Empirical study of LDA for Arabic topic identification

Marwa Naili, Anja Habacha Chaibi and Henda Ben Ghézala

RIADI-ENSI
University of Manouba
Manouba 2010, Tunisia
maroua.naili@riadi.rnu.tn

RÉSUMÉ. Cet article met l'accent sur l'identification thématique pour la langue arabe. Nous étudions l'Allocation de Dirichlet Latente (LDA) comme une méthode non supervisée pour l'identification thématique. Ainsi, une étude approfondie de LDA a été effectuée à deux niveaux: le processus de lemmatisation et le choix des paramètres. Pour le premier niveau, nous étudions l'effet des différents lemmatiseurs sur LDA. Pour le deuxième niveau, nous nous focalisons sur les paramètres de LDA et leurs impacts sur l'identification. Cette étude montre que LDA est une méthode efficace pour l'identification thématique Arabe surtout avec le bon choix des paramètres. Un autre résultat important est l'impact élevé des lemmatiseurs sur l'identification thématique.

ABSTRACT. This paper focuses on the topic identification for the Arabic language. We study the Latent Dirichlet Allocation (LDA) as an unsupervised method for the Arabic topic identification. Thus, a deep study of LDA is carried out at two levels: Stemming process and the choice of LDA parameters. For the first one, we study the effect of different Arabic stemmers on LDA. For the second one, we focus on LDA parameters and their impact on the topic identification. This study shows that LDA is an efficient method for Arabic topic identification especially with the right choice of parameters. Another important result is the high impact of stemming algorithms on topic identification.

MOTS-CLÉS : Identification thématique, Allocation de Dirichlet Latente, paramètres de LDA, lemmatiseurs Arabes.

KEYWORDS: Topic identification, Latent Dirichlet Allocation, LDA parameters, Arabic stemmers.

1. Introduction

During the last few years, the number of textual documents has been vastly increasing. Thus, many techniques have been presented to deal with this big number of documents. However, the real challenge is to manage these documents based on their content, especially the thematic one. For this reason, topic Identification and classification draw a lot of intention in research fields dealing with different types of documents (text [7], XML [2], etc). Yet for Arabic textual documents, there is a flagrant lack of research. This can be explained by the high complexity of this language and the lack of Arabic resources. In this paper, we will focus on topic identification by studding LDA as an unsupervised method for Arabic topic identification.

This paper is organized as follows: Section 2 presents an overview of Arabic topic identification; Section 3 describes some Arabic stemmers; Section 4 deals with LDA; Section 5 is dedicated to the evaluation and the discussion; finally, the conclusion and future works are presented in section 6.

2. Overview of Arabic topic identification

Topic identification is the process of identifying the topic of a textual unity. According to most researchers, a topic is a cluster of words which are closely related to the topic. Clusters depend on the stemming process that specifies the type of words (root, stem, etc). For the Arabic topic identification, some methods have been used as:

- *TF-IDF [7]:* allows the construction of a vector space. Each vector represents a document by the combination between TF(w,d) and IDF(w). The topic with the highest similarity with the document will be considered as the document's topic.

- *SVM and MSVM* [13]: is a supervised method which classifies documents into two classes by constructing a hyperplane separator in the R^N vector space. Yet, when the number of categories is superior than 2, the MSVM is used. In fact, the idea of this method is to find *n* hyperplane with *n* corresponds to the number of categories.

-*TR-Classsifier* [7]: is based on triggers which are identified by using the Average Mutual Information. In fact, topics and documents are presented by triggers which are a set of words that have the highest degree of correlation. Then, based on the TR-distance, the similarity is calculated between triggers to identify the topic of the document.

-*Named Entities approach* [10]: The idea of this approach is to reduce the dimension of vectors by using only the segments bounded by named entities pairs. Then, the mutual information is used to calculate similarity between topics and documents. Besides these methods, we can cite other methods used for topic identification such as

TULM and Neural networks in [7]. However, the major limit of these methods is that a training step is necessary to identify the topics and to construct a vocabulary for each topic. Thus, we opted to use the unsupervised method LDA. That means that there is no need to a training step because topics are identified in the process of topic identification.

