Scalable Cross-lingual Document Similarity through Language-specific Concept Hierarchies

Carlos Badenes-Olmedo
cbadenes@fi.upm.es
Ontology Engineering Group,
Universidad Politécnica de Madrid
Boadilla del Monte, Spain

José Luis Redondo-García
jluisred@amazon.com
Amazon Research
Cambridge, UK

Oscar Corcho
ocorcho@fi.upm.es
Ontology Engineering Group,
Universidad Politécnica de Madrid
Boadilla del Monte, Spain

ABSTRACT

With the ongoing growth in number of digital articles in a wider set of languages and the expanding use of different languages, we need annotation methods that enable browsing multi-lingual corpora. Multilingual probabilistic topic models have recently emerged as a group of semi-supervised machine learning models that can be used to perform thematic explorations on collections of texts in multiple languages. However, these approaches require theme-aligned training data to create a language-independent space. This constraint limits the amount of scenarios that this technique can offer solutions to train and makes it difficult to scale up to situations where a huge collection of multi-lingual documents are required during the training phase. This paper presents an unsupervised document similarity algorithm that does not require parallel or comparable corpora, or any other type of translation resource. The algorithm annotates topics automatically created from documents in a single language with cross-lingual labels and describes documents by hierarchies of multi-lingual concepts from independently-trained models. Experiments performed on the English, Spanish and French editions of JCR-Acquis corpora reveal promising results on classifying and sorting documents by similar content.

CCS CONCEPTS

• **Information systems** → **Digital libraries and archives**; **Information retrieval**.

KEYWORDS

cross-lingual semantic similarity; large-scale text analysis; topic models

1 INTRODUCTION

Cross-language information extraction deals with the retrieval of documents written in languages different from the language of the user's query. At execution time, the query in the source language is typically translated into the target language of the documents with the help of a dictionary or a machine-translation system. But for many languages we may not have access to translation dictionaries or a full translation system, or they can be expensive to apply in an online search system. In such situations it is useful to rely on smaller annotation units derived from the text so the full content doesn't need to be translated, for instance by finding correspondences with regard to the topics discussed. In this case, it may be advisable to automatically learn cross-lingual topics to browse multi-lingual document collections.

Multi-lingual topic models discover language-specific descriptions of each topic from documents in multi-lingual corpora. They are mainly based on Latent Dirichlet Allocation (LDA) [4], adding supervised associations between languages by using *parallel* corpus, with sentence-aligned documents (e.g. Europarl[1] corpora), or *comparable* corpus, with theme-aligned documents (e.g. Wikipedia[2] articles), in multiple languages. These requirements restrict the kind of corpora that can be used for training since large parallel corpora are rare in most of the use cases, especially for languages with fewer resources. Wikipedia, for example, contains texts in 304 languages but 255 of them have less than 3% of articles[3]. Therefore, the requirement of parallel/comparable corpora for multilingual topic models limits their usage in many situations. In addition, these models rely on associations between documents prior to training. So in order to incorporate new languages or update the existing associations, the model must be re-trained with documents from all languages, making it difficult to scale to large corpora [17] [30].

Another approach is to use multi-lingual dictionaries as supervised methods [25][16]. They are usually easier to obtain and more widely available than parallel corpora (e.g. PANLEX[4] covers 5,700 languages and Wiktionary[5] covers 8,070 languages). Models built on dictionaries rather than a parallel/comparable corpora are potentially applicable to many more use cases. And even coherent multi-lingual topics can be learnt from partially and fully incomparable corpora with limited amounts of dictionary resources [18].

---

[1]https://ec.europa.eu/jrc/en/language-technologies/dcep
[2]https://www.wikipedia.org/
[3]https://meta.wikimedia.org/wiki/List_of_Wikipedias
[4]https://panlex.org
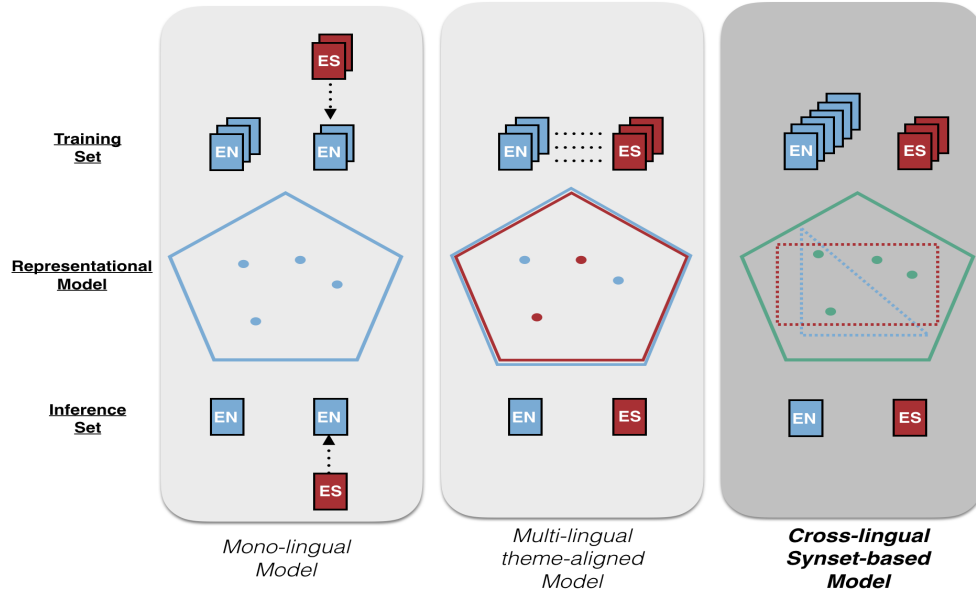[5]https://www.wiktionary.org

