additional insight, thus most users may even prefer "weak" links.

In addition to the manual cross–linking evaluation, we also want to find the optimal threshold values and combination of weights for CCA/Entity vectors. Thus, the result of evaluation would be used to find an optimal combination of parameters W and similarity threshold T.

We calculated precision, recall (from the selected dataset) and F0.5 measure (F–measure giving more importance to precision rather than recall) for each language. Parameters were calculated for threshold settings between 0.2 and 0.6 (step 0.05).

4. Results

In this section, results obtained using the approach described in the Section 3 are presented.

4.1. Clustering – against Google News stories

By performing clustering runs with different parameter values we found an optimal set of parameters that would take into account clustering quality and the speed of execution. We used stop word threshold value 0.0002, smallest cluster size of 1 and initial buffer size of 1000 items. Table 4 shows the F1 measure values obtained for different values of other parameters, namely thresholds, word limit (WL) and cluster expiry in hours (Ex).

For the same set of parameters, we compute the clustering speed, aggregated across all languages. Speed measurements are listed in Table 5.

Except for an outlier (marked with *), all other measurements form a consistent picture. The optimal similarity cutoff is in the range 0.2–0.3 and increasing the cutoff slightly increases clustering speed. Tests show that speed increases considerably (20%–50%) by introducing the two–level threshold hierarchy, whereas F1 measure is not significantly affected by this parameter. Number of words, WL, has almost no effect on F1 score, but the lower value improves clustering speed. As for expiry interval, Ex, clustering gets slower as the interval is increased (more than 2x for 4h compared to 48h), but the optimal value depends on the language. Observing the results, we concluded that for "small" languages we get almost no clusters unless we increase the interval to 24h–48h, while the quality of English clusters drops for intervals longer than 8h–12h.

	Ex:	4		6		8		12	24	48
lang	*WL: thresholds*	200	500	200	500	200	500	200	200	200
deu	(0.1, 0.05)	0.46		0.45		0.45		0.45	0.45	
	(0.2)	0.48	0.48	0.49	0.48	0.49	0.49			
	(0.2, 0.05)	0.48	0.48	0.48	0.48	0.48	0.48	0.49	0.49	0.49
	(0.3)	0.47	0.47	0.48	0.47	0.48	0.47			
	(0.3, 0.05)	0.47	0.47	0.48	0.47	0.48	0.47	0.48	0.49	0.49
	(0.4)	0.44	0.43	0.43	0.43	0.43	0.43			
	(0.4, 0.05)	0.44	0.43	0.43	0.43	0.43	0.43	0.43	0.43	0.44
eng	(0.1, 0.05)	0.35		0.35		0.29				
	(0.2)	0.39	0.4	0.4	0.4	0.37	0.37			
	(0.2, 0.05)	0.39	0.39	0.39	0.4	0.36	0.36	0.35	0.35	0.35
	(0.3)	0.41	0.4	0.41	0.41	0.38	0.38			
	(0.3, 0.05)	0.4	0.4	0.4	0.4	0.38	0.37	0.37	0.36	0.36
	(0.4)	0.4	0.4	0.39	0.39	0.38	0.38			
	(0.4, 0.05)	0.4	0.39	0.39	0.39	0.37	0.37	0.37	0.36	0.35
spa	(0.1, 0.05)	0.46		0.45		0.4				
	(0.2)	0.49	0.49	0.49	0.49	0.48	0.48			
	(0.2, 0.05)	0.49	0.49	0.49	0.48	0.48	0.47	0.46	0.46	0.45
	(0.3)	0.46	0.46	0.45	0.45	0.45	0.45			
	(0.3, 0.05)	0.45	0.45	0.45	0.45	0.45	0.45	0.44	0.43	0.43
	(0.4)	0.4	0.4	0.4	0.4	0.39	0.4			
	(0.4, 0.05)	0.4	0.4	0.4	0.4	0.4	0.39	0.39	0.38	0.38

Table 4: F1 measure for different sets of parameters, per language

Ex:	4		6		8		12	24	48
WL: thresholds	200	500	200	500	200	500	200	200	200
(0.1, 0.05)	27		25		19		16	14	
(0.2)	29	28	23	23	23	22			
(0.2, 0.05)	37	19*	30	29	26	23	19	18	16
(0.3)	26	25	21	21	19	18			
(0.3, 0.05)	38	26	32	30	26	25	23	18	16
(0.4)	25	24	21	19	18	18			
(0.4, 0.05)	39	31	32	31	27	26	24	18	15

Table 5: Clustering speed [item/second]

Measurements show that English gets optimal values in the range 4h–6h, Spanish 4h–8h and German along the whole scale, but with slightly better results in the interval 12h–48h. Optimal threshold values are (0.3, 0.05) and for word limit 200 words. The expiry interval is language–dependent, for English and Spanish at 6h and German 24h.