3. Arabic stemmers

Arabic language is one of the most complex and ambiguous language because of its wide variety of grammatical forms and its complex morphology. Thus, the stemming process is more difficult for the Arabic language than other languages. The stemming process aims to find the lexical root or lemma of words by removing prefixes and suffixes with are attached to its root. As an example of Arabic stemmers we mention:

- *Khoja Stemmer [11]:* it extracts the root of a word by removing the longest suffix and prefix and then by matching the rest with verbal and nouns patterns.
- *ISRI Arabic Stemmer* [5]: it extracts the root of a word. But, unlike Khoja Stemmer, it doesn't use any root dictionary or lexicon.
- *The Buckwalter Arabic Morphological Analyzer* [12]: it returns the stems of words based on lexicons of stems, prefixes, suffixes and morphological compatibility tables.
- *Light Stemmer* [6]: Unlike Khoja Stemmer, it removes some defined prefixes and suffixes instead of extracting the original root words.

According to different studies [5,6] the most efficient stemmers are Khoja and Light Stemmers. These two stemmers are available freely on the web and might be the only available Open Source ones. Thus, we will study Khoja and Light Stemmers to evaluate the effect of the stemming process on the topic identification.

4. Latent dirichlet allocation (LDA)

LDA [3] is a generative model in which documents are represented as a mixture of topic. Each topic is a multinomial distribution over words that depends on the stemming process. Therefore, for each document w in the corpus D, the generative process is:

1. We choose N (a document is a sequence of N words) according to Poisson distribution (N ~ Poisson(ξ))
2. We choose θ (θ_d is the distribution over the topic of the document d) according to dirichlet allocation ($\theta \sim Dirichlet(\alpha)$)
3. For each of the N words w_n : Choose a latent topic z_n according to a multinomial distribution and choose a word w_n from $p\ (w_n|z_n, \beta)$

The θ variable takes values in the *(k-1)* simplex and its density is equal to:

$$p(\theta|\alpha) = \frac{\Gamma\left(\sum_{i=1}^{k} \alpha_i\right)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \theta_1^{\alpha_1 - 1} \dots \theta_k^{\alpha_k - 1} \qquad (1)$$

Where $\alpha \in \mathbb{R}^k$, $\alpha_i > 0$ and $\Gamma(x)$ is the Gamma function.

Therefore, given α and β, the joint distribution of θ, z and w is equal to:

$$p(\theta, z, w|\alpha, \beta) = p(\theta|\alpha) \prod_{n=1}^{N} p(z_n|\theta)\ p(w_n|z_n, \beta) \qquad (2)$$

Finally, by integrating over θ and summing over z, the marginal distribution of a document is as follow (equation 3):

$$p(w|\alpha, \beta) = \int p(\theta|\alpha)\left(\prod_{n=1}^{N} \sum_{z_n} p(z_n|\theta) p(w_n|z_n, \beta)\right) d\theta \qquad (3)$$

According to Steyvers and Griffiths [8], the choice of α and β has an effect on the performance of LDA. Besides, these parameters depend on the number of topics and the vocabulary size. Moreover, Steyvers and Griffiths [8] recommended to use $\alpha = 50/k$ and $\beta = 0.01$. However, Lu et al. [14] conduct an in-depth analysis of the choice of α with $\beta = 0.01$. According to this analysis, the performance of LDA is influenced by the initializing choiceα. This choice also depends on the field of application such as topic classification and information retrieval which are tested in this study. As result, they found that, for the topic classification, the optimal performance is obtained by α between 0.1 and 0.5. Yet, for information retrieval, the optimal performance is obtained by α between 0.5 and 2. However, according to Lu et al. [14], the best value of α is not stable and it depends on the collection of documents used for tests. On the other hand, Heinrich [4] estimated the values of α and β by using the information available from the Gibbs sampler. In fact, Heinrich [4] showed that hyper-parameters are best estimated as parameters of the Dirichlet-multinomial distribution.

Despite the high performance of LDA, few works dealing with LDA were presented in the field of Arabic topic identification [9,1]. According to these works, promising results have been obtained by LDA. However, we note that no one has studied LDA parameters in the field of topic identification. Therefore, in this paper, we will study in depth the LDA by studding the choice of α and more important the effect of different stemming algorithms to enhance the quality of topic identification.