Figure 1: Graphical representation of the model that relies on the latent layer of cross-lingual topics obtained by LDA and hash functions through hierarchies of synsets. Mono-lingual approaches force to translate the documents to the same language to represent them in a unique feature space. Multi-language approaches require previously aligned topics from different languages so that documents can be represented in an equivalent feature space. Cross-lingual Synset-based approach creates a new space by combining the feature spaces of each language (i.e synsets from topn topic words). Documents are then represented in this unique space.

But all these probabilistic topic models are based on prior knowledge. Connections at document-level (by parallel or comparable corpora) or at word-level (by dictionaries) are created in the training-data before building the model. In this way, the pre-established language relations condition the creation of the topics (supervised method), instead of being inferred from the topics themselves as a posteriori knowledge (non-supervised method). We propose a completely unsupervised way of building cross-lingual topic models that uses sets of cognitive synonyms (synsets) to establish relations between language-specific topics once the model is created and does not require parallel or comparable data for training. These models can be used for large-scale multi-lingual (1) document classification and (2) information retrieval tasks. Our main contributions, described in this paper, are:

- a novel **cross-lingual document similarity algorithm** based on hierarchies of synsets.
- an open-source **implementation** [6] of the algorithm
- **data-sets** and **pre-trained models** to facilitate other researchers to replicate our experiments and validate and test their own ideas.

2 RELATED WORK

One of the greatest advantages of using probabilistic topic models (PTM) in large document collections is their ability to represent documents as probability distributions over a fixed number of topics, thereby mapping documents into a low-dimensional latent space

(the K-dimensional probability simplex, where K is the number of topics). A document, represented as a point in this simplex, is considered to have a particular topic distribution. This brings a lot of potential when applied over different information-retrieval (IR) tasks, as evidenced by recent works in different domains such as scholarly [12], health [24] [38], legal [33][14], news [19] and social networks [35] [7].

Multilingual probabilistic topic models (MuPTM) [39] have recently emerged as a group of language-independent generative machine learning models that can be used on large-volume theme-aligned multilingual text. Due to its generic language-independent nature and the power of inference on unseen documents, MuPTM's have found many interesting applications in many different cross-lingual tasks. They have been used on cross-lingual event clustering [8], document classification [9] [32], semantic similarity of words [29] [40], information retrieval [41] [11], document matching [34] [44], and others.

Once a PTM or MuPTM has been generated, documents can be represented by data points in a feature space based on topics to detect similarities among them exploiting inference results and using distance calculation metrics on it. Since exact similarity computations are unaffordable for neighbours detection tasks ($O(n^2)$), some algorithms based on approximate nearest neighbor (ANN) techniques have been proposed to efficiently perform document similarity search based on the low-dimensional latent space created by probabilistic topic models[43] [26]. They transform data points

---

[6]https://github.com/cbadenes/crosslingual-semantic-similarity

from the original feature space into a hash-code space, so that similar data points have larger probability of collision (i.e. having the same hash code). However, the smaller space created by existing hashing methods lose the exploratory capabilities that topic models offer and the explanatory power that topics have to support the document similarity. The notion of topics is discarded and therefore the ability to make thematic explorations of documents. Recently, a hashing algorithm that groups similar documents and preserves the notion of topics has been proposed [2]. It defines a hierarchical set-type data where each level of the hierarchy indicates the importance of the topic according to its distribution. Level 0 contains the topics of the document with the highest score. Level 1 contains the topics with highest score once the first ones have been eliminated, and so on. The knowledge provided by the topics to describe the documents is maintained and an efficient exploration of document collections on a large scale can be performed.

In this paper we take these hierarchies of PTM a step further, to make them cross-lingual. Documents from multi-language corpora can then be efficiently browsed and related without the need for translation. An algorithm that annotates topics with knowledge from a lexical data base and describes documents with hierarchical expressions of multi-lingual concepts is presented in this paper. Hash codes are created from those concept hierarchies to perform document classification and information retrieval tasks on large document collections.

3 AN APPROACH TO CALCULATE CROSS-LINGUAL DOCUMENT SIMILARITY EFFICIENTLY

Our similarity algorithm considers that cross-lingual models can be built from non-parallel or even non-comparable collections of multi-lingual documents. It first creates a probabilistic topic model for each language separately, and then annotates the topics with cross-lingual labels (Fig 1). In the same way, the topic distribution of documents expressed through weighted vectors are first transformed into hierarchies of topics according to their relevance. And then documents are described by a 3-level hierarchy of cross-lingual concepts.

