Automatic Summarization Approaches
to Speed up Topic Model Learning Process

Mohamed Morchid[1], Juan-Manuel Torres-Moreno[1,2,3], Richard Dufour[1],
Javier Ramirez-Rodriguez[4], and Georges Linarès[1]

[1]LIA - Université d'Avignon (France)
{firstname.lastname}@univ-avignon.fr
[2]SFR Agorantic, Université d'Avignon (France)
[3]École Polytechnique de Montréal, Québec (Canada)
[4]Universidad Autónoma Metropolitana–Azcapotzalco (Mexico)
jararo@azc.uam.mx

Abstract. The number of documents available into Internet moves each
day up. For this reason, processing this amount of information effectively
and expressibly becomes a major concern for companies and scientists.
Methods that represent a textual document by a topic representation
are widely used in Information Retrieval (IR) to process big data such
as Wikipedia articles. One of the main difficulty in using topic model on
huge data collection is related to the material resources (CPU time and
memory) required for model estimate. To deal with this issue, we pro-
pose to build topic spaces from summarized documents. In this paper, we
present a study of topic space representation in the context of big data.
The topic space representation behavior is analyzed on different lan-
guages. Experiments show that topic spaces estimated from summaries
are as relevant as those estimated from the complete documents. The real
advantage of such an approach is the processing time gain: we showed
that the processing time can be drastically reduced using summarized
documents (more than 60% in general). This study finally points out the
differences between thematic representations of documents depending on
the targeted languages such as English or latin languages.

1 Introduction

The number of documents available into Internet moves each day up in an ex-
ponential way. For this reason, processing this amount of information effectively
and expressibly becomes a major concern for companies and scientists. An im-
portant part of the information is conveyed through textual documents such as
blogs or micro-blogs, general or advertise websites, and encyclopedic documents.
This last type of textual data increases each day with new articles, which con-
vey large and heterogenous information. The most famous and used collaborative
Internet encyclopedia is Wikipedia, enriched by worldwide volunteers. It is the
12th most visited website in the USA, with around 10.65 million users visiting

the site daily, and a total reaching 39 millions of the estimated 173.3 million Internet users in the USA[1] [2].

The massive number of documents provided by Wikipedia is mainly exploited by Natural Language Processing (NLP) scientists in various tasks such as keyword extraction, document clustering, automatic text summarization, etc. Different classical representations of a document, such as term-frequency based representation [1], have been proposed to extract word-level information from this large amount of data in a limited time. Nonetheless, these straightforward representations obtain poor results in many NLP tasks with respect to more abstract and complex representations. Indeed, the classical term-frequency representation reveals little in way of intra- or inter-document statistical structure, and does not allow us to capture possible and unpredictable context dependencies. For these reasons, more abstract representations based on latent topics have been proposed. The most known and used one is the latent Dirichlet allocation (LDA) [2] approach which outperforms classical methods in many NLP tasks. The main drawback of this topic-based representation is the time needed to learn LDA latent variables. This massive waste of time that occurs during the LDA learning process, is mainly due to the documents size along with the number of documents, which is highly visible in the context of big data such as Wikipedia.

The solution proposed in this article is to summarize text documents contained into a big data corpus (here Wikipedia) and then, learn a LDA topic space. This should answer the these three raised difficulties:

- reducing the processing time during the LDA learning process,
- retaining the intelligibility of documents,
- maintaining the quality of LDA models.

With this text summarization approach, the size of documents will be drastically reduced, the intelligibility of documents will be preserved, and we make the assumption that the LDA model quality will be conserved. Moreover, for all these reasons, the classical term-frequency document reduction is not considered in this paper. Indeed, this extraction of a subset of words to represent the document content allows us to reduce the document size, but does not keep the document structure and then, the intelligibility of each document.

The main objective of the paper is to compare topic space representations using complete documents and summarized ones. The idea behind is to show the effectiveness of this document representation, in terms of performance and time-processing reduction, when summarized documents are used. The topic space representation behavior is analyzed on different languages (English, French and Spanish). In the series of proposed experiments, the topic models built from complete and summarized documents are evaluated using the Jensen-Shannon ($\mathcal{JS}$) divergence measure as well as the perplexity measure. To the best of our knowledge, this is the most extensive set of experiments interpreting the evaluation of topic spaces built from complete and summarized documents without human models.

---

[1] http://www.alexa.com

[2] http://www.metrics2.com

The rest of the paper is organized in the following way: first, Section 2 introduces related work in the areas of topic modeling and automatic text summarization evaluations. Then, Section 3 describes the proposed approach, including the topic representation adopted in our work and the different summarization systems employed. Section 4 presents the topic space quality measures used for the evaluation. Experiments carried out along with with the results presented in Section 5. A discussion is finally proposed in Section 6 before concluding in Section 7.

2   Related work

Several methods were proposed by Information Retrieval (IR) researchers to process large corpus of documents such as Wikipedia encyclopedia. All these methods consider documents as a bag-of-words [1] where the word order is not taken into account.