5. Evaluation and discussion

In this section, we evaluated LDA with different stemmers. Thus, we presented three different versions: LDA-WS (**W**ithout **S**temmer), LDA-KS (**K**hoja **S**temmer) and LDA-LS (**L**ight **S**temmer). For this evaluation, we use the Arabic benchmark Al-Watan

which contains 20291 articles from Watan newspaper and it covers six topics: culture (2782 documents), economy (3468 documents), international news (2035 documents), local news (3596 documents), religion (3860 documents) and sport (4550 documents). To report the evaluation results, we use three metrics: **R**ecall, **P**recision and **F**-measure.

5.1. Identified topics based on different stemmers

	Culture	Economy	International News	Local News	Religion	Sport
LDA-WS	الله (god)	مليون (million)	قال (said)	السلطنة (sultanate)	الله (god)	المباراة (match)
	الإسلام (Islam)	ريال (real)	العراق (Iraq)	العمل (work)	قال (said)	المنتخب (team)
	الحياة (life)	عام (public)	المتحدة (united)	العام (the public)	صلى (pray)	الأول (first)
	الناس (people)	الدول (countries)	الأميركية (American)	العامة (the public)	وسلم (salaam)	المركز (position)
	الإسلامية(Islamic)	السلطنة(sultanate)	عام (public)	محمد (Mohammed)	رسول (prophet)	الثاني (second)
LDA-KS	علم(knowledge)	دول (countries)	عرق (vein)	جمع (collect)	سلم (salaam)	لعب (play)
	كون (universe)	شرك (share)	روس (Russian)	دور (role)	قول (saying)	فرق (teams)
	عمل (work)	عمل (work)	حكم (rule)	علم (knowledge)	صلى (pray)	نخب (pledge)
	كتب (write)	صنع (production)	دول (countries)	عمل (work)	كون (universe)	دور (role)
	جمع (collect)	عوم (launch)	عمل (work)	قوم (nation)	رسل (Russell)	بطل (champion)
LDA-LS	اسلام (Islam)	شرك (share)	عراق (Iraq)	عام (public)	قال (said)	فريق (team)
	عرب (Arab)	عام (public)	اميرك (American)	عمل (work)	صل (pray)	منتخب (team)
	فن (art)	اقتصاد(economy)	دول (countries)	عمان (Amman)	رسول (prophet)	دور (role)
	كتاب (book)	دول (countries)	قال (said)	دور (role)	سلم (salaam)	مبارا (match)
	عالم (world)	قطاع (sector)	رئيس (president)	تعليم (education)	مسلم (Muslim)	بطول(championship)

Figure 1. ***Identified topics based on LDA-WS, LDA-KS and LDA-LS*.**

By conducting the three versions of LDA on AL-Watan corpus, we were able to identify all the six topics. As shown in Figure.1, the identified topics depend on the used stemmer. In fact, without using any stemming algorithms, the different topics were successfully identified by LDA-WS. However, the problem is that some words can figure more than once with different affix or suffix such as العام and العامة which mean public. This problem is resolved by using Khoja stemmer which extracts the root of words. Thus, by employing LDA-KS, the topics are present by roots. The limit of this method is that a root can have several meaning such as علم which has many meaning like: knowledge, flag, aware. Therefore, by using Khoja Stemmer, we might lose the meaning. Yet, Light Stemmer removes only the prefix to maintain the meaning such as the word المنتخب (the team) without stemming, نخب (pledge) with Khoja Stemmer and منتخب (team) with Light Stemmer. As conclusion, all the six topics have been successfully identified by LDA. Moreover, Light Stemmer is the most efficient stemmer because it solves the problem of repetition (which is caused by the absence of stemmer: LDA-WS) and the loss of meaning (which is caused by Khoja Stemmer LDA-KS).

5.2. Study of LDA parameter (α)

We study in depth the α parameter of LDA by using three values 0.1, 0.5 and 50/k (k is number of topics which is 6 in our study). These values are proposed by [8,14]. For

β, we used $\beta = 0.01$ which is recommended in most research. For each value of α, the results of LDA-WS, LDA-KS and LDA-LS are illustrates in table 1. First of all, we remark that LDA-LS is independent of α. Yet, LDA-WS and LDA-KS are strongly influenced by α and the best results are obtained by $\alpha = 0.5$. Furthermore, for $\alpha = 0.5$, the results of LDA-LS and LDA-KS are very close. Based on this result and the results of the stemming process for the topic identification, Light Stemmer is the most efficient stemmer to use with LDA. In the other hand, regardless of the value of α and the stemming algorithm, the well identified topics are: sport (F = 91.86%), religion (F = 82.75%), economy (F = 75.13%). Yet, for the other topics, especially the culture topic, the performance of LDA is not stable. This can be explained by the fact that the vocabularies of these topics (culture, international and local news) are very close.