3.1 Synset-based annotations

Each topic is annotated with a list of synset [5] retrieved from Word-Net[7][27] based on its top*n* words (Fig 2). Word by word are queried in WordNet to retrieve its synsets. The final set of synsets for a topic is the union of the synsets from the individual top-words of a topics. Based on empirical evidence from different executions of the algorithm, n=5 is the configuration that offered the best performance in our tests. Let's look at an example to clarify how it works. Given the topics of Table 1, the EN-Topic (*"communications systems"*) is annotated with the following synset list: *radio.a.01, radio.v.01, radio.n.03, radio.n.01, radio_receiver.n.01, equipment.n.01, network.n.02, network.n.04, network.v.01, network.n.05, network.n.01, net.n.06, communication.n.02, communication.n.03, communication.n.01, regulative.s.01.* The list of synset for the ES-Topic (*"sistema de comunicación"*) is: *kit.n.02,team.n.01, equipment.n.01, net.n.02, net.n.05,*

Cross-lingual Hashing

H={
  h0=(*fruit.n.03, fruit.n.01, fruit.v.02*),
  h1=(*guarantee.n.02, guarantee.n.03, management.n.01, management.n.02*),
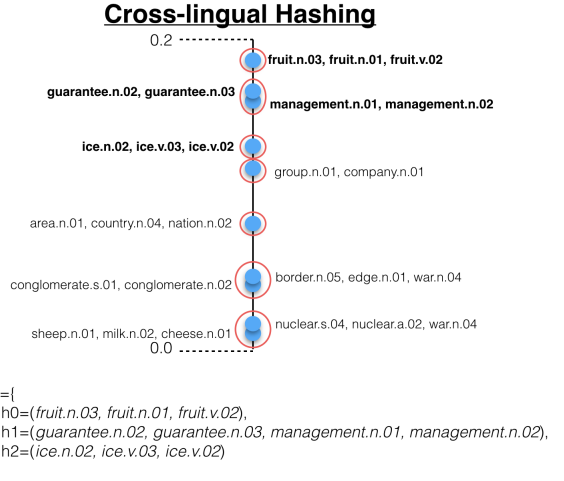 h2=(*ice.n.02, ice.v.03, ice.v.02*)
}

Figure 2: Cross-lingual hash-expression (H) of a document based on WordNet-synset annotations created from the top words of each topic distribution. The most relevant topics are grouped according to their importance in three levels (h0, h1 and h2)

network.n.05, web.n.06, network.n.01, web.n.02, communication.n.02, communication.n.01, announcement.n.02, spectrum.n.02, spectrum.n.01, creep.n.01, ghost.n.01, apparition.n.02, electromagnetic.a.01. And the list for FR-Topic (*"systeme de communication"*) is: *access.n.02, approach.n.07, approach.n.02, access.n.06, access.n.03, access.n.05, assault.n.03, bout.n.02, approach.n.01, entree.n.02, entry.n.01, entrance.n.01, entry.n.03, admission.n.01, submission.n.01, introduction.n.01.* The librAIry NLP service[8] was used to identify the list of synsets from a topic description based on top words. It includes the Open Multi-lingual WordNet[9] [6].

3.2 Document representation

Documents (i.e seen as data points in the generated space) are transformed from the original feature space based on mono-lingual topic distributions into a hierarchical-code space, so that similar data points share relevant cross-lingual concepts. Since topic models create latent themes from word co-occurrence statistics in a corpus, a cross-lingual concept specifies the knowledge about the word-word relations it contains for each language. This abstraction can be extended to cover the knowledge derived from sets of topics. The topics are obtained via state-of-the art methods, collapsed Gibbs sampling[15] for LDA, and hierarchically divided into groups with different degrees of semantic specificity in a document. Documents represented as a weighted mixture of latent topics (per-document topic distributions) are then annotated in these feature spaces with the relation between topics inside each hierarchy level. Regardless of their language, they are then described by cross-lingual concepts (based on WordNet-synset annotations) and hash codes are calculated to summarize their content [2]. The

---

[7]https://wordnet.princeton.edu/

[8]http://librairy.linkeddata.es/nlp
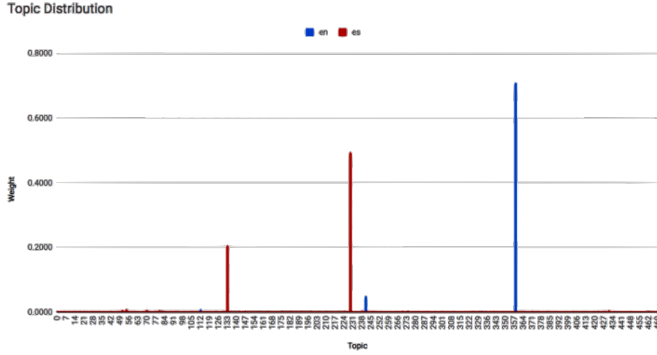[9]http://compling.hss.ntu.edu.sg/omw/