4.2. Manual evaluation: clustering

For manual analysis we used a dataset of news items and clusters from 28/05/2014. Exact numbers of news items and clusters over languages is listed in Table 6.

We manually evaluated the performance of German, English and Spanish clusters for the same date and obtained the following precision per language: German – 0.93%, English – 0.95% and Spanish – 0.97%. Although we did not attempt to evaluate recall, the evaluator has estimated that there were almost no clusters that could be merged into bigger clusters. We believe that the opposite conclusion would point to low recall. However, the evaluator did not have access to discarded clusters – those with less than 5 articles for English and 3 articles for other languages. The evaluator for all three languages estimated the margin of error, while annotating, was around 5%, thus we estimate that for all languages around 90% of precision was obtained. We also conclude that there is no significant difference among clustering quality across the languages.

language	*news items*	*clusters*
deu	2 685	382
eng	10 404	749
spa	2 583	382
total	15 672	1 513

Table 6: Number of news items and clusters per language. Clusters starting on 28/5/2014.

4.3. Manual evaluation: cross–linking

Manual evaluation of cross–links was performed on the same dataset of clusters from 28/05/2014. In total, we detected 20 101 links, spread across languages as shown in Table 7. Each of the methods produced a different subset of these links. Exact numbers of links over methods are listed in Table 8.

language pair	*links*
eng–deu	7 477
spa–deu	6 621
spa–eng	6 003
total	20 101

Table 7: Number of cross–links for language pairs.

language pair	*SUM–CCA*	*MAX–CCA*	*ENTITY*
eng–deu	950	1 158	6 476
spa–deu	665	794	5 969
spa–eng	802	917	5 378

Table 8: Number of links over threshold (0.2).

The evaluator classified the 20 101 cross–links as described in Section 3.4. The obtained classifications are listed in Table 9.

language pair	*1–match*	*2–partial*	*3–wrong link*	*N/A*
eng–deu	434	222	6 815	6
spa–deu	299	105	6 206	11
spa–eng	527	204	5 266	6

Table 9: Annotated set of links.

We have also compared the two methods, SUM–CCA and MAX–CCA, by their F0.5 score. Table 10 shows the best scores across all tested thresholds, which shows that MAX–CCA consistently outperforms SUM–CCA in strict and weak evaluation.

	strict evaluation		*weak evaluation*	
language pair	*MAX*	*SUM*	*MAX*	*SUM*
eng–deu	0.44	0.36	0.46	0.42
spa–deu	0.41	0.36	0.41	0.37
spa–eng	0.46	0.40	0.49	0.45

Table 10: F0.5 maximum values across all thresholds.

Consequently, we were able to estimate for each language pair the optimal threshold (Table 11) for the MAX–CCA vector, using the weak annotation as our criteria because we believe even weak links may give additional clues to news consumers.

threshold	*eng–deu*	*spa–deu*	*spa–eng*
0.2	0.36	0.33	0.45
0.25	0.41	0.37	**0.49**
0.3	0.45	0.37	**0.49**
0.35	**0.46**	0.40	0.47
0.4	0.43	**0.41**	0.43
0.45	0.39	0.38	0.36
0.5	0.33	0.33	0.28
0.55	0.27	0.26	0.19
0.6	0.19	0.18	0.12

Table 11: Optimal threshold values for the F.05 measure. Annotation scores 1 and 2, CCA–MAX feature vector.

For the best F0.5 measure we set the threshold somewhere between 0.30 and 0.45. In real–world applications, we would opt for higher recall values if our user group would consist of news analysts and higher precision values for the public.

4.4. Manual evaluation: cross–linking with entities

Since proper nouns play an important part in journalistic works, we had identified named entities as a potential indicator of good cross–lingual linking performance. In this subsection, we test if adding this semantic information (i.e. named entities) will improve the performance of our system.

We are trying to optimize values for W (entity score weight) and T (total score threshold) for a given dataset. Again, we will analyse F0.5 scores and precision in order to get a good estimate. The evaluation of MAX–CCA in combination with Entity feature vectors turned to perform much better then by using just CCA. In particular, a significant gain in quality can be observed in Table 12.

	strict evaluation		*weak evaluation*	
language pair	*MAX*	*MAX + ENTITY*	*MAX*	*MAX + ENTITY*
eng–deu	0.37	0.52	0.36	0.55
spa–deu	0.34	0.50	0.36	0.55
spa–eng	0.38	0.58	0.40	0.60

Table 12: F0.5 maximum values across all thresholds for MAX–CCA without and with named entity features.