			Culture	Economy	Intern News	Local News	Religion	Sport	Average
LDA-WS	$\alpha = 0.1$	**R**	9.09%	70.10%	95.23%	84.73%	50.34%	85.25%	65.79%
		P	12.02%	**80.95%**	47.53%	**58.73%**	96.00%	**99.59%**	65.80%
		F	10.36%	**75.13%**	63.42%	**69.38%**	66.04%	**91.86%**	62.70%
	$\alpha = 0.5$	**R**	48.56%	70.30%	97.49%	81.01%	61.11%	84.13%	**73.77%**
		P	**46.73%**	79.72%	**67.21%**	56.98%	**97.16%**	99.43%	**74.54%**
		F	**47.63%**	74.72%	**79.57%**	66.90%	**75.03%**	91.14%	**72.50%**
	$\alpha = 50/k$	**R**	46.62%	69.49%	97.59%	80.70%	60.18%	84.28%	73.14%
		P	45.40%	79.04%	66.22%	56.47%	97.11%	99.48%	73.95%
		F	46.00%	73.96%	78.90%	66.44%	74.31%	91.25%	71.81%
LDA-KS	$\alpha = 0.1$	**R**	68.40%	64.27%	78.52%	50.08%	71.35%	75.82%	68.07%
		P	55.53%	57.72%	52.62%	50.75%	93.58%	99.34%	68.26%
		F	61.30%	60.82%	63.01%	50.41%	80.96%	86.00%	67.08%
	$\alpha = 0.5$	**R**	69.55%	54.67%	95.92%	78.28%	73.70%	79.98%	**75.35%**
		P	**55.76%**	**82.87%**	**76.28%**	**53.18%**	**94.33%**	99.29%	**76.95%**
		F	**61.90%**	**65.88%**	**84.98%**	**63.34%**	**82.75%**	**88.59%**	**74.59%**
	$\alpha = 50/k$	**R**	68.44%	63.98%	90.47%	50.78%	70.72%	75.54%	69.99%
		P	54.84%	57.79%	61.02%	50.79%	93.85%	**99.39%**	69.61%
		F	60.89%	60.73%	72.88%	50.78%	80.66%	85.84%	68.63%
LDA-LS	$\alpha = 0.1$	**R**	60.71%	63.32%	97.00%	77.11%	59.09%	83.49%	73.45%
		P	49.38%	**75.88%**	74.18%	54.20%	96.24%	**99.19%**	74.84%
		F	54.47%	**69.03%**	84.07%	63.66%	73.23%	90.67%	72.52%
	$\alpha = 0.5$	**R**	63.73%	62.51%	96.36%	77.14%	65.72%	83.54%	**74.83%**
		P	**54.19%**	75.54%	**75.60%**	**54.57%**	**96.10%**	**99.19%**	**75.86%**
		F	**58.57%**	68.41%	**84.73%**	**63.92%**	78.06%	**90.69%**	**74.06%**
	$\alpha = 50/k$	**R**	62.98%	62.92%	96.46%	76.42%	65.78%	83.36%	74.65%
		P	54.12%	75.47%	75.50%	53.97%	**96.10%**	99.06%	75.70%
		F	58.21%	68.63%	84.70%	63.26%	**78.10%**	90.53%	73.90%

Table 1. ***LDA-WS, LSA-KS and LDA-LS results with*** $\alpha = 0.1$, $\alpha = 0.5$ **and** $\alpha = 50/k$.

But the vocabularies of sport, religion and economy are more representative and unique for each topic which leads to an efficient topic identification.