Figure 3: topic distributions from the same document in English ($h_{EN} = \{(t3062), (t335), (t8278)\}$) and Spanish ($h_{ES} = \{(t335), (t4060), (t5769)\}$).

hash expression sets a 3-level hierarchy of cross-lingual concepts. Topics with similar presence in a document are grouped together in the same hierarchical level (Fig 2). Each level of the hierarchy indicates the importance of the topic according to its distribution. Level 0 describes the topics with the highest score. Level 1 describes the topics with highest score once the first ones have been eliminated, and so on. Documents are described by vectors containing set of topics (i.e. set of synsets), where each dimension means a topic relevance. Given a document d with a topic distribution $q = [t0 = 0.28, t1 = 0.05, t2 = 0.44, t3 = 0.23]$, the hash expression may be $H_d = (ts2), (ts0, ts3), (ts1)$. It means that topic $t2$ described by the synset $ts2$ is the most relevant (i.e 0.44 score), then topics $t0$ and $t3$ described by synsets $ts0$ and $ts3$ (i.e 0.28 and 0.23 scores) and, finally, topic $t1$ described by synset $ts1$ (i.e 0.05).

3.3 Similarity metric

Since documents are described by set-type data, the proposed distance metric is based on the Jaccard coefficient. This metric is mainly used for this type of data [23] [21] [22] [42] and computes the similarity of sets by looking at the relative size of their intersection ($\frac{|A \cap B|}{|A \cup B|}$). More specifically, the similarity metric used to compare the hash codes created from set of topics is the sum of the Jaccard distances for each hierarchy level, i.e. for each set of topics [2]:

$$d_H(H_A, H_B) = \sum_{l=1}^{L} \left(d_J(H_A(h_l), H_B(h_l)) \right) = \sum_{l=1}^{L} \left(1 - \frac{H_A(h_l) \cap H_B(h_l)}{H_A(h_l) \cup H_B(h_l)} \right) \quad (1)$$

where H_A and H_B are hash codes, $H_A(h_l)$ and $H_B(h_l)$ are the set of topics up to level l for each hash code H, and L is the maximum hierarchy level. A corner case is $L = T$, where T is the number of topics in the model.

4 EVALUATION

A way to evaluate our cross-lingual document similarity algorithm is to test how well it performs in practice for different real-life tasks: document classification and information retrieval. Evaluation is done using the B-Cubed metrics [3] to estimate the fit between two

clusters, the one obtained from a supervised category-based topic alignment algorithm and the one obtained from our unsupervised synset-based topic alignment algorithm.

Let CL_i be the cluster that document t_i gets clustered in, and G_i its correct cluster from the ground truth. The B-Cubed metric then calculates $precision = \frac{|CL_i \cap G_i|}{|CL_i|}$ and $recall = \frac{|CL_i \cap G_i|}{|G_i|}$. The total precision and recall of the clustering are taken as the average of the precision and recall scores over all documents. Results are also presented in terms of the F_1 measure to balance between precision and recall: $F_1 = \frac{2 \cdot precision \cdot recall}{precision + recall}$. The aim is to measure the performance of the algorithm taking into account documents with manual category assignments.

4.1 Data Sets

A multilingual corpora is required to create the cross-lingual topic models that support our document similarity algorithm. The key feature is that it does not need to be parallel or comparable. However, in order to be able to compare the performance of our unsupervised algorithm with a semi-supervised algorithm (MuPTM-based) it is necessary to use theme-aligned corpora that map topics across languages. We used the JRC-Acquis[10] corpus [37]. It is a collection of legislative texts written in 23 languages, although we only use English, Spanish and French for the tests. Most texts have been manually classified into subject domains according to the EU-ROVOC[11] thesaurus [10], which exists in one-to-one translations into approximately twenty languages and distinguishes about 6,000 hierarchically organised descriptors (subject domains). More than 20k documents were used for each language-specific model, a total of 82,140 texts are included in the training-test package, which is publicly available[12] for reuse.

4.2 Cross-lingual Models

The JRC-Acquis corpus is annotated with EUROVOC categories. These categories are shared among languages and will serve as support for building the topic models. Moreover, the topic independence assumption [4] of LDA models should be also satisfied, so the categories must first be moved to their base concepts and therefore disjointed categories. The EUROVOC taxonomy has 7,193 concepts/labels from 21 domain areas such as politics, international relations, european union, law, economics, etc. There are 4,904 reciprocal hierarchical relationships (no polyhierarchy) and 6,992 reciprocal associative relationships. Using hierarchical relations, we identified the root concepts from which all other categories derive. The initial 7,193 labels were then reduced to 452 labels, which are independent (topic independence assumption from LDA is satisfied), and can be used to train the topic models.