Among the first methods proposed in IR, [3] propose to reduce each document from a discrete space (words and documents) to a vector of numeral values represented by the word counts (number of occurrences) in the document named TF-IDF [4]. This approach showed its effectiveness in different tasks, and more precisely in the basic identification of discriminative words for a document [5]. However, this method has many weaknesses such as the small amount of reduction in description length, or the weak of inter- or intra-statistical structure of documents in the text corpus.

To substantiate the claims regarding TF-IDF method, IR researchers have proposed several other dimensionality reductions such as Latent Semantic Analysis (LSA) [6, 7] which uses a *singular value decomposition* (SVD) to reduce the space dimension.

This method was improved by [8] which proposed a Probabilistic LSA (PLSA). PLSA models each word in a document as a sample from a mixture model, where the mixture components are multinomial random variables that can be viewed as representations of topics. This method demonstrated its performance on various tasks, such as sentence [9] or keyword [10] extraction. In spite of the effectiveness of the PLSA approach, this method has two main drawbacks. The distribution of topics in PLSA is indexed by training documents. Thus, the number of its parameters grows with the training document set size and then, the model is prone to overfitting which is a main issue in an IR task such as documents clustering. However, to address this shortcoming, a tempering heuristic is used to smooth the parameter of PLSA models for acceptable predictive performance: the authors in [11] showed that overfitting can occur even if tempering process is used.

To overcome these two issues, the latent Dirichlet allocation (LDA) [2] method was proposed. Thus, the number of LDA parameters does not grow with the size of the training corpus and LDA is not candidate for overfitting. Next section describes more precisely the LDA approach that will be used in our experimental study.

[12] evaluated the effectiveness of the Jensen-Shannon ($\mathcal{JS}$) theoretic measure [13] in predicting systems ranks in two text summarization tasks: *query-focused* and *update summarization*. They have shown that ranks produced by PYRAMIDS and those produced by $\mathcal{JS}$ measure correlate. However, they did not investigate the effect of the measure in other complex summarization tasks. [14, 15] using the FRESA system, reported evaluation results without human references on several tasks: generic and multi-document summarization (DUC 2004 Task 2), biographical summarization (DUC 2004 Task 5), opinion summarization (TAC 2008 OS) and summarization in languages other than English.

Next section describes the proposed approach followed in this article, including the topic space representation with the LDA approach and its evaluation with the perplexity and the FRESA divergence metrics.

3 Overview of the proposed approach

Figure 1 describes the approach proposed in this paper to evaluate the quality of a topic model representation with and without automatic summarization systems. The latent Dirichlet allocation (LDA) approach, described in details in the next section, is used for topic representation, in conjunction with different state-of-the-art summarization systems presented in Section 3.2.

3.1 Topic representation: latent Dirichlet allocation

LDA is a generative model which considers a document, seen as a *bag-of-words* [1], as a mixture of latent topics. In opposition to a multinomial mixture model, LDA considers that a theme is associated to each occurrence of a word composing the document, rather than associate a topic with the complete document. Thereby, a document can change of topics from a word to another. However, the word occurrences are connected by a latent variable which controls the global respect of the distribution of the topics in the document. These latent topics are characterized by a distribution of word probabilities associated with them. PLSA and LDA models have been shown to generally outperform LSA on IR tasks [16]. Moreover, LDA provides a direct estimate of the relevance of a topic knowing a word set.

Figure 2 shows the LDA formalism. For every document d of a corpus D, a first parameter θ is drawn according to a Dirichlet law of parameter α. A second parameter ϕ is drawn according to the same Dirichlet law of parameter β. Then, to generate every word w of the document c, a latent topic z is drawn from a multinomial distribution on θ. Knowing this topic z, the distribution of the words is a multinomial of parameters ϕ. The parameter θ is drawn for all the documents from the same *prior* parameter α. This allows to obtain a parameter binding all the documents together [2].

Several techniques have been proposed to estimate LDA parameters, such as Variational Methods [2], Expectation-propagation [17] or Gibbs Sampling [18]. Gibbs Sampling is a special case of Markov-chain Monte Carlo (MCMC) [19] and
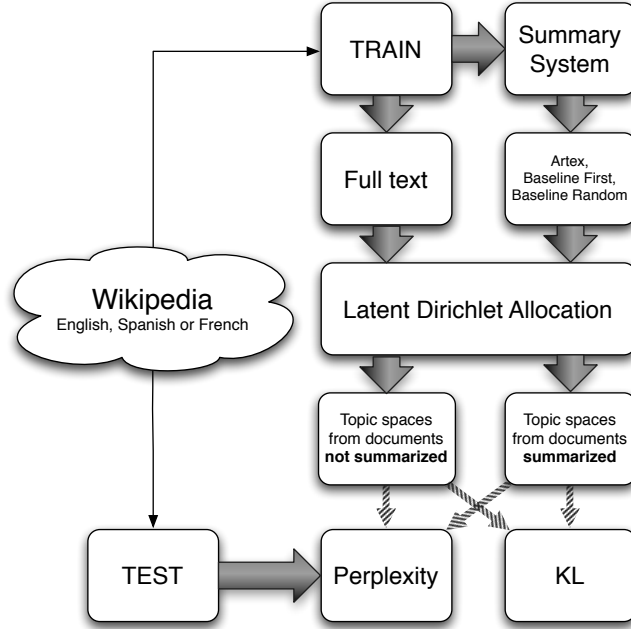
Fig. 1. Overview of the proposed approach.