5.3. Comparison with related works

To evaluate our work, we choose to compare our methods (LDA-KS and LDA-LS) with the works of Abbas et al. [7] and Koulali and Meziane [10]. The reason for this choice is that we used the same test corpus for the evaluation. Yet, we note that in these works [7,10], 90% of the corpus is used for the training step and only 10% for the test. This can explain the high performance of TF-IDF [7], MSVM [7], TR-Classifier [7] and the Named Entities approach (NE) [10]. However, as an unsupervised method which does not need any kind of training step, the results of LDA-KS and LDA-LS are promising. In fact, dispute culture and economy topics, the result for the rest of topics are comparable and even better some times. For example, for the international news topic, LDA-KS and LDA-LS are better than TF-IDF, MSVM and TR-classifier.

Works	**Culture**	**Economy**	**Intern News**	**Local News**	**Religion**	**Sport**	**Average**
TF-IDF	78.96%	90.03%	81.96%	78.43%	88.60%	96.91%	**86.04%**
MSVM	76.47%	95.50%	79.02%	68.64%	84.83%	89.75%	82.44%
TR-Classifier	81.60%	89.50%	83.77%	84.35%	91.97%	96.66%	88.02%
NE	75.66%	78.14%	90.15%	77.08%	88.26%	95.46%	84.15%
LDA-KS	61.90%	65.88%	84.98%	63.34%	82.75%	88.59%	74.59%
LDA-LS	58.57%	68.41%	84.73%	63.92%	78.06%	90.69%	74.06%

Table 2. ***Comparison with related works*****.**

6. Conclusion

In this paper, we presented a deep study of LDA in the field of Arabic topic identification. In fact, we studied the effect of the stemming process on topic identification by using Arabic stemmers (Khoja and Light Stemmers). Besides, we studied in depth the parameters of LDA. As result, we showed that the choice of parameters influence the performance of LDA and the best result are obtained by $\alpha = 0.5$. Moreover, LDA depends on the stemming algorithms. Based on our evaluation, Light Stemmer is the best stemmer for the topic identification. Thus, based on the best choice of parameters and the stemming algorithm, the result of LDA is very promising in the field of topic identification. For further studies, we will use LDA for topic segmentation to realize a complete topic analysis of Arabic documents.

6. References

[1] A. Kelaiaia and H.F. Merouani. “Clustering with Probabilistic Topic Models on Arabic Texts”. In *Modeling Approaches and Algorithms for Advanced Computer Applications, Springer,* 65-74, 2013.

[2] A.A.Y. Yassine, and K. Amrouche. "Réseaux bayésiens jumelés et noyau de Fisher pondéré pour la classification de documents XML.", ARIMA Journal, Special issue CARI’12, 17:141-154, 2014.

[3] D.M. Blei, A.Y. Ng and M.I. Jordan, "Latent dirichlet allocation". The Journal of machine Learning research, 3, 993-1022, 2003.

[4] G. Heinrich. "Parameter estimation for text analysis". *University of Leipzig, Tech. Rep*, 2008.

[5] K. Taghva, R. Elkhoury and J. Coombs, "Arabic stemming without a root dictionary". International conference on Information Technology, 1:52-57, 2005.

[6] L. Larkey, L. Ballesteros and M. Connell, *Light stemming for Arabic information retrieval*. Arabic Computational Morphology, book chapter, Springer, 2007.

[7] M. Abbas, K. Smaïli and D. Berkani. "Evaluation of Topic Identification Methods on Arabic Corpora". *JDIM*, *9*(5), 185-192, 2011.

[8] M. Steyvers and T. Griffiths. *Probabilistic topic models*. Handbook of latent semantic analysis, 427(7):424-440, 2007.

[9] M. Zrigui, R. Ayadi, M. Mars and M. Maraoui, "Arabic text classification framework based on latent dirichlet allocation". *CIT.* Journal of Computing and Information Technology, 20(2): 125-140, 2012.

[10] R. Koulali and A. Meziane, "Feature Selection for Arabic Topic Detection Using Named Entities". In Proceeding of CITALA, Oujda, Morocco, pp. 243-246, 2014.

[11] S. Khoja and R. Garside, "Stemming Arabic text". Computer science, UK, 1999.

[12] T. Buckwalter, "Buckwalter Arabic morphological analyser version 2.0". LDC2004L02, ISBN 1-58563-324-0, 2004.

[13] V. Vapnik, "The natural of statitical learning theory". Springer, New York, 1995.

[14] Y. Lu, M. Qiaozhu and Z. ChengXiang. "Investigating task performance of probabilistic topic models: an empirical study of PLSA and LDA."*Information Retrieval* 14(2):178-203, 2011.