A pre-processing of the documents was required to clean texts and to build a suitable data set for the model. We assume that terms with high frequency are not specific to a particular topic, so words present in more than 90% of the corpus are considered stopwords and removed from the model. Also, rare terms that occur infrequently are considered not representative of a single topic since they do not appear enough to infer that it is salient for a

---

[10]https://ec.europa.eu/jrc/en/language-technologies/jrc-acquis
[11]http://eurovoc.europa.eu/
[12]http://librairy.linkeddata.es/data/jrc/select?q=*:*

EN-Topic 3	ES-Topic 3	FR-Topic 26
"communications systems"	*"sistema de comunicación"*	*"systeme de communication"*
radio	equipo	communications
equipment	red	reseaux
network	comunicación	electroniques
communication	espectro	acces
regulatory	electromagnético	telecommunications
spectrum	electrónico	service
electronic	reglamentación	universel
access	banda	reglamentaires
standard	etsir	nationales
mobile	compatibilidad	fourniture

Table 1: Randonly selected theme-aligned topics described by top 10 words based on EUROVOC annotations from JRC-Acquis dataset

topic. Then, words present in less than 0.5% of the corpus are also removed from the model. Lemmatized expressions of names, verbs and adjectives were used to create the bag-of-words, and documents with less than 100 characters were discarded since LDA has proven to has lower performance with these type of texts [7].

Then, we set the number of topics $K = 500$ (several configurations were evaluated, but this was the closest to the performance obtained with the supervised model based on categories). We run the Gibbs samplers for 1000 training iterations on LDA from the open-source librAIry [1] software. The Dirichlet priors $\alpha = 0.1$ and $\beta = 0.01$ were set following [20]. Once the word distributions for each topic is available, the list of synsets related with the top5 words for each topic are identified (this number is set to offer better performance after trying several alternatives). Finally, the 3-level hierarchy of topics per document is replaced by a 3-level hierarchy of synsets. Probabilistic topic models in Spanish[13], English [14] and French[15] were created independently without previously establishing any type of alignment between their topics.

In order to compare the performance of this non-supervised approach with approaches based on aligned topics, we need to use a variant of LDA to force the correspondence between the 452 root categories identified in the EUROVOC thesaurus and the latent topics of the model. Thus, LabeledLDA [36], a supervised version of LDA, was used to perform parameter estimation. Theme-aligned probabilistic topic models in Spanish[16], English [17] and French[18]) were created sharing the topics but not its definitions (i.e. vocabulary) (see table 1).

A simple way of looking at the output quality of the topic models is by simply inspecting top words associated with a particular topic learned during training. A latent topic is semantically coherent if it assigns high probability scores to words that are semantically related [13] [31] [28]. It is much easier for humans to judge semantic coherence of cross-lingual topics and their alignment across languages when observing the actual words constituting a topic.

These words provide a shallow qualitative representation of the latent topic space, and could be seen as direct and comprehensive word-based summaries of a large document collection.

Samples of cross-lingual topics are provided in Table 1. We may consider this visual inspection of the top words associated with each topic as an initial qualitative evaluation, suitable for human judges. Documents present similar topic distributions when projecting their content on topics according to their language as can be seen in fig 3. Since the topic identifiers are not aligned, the graphs appear displaced.

4.3 Cross-lingual Document Classification

A random group of 1k documents, which have not been used to train the models, is considered for evaluation as they are manually tagged with EUROVOC categories. For each document, the cluster to which it belongs is identified from its categories. This cluster is then compared (B-Cubed metrics) with the one obtained from the labels generated from its most representative topics (*cat*) and with the one obtained from the labels generated with the WordNet-Synsets of those topics (*syn*). Algorithm performance is evaluated in monolingual, bilingual, and multilingual document collections (tables 2 and 3) .

The results show a higher performance of the semi-supervised algorithm (categories-based topic alignment) in terms of precision, and of the unsupervised algorithm (synset-based topic alignment) in terms of coverage. The cause lies in the set of synonyms generated by WordNet, being able to share the same synset for two different topics. From a more general point of view (fMeasure), the benefit obtained by the increase in coverage (recall) is greater than by the loss of accuracy (precision).

4.4 Cross-lingual Information Retrieval

Given a set of documents and a text, the task is to rank the documents according to their relevance to the query text regardless of the language used. The JRC-Acquis corpus is used because by having texts tagged with EUROVOC categories we can build a ground-truth set grouping the documents that share the same codes as those used in the query document. A collection of 1k randomly selected documents (monolingual, bi-lingual and multi-lingual) are

---

JRC-Acquis Corpora							
		en		**es**		**fr**	
		cat	*syn*	*cat*	*syn*	*cat*	*syn*
prec	*min*	0.01	0.01	0.01	0.01	0.01	0.01
	max	1.00	0.95	1.00	0.87	1.00	0.87
	mean	**0.58**	0.48	**0.55**	0.48	**0.55**	0.41
	dev	0.27	0.23	0.27	0.22	0.26	0.20
rec	*min*	0.01	0.03	0.01	0.04	0.01	0.05
	max	0.96	1.00	0.93	1.00	0.95	1.00
	mean	0.39	**0.52**	0.36	**0.49**	0.42	**0.51**
	dev	0.24	0.20	0.23	0.20	0.23	0.23
f1	*min*	0.02	0.03	0.01	0.02	0.02	0.03
	max	0.70	0.75	0.70	0.71	0.70	0.73
	mean	0.35	**0.42**	0.32	**0.41**	0.37	**0.39**
	dev	0.16	0.15	0.15	0.15	0.17	0.17