The optimal weight, W, and threshold, T were determined by exploring the space of their values for each language pair as shown in Tables 13, 14 and 15, where the weight of the entity similarity increases from left to right, taking into account only entities in the last column, where $W = 1.0$. For brevity we omit rows at $T \in \{0.25, 0.35, 0.45, 0.55\}$. The values in bold are maxima over all values.

	W									
T	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
0.20	0.40	0.44	0.43	0.38	0.30	0.22	0.17	0.13	0.11	0.09
0.30	0.48	0.53	**0.55**	0.53	0.46	0.36	0.29	0.23	0.18	0.15
0.40	0.43	0.42	0.43	0.43	0.40	0.36	0.32	0.28	0.25	0.21
0.50	0.32	0.29	0.24	0.20	0.21	0.22	0.25	0.23	0.25	0.24
0.60	0.14	0.10	0.09	0.09	0.08	0.08	0.12	0.14	0.18	0.18

Table 13: F0.5 scores for eng–deu cross–links, weak evaluation, MAX–CCA + ENTITY

	W									
T	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
0.20	0.38	0.39	0.41	0.39	0.29	0.21	0.15	0.11	0.08	0.06
0.30	0.42	0.48	0.54	**0.55**	0.50	0.42	0.33	0.24	0.18	0.14
0.40	0.44	0.44	0.43	0.45	0.39	0.37	0.33	0.32	0.27	0.21
0.50	0.28	0.26	0.21	0.20	0.21	0.22	0.27	0.26	0.24	0.25
0.60	0.14	0.10	0.09	0.06	0.05	0.07	0.10	0.09	0.12	0.14

Table 14: F0.5 scores for spa–deu cross–links, weak evaluation, MAX–CCA + ENTITY

	W									
T	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
0.20	0.49	0.51	0.53	0.48	0.39	0.31	0.24	0.20	0.16	0.14
0.30	0.54	0.58	0.59	**0.60**	0.57	0.49	0.41	0.34	0.28	0.24
0.40	0.44	0.44	0.48	0.50	0.53	0.52	0.49	0.44	0.39	0.34
0.50	0.28	0.28	0.28	0.31	0.33	0.39	0.43	0.43	0.42	0.39
0.60	0.13	0.13	0.12	0.14	0.16	0.19	0.25	0.32	0.36	0.36

Table 15: F0.5 scores for spa–eng cross–links, weak evaluation, MAX–CCA + ENTITY

For evaluated language pairs an optimal combination would be to use T in the range [0.2–0.4] and W [0.3–0.4].

5. Conclusion and future work

A new system has been developed to improve the linking of cross-lingual clusters, based on the use of the semantic part of news i.e. entities. Several techniques were assessed, but the novel approach used proved that the use of named entities (which articles have in common across languages) is a good source that can support linking of multilingual clusters. The evaluation has shown that the results are good, the use of semantic features improved the scores by almost 50%. Thus, using named entities is a promising strategy for improving cross–lingual linking of clusters.

Future work will include taking into account the distribution of the named entities in the clusters, using different named entities such as date expressions, as well as experimenting and adapting our system to different language pairs.

6. Acknowledgments

This work was funded by the European Union through project XLike (FP7-ICT-2011-288342).

References

[1] R. Steinberger, A survey of methods to ease the development of highly multilingual text mining applications, Lang. Resour. Eval. 46 (2) (2012) 155–176.

[2] X. Carreras, L. Padró, L. Zhang, A. Rettinger, Z. Li, E. García-Cuesta, v. Agić, B. Bekavec, B. Fortuna, T. Štajner, Xlike project language analysis services, in: Proceedings of EACL '14: demos, 2014, pp. 9–12.

[3] W.-c. Wong, A.-c. Fu, Incremental document clustering for web page classification, in: Q. Jin, J. Li, N. Zhang, J. Cheng, C. Yu, S. Noguchi (Eds.), Enabling Society with Information Technology, 2002, pp. 101–110.

[4] E. Filatova, Multilingual wikipedia, summarization, and information trustworthiness, in: SIGIR workshop on information access in a multilingual world, 2009.

[5] J. Van Gael, X. Zhu, Correlation clustering for crosslingual link detection, in: Proceedings of IJCAI '07, 2007, pp. 1744–1749.