gives a simple algorithm to approximate inference in high-dimensional models such as LDA [20]. This overcomes the difficulty to directly and exactly estimate parameters that maximize the likelihood of the whole data collection defined as: $p(W|\overrightarrow{\alpha}, \overrightarrow{\beta}) = \prod_{m=1}^{M} p(\overrightarrow{w}_m|\overrightarrow{\alpha}, \overrightarrow{\beta})$ for the whole data collection $W = \{\overrightarrow{w}_m\}_{m=1}^{M}$ knowing the Dirichlet parameters $\overrightarrow{\alpha}$ and $\overrightarrow{\beta}$.

The first use of Gibbs Sampling for estimating LDA is reported in [18] and a more comprehensive description of this method can be found in [20].

The next section describes the income of the LDA technique. The input of the LDA method is an automatic summary of each document of the train corpus. These summaries are built with different systems.

3.2 Automatic text summarization systems

Various text summarization systems have been proposed over the years [21]. Two baseline systems as well as the ARTEX summarization system, that reaches state-of-the-art performance [22], are presented in this section.

Baseline first (BF) The *Baseline first* (or leadbase) selects the n first sentences of the documents, where n is determined by a *compression rate*. Although very
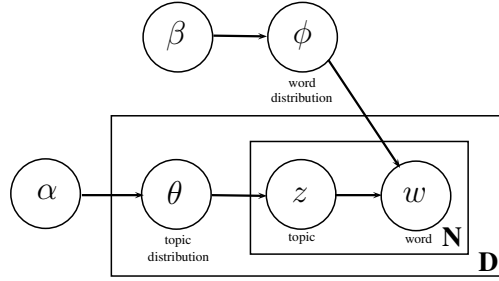
Fig. 2. LDA Formalism.

simple, this method is a strong baseline for the performance of any automatic summarization system [23, 24]. This very old and very simple sentence weighting heuristic does not involve any terms at all: it assigns highest weight to the first sentences of the text. Texts of some genres, such as news reports or scientific papers, are specifically designed for this heuristic: *e.g.*, any scientific paper contains a ready summary at the beginning. This gives a baseline [25] that proves to be very hard to beat on such texts. It is worth noting that in Document Understanding Conference (DUC) competitions [25] only five systems performed above this baseline, which does not demerit the other systems because this baseline is genre-specific.

Baseline random (BR) The *Baseline random* [23] randomly selects n sentences of the documents, where n is also determined by a *compression rate*. This method is the classic baseline for measuring the performance of automatic summarization systems.

ARTEX **A***nothe***R TEX***t* (ARTEX) algorithm [22] is another simple extractive algorithm. The main idea is to represent the text in a suitable space model (VSM). Then, an average document vector that represents the average (the "global topic") of all sentence vectors is constructed. At the same time, the "lexical weight" for each sentence, *i.e.* the number of words in the sentence, is obtained. After that, the angle between the average document and each sentence is calculated. Narrow angles α indicate that the sentences near the "global topic" should be important and are therefore extracted. See Figure 3 for the VSM of words: p vector sentences and the average "global topic" are represented in a N dimensional space of words. The angle α between the sentence $\overrightarrow{s_\mu}$ and the global topic $\overrightarrow{b}$ is processed as follow:

$$cos(\alpha) = \frac{\overrightarrow{b} \times \overrightarrow{s_\mu}}{|| \overrightarrow{b} ||.||\overrightarrow{s_\mu}||} \tag{1}$$
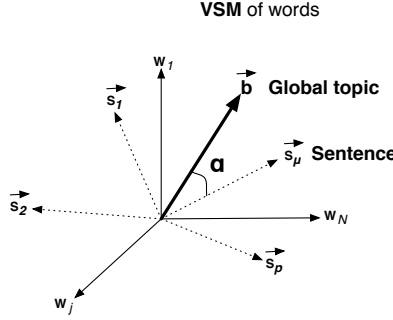
Fig. 3. The "global topic" in a Vector Space Model of N words.