Table 2: Document classification performance (precision-'prec', recall-'rec' and fMeasure-'f1') of the categories-based (*cat*) and synset-based (*syn*) topic alignment algorithms in monolingual document collections (en, es, fr)

JRC-Acquis Corpora									
		en-es		**en-fr**		**es-fr**		**en-es-fr**	
		cat	*syn*	*cat*	*syn*	*cat*	*syn*	*cat*	*syn*
prec	*min*	0.01	0.01	0.01	0.01	0.01	0.01	0.02	0.02
	max	1.00	0.97	1.00	0.98	1.00	0.97	1.00	0.98
	mean	**0.62**	0.55	**0.62**	0.56	**0.61**	0.56	**0.59**	0.52
	dev	0.26	0.23	0.25	0.23	0.26	0.23	0.26	0.23
rec	*min*	0.01	0.09	0.00	0.06	0.01	0.07	0.01	0.07
	max	1.00	1.00	0.94	0.97	0.91	0.93	0.86	0.93
	mean	0.33	**0.57**	0.36	**0.50**	0.30	**0.40**	0.25	**0.39**
	dev	0.16	0.23	0.17	0.19	0.13	0.13	0.13	0.15
f1	*min*	0.02	0.02	0.01	0.02	0.02	0.02	0.02	0.05
	max	0.75	0.81	0.76	0.81	0.68	0.72	0.62	0.66
	mean	0.36	**0.49**	0.38	**0.47**	0.35	**0.41**	0.30	**0.38**
	dev	0.16	0.18	0.15	0.18	0.14	0.14	0.11	0.12

Table 3: Document classification performance (precision-'prec', recall-'rec' and fMeasure-'f1') of the categories-based (*cat*) and synset-based (*syn*) topic alignment algorithms in multi-lingual document collections (en-es, en-fr, es-fr, en-es-fr)

annotated by the category-based and synset-based topic alignment algorithms. Then, we randomly take articles to search in D for documents that share the same categories than the query document (i.e the ground-truth set). Next, the query text is used to search in D for similar documents using category-based annotations and synset-based annotations. We evaluate the performance of the algorithms in terms of precision@3, precision@5 and precision@10 (tables 4 and 5) .

Although the precision values are lower than those obtained by semi-supervised approximation, they are sufficiently promising (around 0.75) to think that introducing improvements in the lemmatization process would increase the quality of the WordNet-synset annotations derived from the most representative words of each topic (precision values close to 0.8 in the English corpus).

JRC-Acquis Corpora							
		en		**es**		**fr**	
		cat	*syn*	*cat*	*syn*	*cat*	*syn*
p@3	*mean*	**0.84**	0.83	**0.81**	0.78	**0.83**	0.74
	dev	0.26	0.26	0.27	0.29	0.26	0.32
p@5	*mean*	**0.82**	0.80	**0.79**	0.75	**0.80**	0.72
	dev	0.25	0.25	0.25	0.27	0.25	0.29
p@10	*mean*	**0.77**	0.76	**0.75**	0.73	**0.77**	0.68
	dev	0.23	0.25	0.25	0.27	0.24	0.27

Table 4: Information retrieval performance (precision@3, precision@5 and precision@10) of the categories-based (*cat*) and synset-based (*syn*) topic alignment algorithms in monolingual document collections (en, es, fr)

JRC-Acquis Corpora									
		en-es		**en-fr**		**es-fr**		**en-es-fr**	
		cat	*syn*	*cat*	*syn*	*cat*	*syn*	*cat*	*syn*
p@3	*mean*	**0.84**	0.79	**0.86**	0.77	**0.85**	0.78	**0.85**	0.75
	dev	0.25	0.28	0.23	0.28	0.25	0.29	0.24	0.31
p@5	*mean*	**0.82**	0.76	**0.84**	0.75	**0.82**	0.76	**0.81**	0.72
	dev	0.24	0.26	0.23	0.27	0.23	0.27	0.23	0.28
p@10	*mean*	**0.78**	0.73	**0.80**	0.70	**0.77**	0.72	**0.76**	0.67
	dev	0.22	0.24	0.22	0.24	0.23	0.26	0.23	0.26

Table 5: Information retrieval performance (precision@3, precision@5 and precision@10) of the categories-based (*cat*) and synset-based (*syn*) topic alignment algorithms in multi-lingual document collections (en-es, en-fr, es-fr, en-es-fr)

5 CONCLUSIONS

In this paper we present a first approach towards the calculation of cross-lingual document similarity through unsupervised probabilistic topic models and WordNet-synsets without the need for parallel or comparable corpora.

As expected, the performance of our algorithm in terms of accuracy is not as good as that of the algorithm based on topics previously aligned by documents annotated with categories (theme-aligned training data). However, in terms of coverage, the performance of the unsupervised approach is much greater than that offered by the semi-supervised approach, to the point of offering better overall performance (i.e f1) in classification tasks. In addition, the algorithm has proved to perform close to the semi-supervised algorithm in information retrieval task, which makes us think that the process of topic annotation by set of synonyms should be improved to filter those elements that are not sufficiently representative. Our future lines of work will go in that direction, incorporating context information to identify the most representative synset for each topic.