[6] B. Pouliquen, R. Steinberger, C. Ignat, Automatic linking of similar texts across languages, Recent Advances in Natural Language Processing III. Selected Papers from RANLP 2003 (2003) 307–316.

[7] J. Piskorski, J. Belayeva, M. Atkinson, On refining real–time multilingual news event extraction through deployment of cross–lingual information fusion techniques, in: Proceedings of EISIC '11, 2011, pp. 38–45.

[8] M. Trampuš, B. Novak, The internals of an aggregated web news feed, in: Proceedings of IS '12, 2012.

[9] M. Trampuš, F. Fuart, J. Berčič, D. Rusu, L. Stopar, T. Štajner, (i)diversinews – a stream–based, on–line service for diversified news, in: Proceedings of SiKDD 2013, 2013, pp. 184–187.

[10] G. Leban, B. Fortuna, J. Brank, M. Grobelnik, Event registry: Learning about world events from news, in: Proceedings of WWW Companion '14, 2014, pp. 107–110.

[11] M. Hassel, Exploitation of named entities in automatic text summarization for swedish, in: Proceedings of NODALIDA 03, 2003.

[12] F. C. Gey, Research to improve cross–language retrieval – position paper for clef, in: Revised Papers from the CLEF '00 Workshop, 2001, pp. 83–88.

[13] E. Gabrilovich, S. Markovitch, Computing semantic relatedness using wikipedia–based explicit semantic analysis, in: Proceedings of IJCAI '07, 2007, pp. 1606–1611.

[14] Z. Wang, J. Li, J. Tang, Boosting cross–lingual knowledge linking via concept annotation, in: Proceedings of IJCAI '13, 2013, pp. 2733–2739.

[15] J. Rupnik, A. Muhič, P. Škraba, Cross–lingual document analysis, in: Proceedings of MLCOGS 2013, 2013.

[16] L. Zhang, A. Rettinger, M. Frber, M. Tadi, A comparative evaluation of cross–lingual text annotation techniques, in: Information Access Evaluation. Multilinguality, Multimodality, and Visualization, Vol. 8138, 2013, pp. 124–135.

[17] H. Hoang, A. Birch, C. Callison-burch, R. Zens, R. Aachen, A. Constantin, M. Federico, N. Bertoldi, C. Dyer, B. Cowan, W. Shen, C. Moran, O. Bojar, Moses: Open source toolkit for statistical machine translation, 2007, pp. 177–180.

[18] D. Mimno, H. M. Wallach, J. Naradowsky, D. A. Smith, A. McCallum, Polylingual topic models, in: Proceedings of EMNLP '09, 2009, pp. 880–889.

[19] D. Zhang, Q. Mei, C. Zhai, Cross-lingual latent topic extraction, in: Proceedings of ACL '10', 2010, pp. 1128–1137.

[20] X. Wan, Co-training for cross-lingual sentiment classification, in: Proceedings of ACL '09, 2009, pp. 235–243.

[21] M. Xiao, Y. Guo, A novel two-step method for cross language representation learning, in: Advances in Neural Information Processing Systems, 2013, pp. 1259–1267.

[22] S. Dumais, T. Letsche, M. Littman, T. Landauer, Automatic cross–language retrieval using latent semantic indexing, in: AAAI'97 Spring Symposium Series: Cross-Language Text and Speech Retrieval, 1997, pp. 18–224.

[23] T. Nakamura, M. Shirakawa, T. Hara, S. Nishio, Semantic similarity measurements for multi-lingual short texts using wikipedia, in: Proceedings of WI '14 and IAT '14, 2014, pp. 22–29.

[24] M. Potthast, B. Stein, M. Anderka, A Wikipedia-Based Multilingual Retrieval Model, in: Proceedings of ECIR '08, 2008, pp. 522–530.

[25] G. Salton, C. Buckley, Term–weighting approaches in automatic text retrieval, in: Information Processing And Management, 1988, pp. 513–523.

[26] D. R. Hardoon, S. Szedmak, O. Szedmak, J. Shawe-Taylor, Canonical correlation analysis; an overview with application to learning methods, Tech. rep. (2007).

[27] `http://www.wikipedia.org`.

[28] J. Rupnik, A. Muhic, P. Skraba, Cross–lingual document retrieval through hub languages., xLiTe: Cross–Lingual Technologies, NIPS 2012 Workshop.

[29] J. Azzopardi, C. Staff, Incremental clustering of news reports, Algorithms 5 (3) (2012) 364–378.