Next, a weight for each sentence is calculated using their proximity with the "global topic" and their "lexical weight". In Figure 4, the "lexical weight" is represented in a VSM of p sentences. Narrow angles indicate that words closest to the "lexical weight" should be important. Finally, the summary is generated concatenating the sentences with the highest scores following their order in the original document. Formally, ARTEX algorithm computes the score of each sentence by calculating the inner product between a sentence vector, an average pseudo-sentence vector (the "global topic") and an *average pseudo-word* vector(the"lexical weight"). Once the pre-processing is complete, a matrix $S_{[pN]}$ (N words and p sentences) is created. Let $\overrightarrow{s_\mu} = (s_{\mu,1}, s_{\mu,2}, ..., s_{\mu,N})$ be a vector of the sentence $\mu = 1, 2, ..., p$. The *average pseudo-word* vector $\overrightarrow{a} = [a\mu]$ was defined as the average number of occurrences of N words used in the sentence $\overrightarrow{s_\mu}$:

$$a_\mu = \frac{1}{N} \sum_j s_{\mu,j} \tag{2}$$

and the *average pseudo-sentence* vector $\overrightarrow{b} = [b_j]$ as the average number of occurrences of each word j used through the p sentences:

$$b_j = \frac{1}{p} \sum_\mu s_{\mu,j} \tag{3}$$

The weight of a sentence $\overrightarrow{s_\mu}$ is calculated as follows:

$$w(\overrightarrow{s_\mu}) = (\overrightarrow{s} \times \overrightarrow{b}) \times \overrightarrow{a}$$

$$= \frac{1}{N_p} \left(\sum_{j=1}^{N} s_{\mu,j} \times b_j \right) \times a_\mu \; ; \mu = 1, 2, \ldots, p \tag{4}$$

The $w(\bullet)$ computed by Equation 4 must be normalized between the interval $[0, 1]$. The calculation of $(\overrightarrow{s} \times \overrightarrow{b})$ indicates the proximity between the sentence
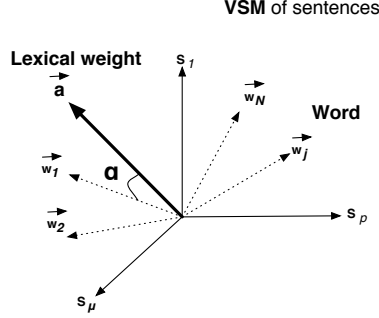
Fig. 4. The "lexical weight" in a Vector Space Model of p sentences.

$\overrightarrow{s_\mu}$ and the *average pseudo-sentence* $\overrightarrow{b}$. The product $(\overrightarrow{s} \times \overrightarrow{b}) \times \overrightarrow{a}$ weight this proximity using the *average pseudo-word* $\overrightarrow{a}$. If a sentence $\overrightarrow{s_\mu}$ is near $\overrightarrow{b}$ and their corresponding element a_μ has a high value, therefore $\overrightarrow{s_\mu}$ will have a high score. Moreover, a sentence $\overrightarrow{s_\mu}$ far from a main topic (i.e. $\overrightarrow{s_\mu} \times \overrightarrow{b}$ is near 0) or their corresponding element $a_m u$ has a low value, (i.e. $a_m u$ are near 0), therefore $\overrightarrow{s_\mu}$ will have a low score.

It is not really necessary to divide the scalar product by the constant $\frac{1}{N_p}$, because the angle α between $\overrightarrow{b}$ and $\overrightarrow{s_\mu}$ is the same if $\overrightarrow{b} = \overrightarrow{b'} = \sum_\mu s_{\mu,j}$. The element a_μ is only a scale factor that does not modify α [22]:

$$w(\overrightarrow{s_\mu})* = \frac{1}{\sqrt{N^5 p^3}} \left(\sum_{j=1}^{N} s_{\mu,j} \times b_j \right) \times a_\mu \; ; \mu = 1, 2, \dots, p \qquad (5)$$

The term $1/\sqrt{N^5 p^3}$ is a constant value, and then $w(\bullet)$ (Equation 4) and $w(\bullet)*$ (Equation 5) are both equivalent.

This summarization system outperforms the CORTEX [26] one with the FRESA [15, 14] measure. ARTEX is evaluated with several corpus such as the *Medecina Clinica* [22]. ARTEX performance is then better than CORTEX on English, Spanish or French, which are the targeted languages in this study.

4 Evaluation of LDA model quality

The previous section described different summarization systems to reduce the size of train corpus and to retain only relevant information contained into the train documents. This section proposes a set of metrics to evaluate the quality of topic spaces generated from summaries of the train documents. The first one is the perplexity. This score is the most popular one. We also propose to study

another measure to evaluate the dispersion of each word into a given topic space. This measure is called the Jensen-Shannon ($\mathcal{JS}$) divergence.

4.1 Perplexity

Perplexity is a standard measure to evaluate topic spaces, and more generally a probabilistic model. A topic model $\mathcal{Z}$ is effective if it can correctly predict an unseen document from the test collection. The perplexity used in language modeling is monotonically decreasing in the likelihood of the test data, and is algebraically equivalent to the inverse of the geometric mean per-word likelihood. A lower perplexity score indicates better generalization performance [2]:

$$perplexity(\mathcal{B}) = exp\left\{-\frac{1}{N_\mathcal{B}} \sum_{d=1}^{M} \log P(\mathbf{w})\right\} \qquad (6)$$

with

$$N_\mathcal{B} = \sum_{d=1}^{M} N_d \qquad (7)$$

where $N_\mathcal{B}$ is the combined length of all M testing terms and N_d is the number of words in the document d; $P(\mathbf{w})$ is the likelihood that the generative model will be assigned to an unseen word $\mathbf{w}$ of a document d in the test collection. The quantity inside the exponent is called the *entropy* of the test collection. The logarithm enables to interpret the entropy in terms of *bits* of information.