REFERENCES

[1] C. Badenes-Olmedo, J.L. Redondo-Garcia, and O. Corcho. 2017. Distributing Text Mining tasks with librAIry. In *17th ACM Symposium on Document Engineering (DocEng)*. https://doi.org/10.1145/3103010.3121040

[2] C. Badenes-Olmedo, J.L. Redondo-García, and O. Corcho. 2019. Large-Scale Semantic Exploration of Scientific Literature using Topic-based Hashing Algorithms. *Semantic Web Journal (under review)* (2019).

[3] A. Bagga and B. Baldwin. 1998. Algorithms for scoring coreference chains. In *Proceedings of the 1st international conference on language resources and evaluation workshop on linguistics coreference*. 563âĂŞ566.

[4] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research* 3, 4-5 (2003), 993–1022. https://doi.org/10.1162/jmlr.2003.3.4-5.993

[5] Francis Bond and Ryan Foster. 2013. Linking and Extending an Open Multilingual Wordnet. *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL)* (2013), 1352–1362.

[6] Francis Bond and Kyonghee Paik. 2012. A survey of wordnets and their licenses. In *Proceedings of the 6th Global WordNet Conference (GWC 2012)*. 64–71.

[7] Xueqi Cheng, Xiaohui Yan, Yanyan Lan, and Jiafeng Guo. 2014. BTM : Topic Modeling over Short Texts. *IEEE Transactions on Knowledge and Data Engineering* 26, 12 (2014), 2928–2941. https://doi.org/10.1109/TKDE.2014.2313872

[8] Wim De Smet and Marie-Francine Moens. 2009. Cross-language linking of news stories on the web using interlingual topic modelling. In *Proceedings of the 2nd ACM workshop on Social web search and mining*. 57. https://doi.org/10.1145/1651437.1651447

[9] Wim De Smet, Jie Tang, and Marie-Francine Moens. 2011. Knowledge Transfer across Multilingual Corpora via Latent Topics. In *Advances in Knowledge Discovery and Data Mining*. 549–560.

[10] Eurovoc. 1995. Thesaurus EUROVOC - Volume 2: Subject-Oriented Version. Ed. 3/English Language. Annex to the index of the Official Journal of the EC. In *Luxembourg, Office for Official Publications of the European Communities*.

[11] Debasis Ganguly, Johannes Leveling, and Gareth Jones. 2012. Cross-Lingual Topical Relevance Models. In *Proceedings of COLING 2012*. 927–942.

[12] Christopher J. Gatti, James D. Brooks, and Sarah G. Nurre. 2015. A Historical Analysis of the Field of OR/MS using Topic Models. *CoRR* abs/1510.0 (2015).

[13] Alfio Massimiliano Gliozzo. 2007. The Domain Restriction Hypothesis: Relating Term Similarity and Semantic Consistency. *In. Proceedings of NAACL HLT* April (2007), 131.

[14] Derek Greene and James P Cross. 2016. Exploring the political agenda of the european parliament using a dynamic topic modeling approach. *Political Analysis* 25, 1 (2016), 77–94. https://doi.org/10.1017/pan.2016.7

[15] Thomas L Griffiths and Mark Steyvers. 2004. Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America* 101 Suppl (2004), 5228–35. https://doi.org/10.1073/pnas.0307752101

[16] E. Dario Gutierrez, Ekaterina Shutova, Patricia Lichtenstein, Gerard de Melo, and Luca Gilardi. 2016. Detecting Cross-cultural Differences Using a Multilingual Topic Model. *Transactions of the Association for Computational Linguistics* 4 (2016), 47–60.

[17] Shudong Hao, Jordan L. Boyd-Graber, and Michael J. Paul. 2018. Lessons from the Bible on Modern Topics: Adapting Topic Model Evaluation to Multilingual and Low-Resource Settings. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguis- tics: Human Language Technologies, NAACL-HLT*. 1090âĂŞ1100.

[18] Shudong Hao and Michael J. Paul. 2018. Learning Multilingual Topics from Incomparable Corpus. In *COLING*.

[19] Jin He, Lei Li, and Xindong Wu. 2017. A self-adaptive sliding window based topic model for non-uniform texts. In *Proceedings - IEEE International Conference on Data Mining, ICDM*, Vol. 2017-Novem. 147–156. https://doi.org/10.1109/ICDM.2017.24

[20] Yuening Hu, Ke Zhai, Vladimir Eidelman, and Jordan L. Boyd-Graber. 2014. Polylingual Tree-Based Topic Models for Translation Domain Adaptation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*. 1166âĂŞ1176.

[21] Jianqiu Ji, Jianmin Li, Shuicheng Yan, Qi Tian, and Bo Zhang. 2013. Min-Max Hash for Jaccard Similarity. In *2013 IEEE 13th International Conference on Data Mining*. IEEE, 301–309. https://doi.org/10.1109/ICDM.2013.119

[22] Ping Li and Christian König. 2010. b-Bit minwise hashing. In *Proceedings of the 19th international conference on World wide web - WWW '10*. ACM Press, 671. https://doi.org/10.1145/1772690.1772759

[23] Ping Li, Art B. Owen, and Cun-Hui Zhang. 2012. One Permutation Hashing. *Advances in Neural Information Processing* (2012).