4.2 Jensen-Shannon ($\mathcal{JS}$) divergence

The perplexity evaluates the performance of a topic space. Another important information is the distribution of words in each topic. The Kullback-Leibler divergence ($\mathcal{KL}$) estimates how much a topic is different from the N topics contained in the topic model. This distribution is defined hereafter:

$$\mathcal{KL}(z_i, z_j) = \sum_{w \in \mathcal{A}} p_i \log \frac{p_i}{p_j} \qquad (8)$$

where $p_i = P(w|z_i)$ and $p_j = P(w|z_j)$ are the probabilities that the word w is generated by the topic z_i or z_j. Thus, the symmetric $\mathcal{KL}$ divergence is named Jensen-Shannon ($\mathcal{JS}$) divergence metric. It is the mid-point measure between $\mathcal{KL}(z_i, z_j)$ and $\mathcal{KL}(z_j, z_i)$. $\mathcal{JS}$ is then defined with equation 8 as the mean of the divergences between (z_i, z_j) and (z_j, z_i) as:

$$\begin{aligned}
\mathcal{JS}(z_i, z_j) &= \frac{1}{2} \left(\mathcal{KL}(z_i, z_j) + \mathcal{KL}(z_j, z_i)\right) \\
&= \frac{1}{2} \sum_{w \in \mathcal{A}} \left(p_i \log \frac{p_i}{p_j} + p_j \log \frac{p_j}{p_i}\right) .
\end{aligned} \qquad (9)$$

The $\mathcal{JS}$ divergence for the entire topic space is then defined as the divergence between each pair of topics composing the topic model $\mathcal{Z}$, defined in equation 9 as:

$$\mathcal{JS}(\mathcal{Z}) = \sum_{z_i \in \mathcal{Z}} \sum_{z_j \in \mathcal{Z}} \mathcal{JS}(z_i, z_j)$$
$$= \frac{1}{2} \sum_{z_i \in \mathcal{Z}} \sum_{z_j \in \mathcal{Z}} \sum_{w \in \mathcal{A}} p_i \log \frac{p_i}{p_j} + p_j \log \frac{p_j}{p_i}. \tag{10}$$

if $i = j \Rightarrow \log \frac{p_j}{p_i} = 0$ ($log1 = 0$). After defining the metrics to evaluate the quality of the model, the next section describes the experiment data sets and the experimental protocol.

5 Experiments

These summarization systems are used to compress and retain only relevant information into train text collection in each language. This section presents the experiments processed to evaluate the relevance and the effectiveness of the proposed system of fast and robust topic space building. First of all, the experimental protocol is presented, and then a qualitative analysis of obtained results is performed using evaluation metrics described in Section 4.

5.1 Experimental protocol

In order to train topic spaces, a large corpus of documents is required. Three corpus was used. Each corpus $\mathcal{C}$ is in a particular language (English, Spanish and French), and is composed of a training set $\mathcal{A}$ and a testing set $\mathcal{B}$. The corpus are composed of articles from Wikipedia. Thus, for each of the three languages, a set of 100,000 documents is collected. 90% of the corpus is summarized and used to build topic spaces, while 10% is used to evaluate each model (no need to be summarized).

Table 1 shows that the latin languages (French and Spanish) have a similar size (a difference of less than 4% is observed), while the English one is bigger than the others (English text corpus is 1.37 times bigger than French or Spanish corpus). In spite of the size difference of corpus, both of them have more or less the same number of words and sentences in an article. We can also note that the English vocabulary size is roughly the same (15%) than the latin languages. Same observations can be made in Table 2, that presents statistics at document level (mean on the whole corpus). In next section, the outcome of this fact is seen during the perplexity evaluation of topic spaces built from English train text collection.

As set of topic spaces is trained to evaluate the perplexity and the Jensen-Shannon ($\mathcal{JS}$) scores for each language, as well as the processing time to summarize and compress documents from the train corpus. Following a classical

Table 1. Dataset statistics of the Wikipedia corpus.

Language	#Words	#Unique Words	#Sentences
English	30,506,196	2,133,055	7,271,971
Spanish	23,742,681	1,808,828	5,245,507
French	25,545,425	1,724,189	5,364,825

Table 2. Dataset statistics per document of the Wikipedia corpus.

Language	#Words	#Unique Words	#Sentences
English	339	24	81
Spanish	264	20	58
French	284	19	56

study of LDA topic spaces quality [27], the number of topics by model is fixed to $\{5, 10, 50, 100, 200, 400\}$. These topic spaces are built with the MALLET toolkit [28].

5.2 Results

The experiments conducted in this paper are topic-based concern. Thus, each metric proposed in Section 4 (Perplexity and $\mathcal{JS}$) are applied for each language (English, Spanish and French), for each topic space size ($\{5, 10, 50, 100, 200, 400\}$), and finally, for each compression rate during the summarization process (10% to 50% of the original size of the documents). Figures 5 and 6 present results obtained by varying the number of topics (Figure (a) to (c)) and the percentage of summary (Figure 6), respectively for perplexity and Jensen-Shannon ($\mathcal{JS}$) measures. Results are computed with a mean among the various topic spaces size and a mean among the different reduced summaries size. Moreover, each language was study separately to point out differences of topic spaces quality depending on the language.
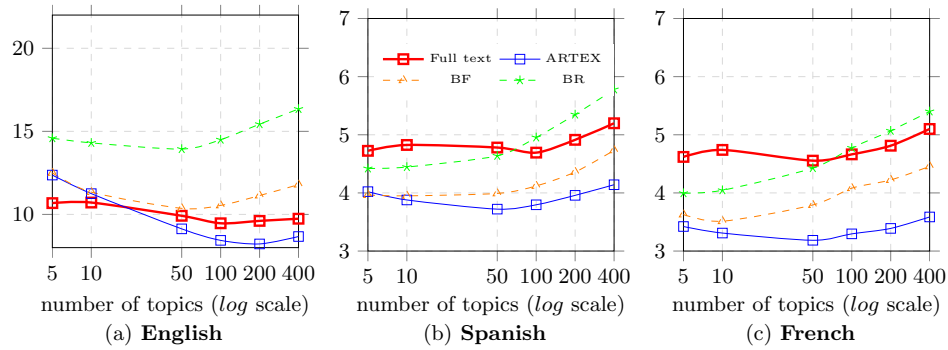
Fig. 5. Perplexity ($\times 10^{-3}$) by varying the number of topics for each corpus.
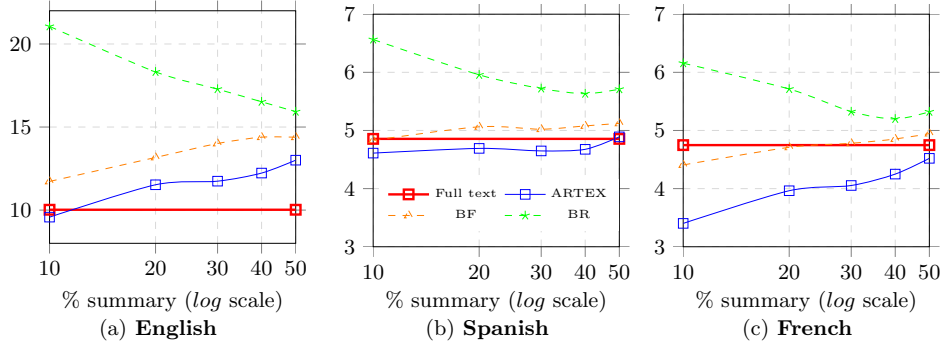
(a) **English** (b) **Spanish** (c) **French**

Fig. 6. Perplexity ($\times 10^{-3}$) by varying the % summary for each corpus.

6 Discussions

The results reported in Figures 5 and 6 allow us to point out a first general remark, already observed in section 5.1: the two latin languages have more or less the same tendencies. This should be explained by the root of these languages, which are both latins.

Figure 5 shows that the Spanish and French corpus obtain a perplexity between 3,000 and 6,100 when the number of classes in the topic space varies. Another observation is that, for these two languages, topic spaces obtained with summarized documents, outperform the ones obtained with complete documents when at least 50 topics are considered (Figures 5-b and -c). The best system for all languages is ordered in the same way. Systems are ordered from the best to the worst in this manner: ARTEX, BF (this fact is explained in the next part and is noted into $\mathcal{JS}$ measure curves in Figures 7 and 8), and then BR. If we considerer a number of topics up to 50, we can note that the topic spaces, from full text documents (*i.e.* not summarized) with an English text corpus, obtain a better perplexity (smaller) than documents processed with a summarization system, that is particularly visible into Figures 6.

To address the shortcoming due to the size of the English corpus (1.37 times bigger than latin languages), the number of topics contained into the thematic space has to be increased to effectively disconnect words into topics. In spite of moving up, the number of topics move down the perplexity of topic spaces for all summarization systems (except random baseline (RB)), the perplexity obtained with the English corpus being higher than those obtained from the Spanish and French corpus.

Among all summarization systems used to reduce the documents from the train corpus, the baseline first (BF) obtains good results for all languages. This performance is due to the fact that BF selects the first paragraph of the document as a summary: when a Wikipedia content provider writes a new article, he exposes the main idea of the article in the first sentences. Furthermore, the rest of the document relates different aspects of the article subject, such as historical or