[24] Hsin Min Lu, Chih Ping Wei, and Fei Yuan Hsiao. 2016. Modeling healthcare data using multiple-channel latent Dirichlet allocation. *Journal of Biomedical Informatics* 60 (2016), 210–223. https://doi.org/10.1016/j.jbi.2016.02.003

[25] Tengfei Ma and Tetsuya Nasukawa. 2017. Inverted Bilingual Topic Models for Lexicon Extraction from Non-parallel Data. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*. 4075–4081.

[26] Xianling Mao, Bo-Si Feng, Yi-Jing Hao, Liqiang Nie, Heyan Huang, and Guihua Wen. 2017. S2JSD-LSH: A Locality-Sensitive Hashing Schema for Probability Distributions. In *AAAI*.

[27] George A. Miller. 1995. WordNet: A Lexical Database for English. *Commun. ACM* 38, 11 (1995), 39–41.

[28] David Mimno, Hanna Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. 2011. Optimizing Semantic Coherence in Topic Models. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Edinburgh, Scotland, UK., 262–272.

[29] David Mimno, Hanna M Wallach, Jason Naradowsky, David A Smith, and Andrew McCallum. 2009. Polylingual Topic Models. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2 - Volume 2 (EMNLP '09)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 880–889.

[30] Maria Moritz and Marco Bu İLchler. 2017. Ambiguity in Semantically Related Word Substitutions: an Investigation in Historical Bible Translations.. In *Proceedings of the NoDaLiDa 2017 Workshop on Processing Historical Language*. 18–23.

[31] David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. 2010. Automatic Evaluation of Topic Coherence. In *Human Language Technologies: The 2010 Annual Conference of the North {A}merican Chapter of the Association for Computational Linguistics*. 100–108.

[32] Xiaochuan Ni, Jian-Tao Sun, Jian Hu, and Zheng Chen. 2011. Cross Lingual Text Classification by Mining Multilingual Topics from Wikipedia. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*. 375–384. https://doi.org/10.1145/1935826.1935887

[33] James OâĂŹNeill, CÃĺcile Robin, Leona OâĂŹBrien, and Paul Buitelaar. 2017. An analysis of topic modelling for legislative texts. *CEUR Workshop Proceedings* 2143 (2017). https://doi.org/10.475/123

[34] John C Platt, Kristina Toutanova, and Wen-tau Yih. 2010. Translingual Document Representations from Discriminative Projections. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP '10)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 251–261.

[35] Daniel Ramage, Susan Dumais, and Dan Liebling. 2010. Characterizing Microblogs with Topic Models. In *Proceedings of the Fourth International Conference on Weblogs and Social Media*. 1–8.

[36] Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D Manning. 2009. Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*. 248–256. https://doi.org/10.3115/1699510.1699543

[37] Ralf Steinberger, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaz Erjavec, Dan Tufis, and Daniel Varga. 2006. The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'2006)*, Vol. 4. 2142–2147.

[38] Mike Donald Tapi Nzali, Sandra Bringay, Christian Lavergne, Caroline Mollevi, and Thomas Opitz. 2017. What Patients Can Tell Us: Topic Analysis for Social Media on Breast Cancer. *JMIR medical informatics* 5, 3 (7 2017), e23. https://doi.org/10.2196/medinform.7779

[39] Ivan Vulić, Wim De Smet, Jie Tang, and Marie Francine Moens. 2015. Probabilistic topic modeling in multilingual settings: An overview of its methodology and applications. *Information Processing and Management* 51, 1 (2015), 111–147. https://doi.org/10.1016/j.ipm.2014.08.003

[40] Ivan Vulić and Marie-Francine Moens. 2012. Detecting Highly Confident Word Translations from Comparable Corpora Without Any Prior Knowledge. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*. 449–459.

[41] Ivan Vulić and Marie-Francine Moens. 2013. A Unified Framework for Monolingual and Cross-Lingual Relevance Modeling Based on Probabilistic Topic Models. In *Advances in Information Retrieval*. 98–109.

[42] Wan-Lei Zhao, HervÃĺ Jégou, and Guillaume Gravier. 2013. Sim-min-hash. In *Proceedings of the 21st ACM international conference on Multimedia - MM '13*. 577–580. https://doi.org/10.1145/2502081.2502152

[43] Yi Zhen, Yue Gao, Dit-Yan Yeung, Hongyuan Zha, and Xuelong Li. 2016. Spectral Multimodal Hashing and Its Application to Multimedia Retrieval. *IEEE Transactions on Cybernetics* 46, 1 (2016), 27–38. https://doi.org/10.1109/TCYB.2015.2392052

[44] Zede Zhu, Miao Li, Lei Chen, and Zhenxin Yang. 2013. Building Comparable Corpora Based on Bilingual {LDA} Model. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. 278–282.