economical details, which are not useful to compose a relevant summary. Thus, this baseline is quite hard to outperform when the documents to summarize are from encyclopedia such as Wikipedia.
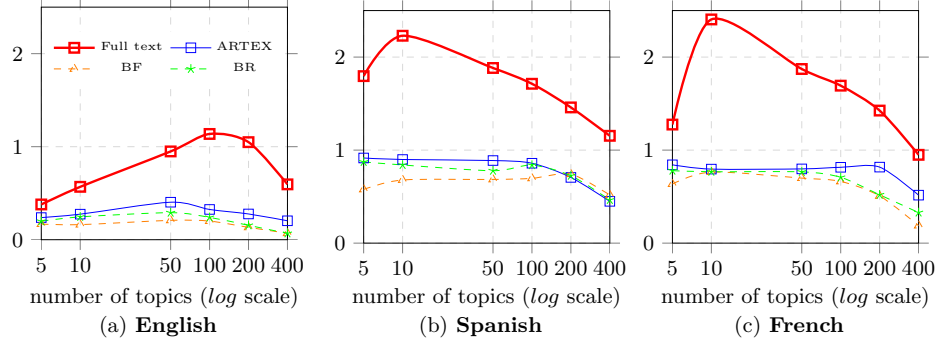
Fig. 7. Jensen-Shannon $(\times 10^3)$ measure by varying the number of topics for each corpus.
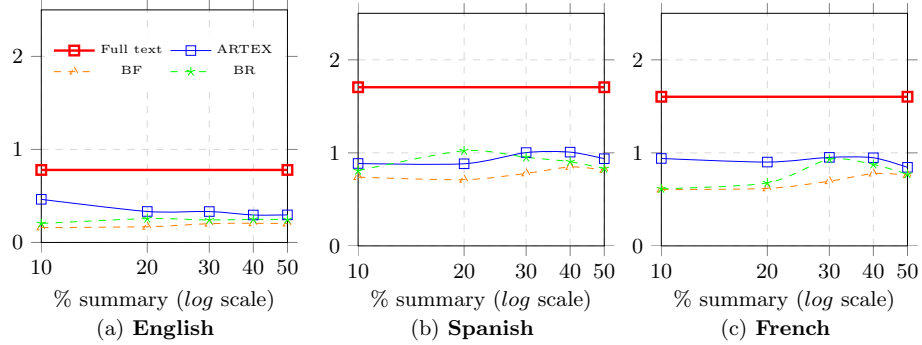
Fig. 8. Jensen-Shannon $(\times 10^3)$ measure by varying the % summary for each corpus.

The random baseline (RB) composes its summary by randomly selecting a set of sentences in an article. This kind of system is particularly relevant when the main ideas are disseminated in the document such as a blog or a website. This is the main reason why this baseline did not obtain good results except for $\mathcal{JS}$ divergence measure (see Figures 7 and 8). This can be explained by the fact that this system selects sentences at different places, and then, selects a variable set of words. Thus, topic spaces from these documents contain a variable vocabulary. The $\mathcal{JS}$ divergence evaluates how much a word contained

in a topic is discriminative, and allows to distinguish this topic from the others that compose the thematic representation.

Figures 7 and 8 also show that Jensen-Shannon ($\mathcal{JS}$) divergence scores between topics obtained a similar performance order of summarization systems for all languages corpus. Moreover, full text documents always outperform all topic spaces representation for all languages and all summary rates. The reason is that full text documents contain a larger vocabulary, and $\mathcal{JS}$ divergence is sensitive to the vocabulary size, especially when the number of topics is equal for summarized and full text documents. This observation is pointed out by Figures 8-b and -c where the means among topic spaces for each summary rate of full text documents are beyond all summarization systems. Last points of the curves show that topic spaces, with a high number of topics and estimated from summaries, do not outperform those estimated from full text documents, but become more and more closer to these ones: this confirms the original idea that have motivated this work.

Tables 3 and 4 finally present the processing time, in seconds, by varying the number of topics for each language corpus, respectively with the full text and the summarized documents. We can see that processing time is saved when topic spaces are learned from summarized documents. Moreover, tables show that the processing times follow an exponential curve, especially for the full text context. For this reason, we can easily imagine the processing time that can be saved using summaries instead of the complete documents, which inevitably contain non informative and irrelevant terms.

Table 3. Processing time (in seconds) by varying the number of topics for each corpus.

System	Language		
Full Text	English	Spanish	French
5	1,861	1,388	1,208
10	2,127	1,731	1,362
50	4,194	2,680	2,319
100	5,288	3,413	3,323
200	6,364	4,524	4,667
400	8,654	6,625	6,751

Table 4. Processing time (in seconds) by varying the number of topics for each corpus.

System	Language			System	Language			System	Language		
ARTEX	English	Spanish	French	BR	English	Spanish	French	BF	English	Spanish	French
5	514	448	394	5	318	265	238	5	466	301	276
10	607	521	438	10	349	298	288	10	529	348	317
50	1,051	804	709	50	466	418	465	50	1031	727	459
100	1,565	1,303	1,039	100	652	602	548	100	1,614	737	680
200	2,536	2,076	1,573	200	919	863	838	200	2,115	814	985
400	3,404	2,853	2,073	400	1,081	988	978	400	2,784	1,448	988

A general remark is that the best summarization system is ARTEX, but if we take into account the processing time during the topic space learning, the baseline first (BF) is the best agreement. Indeed, if one want to find a common ground between a low perplexity, a high $\mathcal{JS}$ divergence between topics and a fast learning process, the BF method should be chosen.

7   Conclusions

In this paper, a qualitative study of the impact of documents summarization in topic space learning is proposed. The basic idea that learning topic spaces from compressed documents is less time consuming than learning topic spaces from the full documents is noted. The main advantage to use the full text document in text corpus to build topic space is to move up the semantic variability into each topic, and then increase the divergence between these ones. Experiments show that topic spaces with enough topics size have more or less (roughly) the same divergence.

Thus, topic spaces with a large number of topics, *i.e.* suitable knowing the size of the corpus (more than 200 topics in our case), have a lower perplexity, a better divergence between topics and are less time consuming during the LDA learning process. The only drawback of topic spaces learned from text corpus of summarized documents disappear when the number of topics comes up suitable for the size of the corpus whatever the language considered.

References

1. Salton, G.: Automatic text processing: the transformation. Analysis and Retrieval of Information by Computer (1989)
2. Blei, D., Ng, A., Jordan, M.: Latent dirichlet allocation. The Journal of Machine Learning Research **3** (2003) 993–1022
3. Baeza-Yates, R., Ribeiro-Neto, B., et al.: Modern information retrieval. Volume 463. ACM press New York (1999)
4. Salton, G., McGill, M.J.: Introduction to modern information retrieval. (1983)
5. Salton, G., Yang, C.S.: On the specification of term values in automatic indexing. Journal of documentation **29** (1973) 351–372
6. Deerwester, S., Dumais, S., Furnas, G., Landauer, T., Harshman, R.: Indexing by latent semantic analysis. Journal of the American society for information science **41** (1990) 391–407
7. Bellegarda, J.: A latent semantic analysis framework for large-span language modeling. In: Fifth European Conference on Speech Communication and Technology. (1997)
8. Hofmann, T.: Probabilistic latent semantic analysis. In: Proc. of Uncertainty in Artificial Intelligence, UAI ' 99, Citeseer (1999) 21
9. Bellegarda, J.: Exploiting latent semantic information in statistical language modeling. Proceedings of the IEEE **88** (2000) 1279–1296
10. Suzuki, Y., Fukumoto, F., Sekiguchi, Y.: Keyword extraction using term-domain interdependence for dictation of radio news. In: 17th international conference on Computational linguistics. Volume 2., ACL (1998) 1272–1276

11. Popescul, A., Pennock, D.M., Lawrence, S.: Probabilistic models for unified collaborative and content-based recommendation in sparse-data environments. In: Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence, Morgan Kaufmann Publishers Inc. (2001) 437–444
12. Louis, A., Nenkova, A.: Automatically Evaluating Content Selection in Summarization without Human Models. In: Empirical Methods in Natural Language Processing, Singapore (2009) 306–314
13. Lin, J.: Divergence Measures based on the Shannon Entropy. IEEE Transactions on Information Theory **37** (1991)
14. Saggion, H., Torres-Moreno, J.M., da Cunha, I., SanJuan, E.: Multilingual summarization evaluation without human models. In: 23rd Int. Conf. on Computational Linguistics. COLING '10, Beijing, China, ACL (2010) 1059–1067
15. Torres-Moreno, J.M., Saggion, H., Cunha, I.d., SanJuan, E., Velázquez-Morales, P.: Summary evaluation with and without references. Polibits (2010) 13–20
16. Hofmann, T.: Unsupervised learning by probabilistic latent semantic analysis. Machine Learning **42** (2001) 177–196
17. Minka, T., Lafferty, J.: Expectation-propagation for the generative aspect model. In: Proceedings of the Eighteenth conference on Uncertainty in artificial intelligence, Morgan Kaufmann Publishers Inc. (2002) 352–359
18. Griffiths, T.L., Steyvers, M.: Finding scientific topics. Proceedings of the National academy of Sciences of the United States of America **101** (2004) 5228–5235
19. Geman, S., Geman, D.: Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. IEEE Transactions on Pattern Analysis and Machine Intelligence (1984) 721–741
20. Heinrich, G.: Parameter estimation for text analysis. Web: http://www. arbylon. net/publications/text-est. pdf (2005)
21. Torres-Moreno, J.M.: Automatic Text Summarization. Wiley and Sons (2014)
22. Torres-Moreno, J.M.: Artex is another text summarizer. arxiv:1210.3312 [cs.ir] (2012)
23. Ledeneva, Y., Gelbukh, A., García-Hernández, R.A.: Terms derived from frequent sequences for extractive text summarization. In: Computational Linguistics and Intelligent Text Processing. Springer (2008) 593–604
24. Manning, C.D., Schütze, H.: Foundations of Statistical Natural Language Processing. The MIT Press, Cambridge, Massachusetts (1999)
25. DUC: Document Understanding Conference. (2002)
26. Torres-Moreno, J.M., Velázquez-Morales, P., Meunier, J.G.: Cortex : un algorithme pour la condensation automatique des textes. In: ARCo'01. Volume 2., Lyon, France (2001) 365–366
27. Rosen-Zvi, M., Griffiths, T., Steyvers, M., Smyth, P.: The author-topic model for authors and documents. In: Proceedings of the 20th conference on Uncertainty in artificial intelligence, AUAI Press (2004) 487–494
28. McCallum, A.K.: Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu (2002)