How the ExpLSA Approach Impacts the Document classification Tasks

Nicolas Béchet
bechet@lirmm.fr

Mathieu Roche
mroche@lirmm.fr

Jacques Chauché
chauche@lirmm.fr
LIRMM - UMR 5506, CNRS
Univ. Montpellier 2, 34392 Montpellier Cedex 5 - France

Abstract

Latent Semantic Analysis (LSA) is a statistical method which can be used to classify texts. This paper proposes a sentence expansion method (ExpLSA) to improve document classification tasks. We propose to study the impact of ExpLSA on the size and on the type of corpora.

1 Introduction

Text classification based on document containing few words can be a difficult task. In this paper, we focus on document expansion to improve categorization task.
After the text expansion, we can use different vector space model of documents, as the Salton vectors described in [13]. In a Salton vector space model, rows are relating to the words and columns are the various contexts (document, section, sentence, etc). Every cells of matrix represent the number of words in the contexts. Two semantically close words (or contexts) are represented by close vectors. The proximity measure is generally defined by the cosine between the two vectors.
Our approach relies on other vector space models to represent a text to categorize. In this paper we deal with the application of the Latent Semantic Analysis (LSA) [8] vector space representation.

The LSA method is based on the fact that words which appear in the same context are semantically close. Corpus is firstly represented by a Salton matrix. Then, a Singular Value Decomposition (SVD) is applied. This one is described in [8]. Then, the resulting matrix of SVD is a reduced matrix of the original, by keeping k singular values. Experiments presented in the section 5 were performed with a factor k stated at 100.
The punctuations and irrelevant words in a semantical point of view (stop words: “and”, “a”, “with”, etc.) are not taken into account.
LSA gives many advantages like the notion of independence of the language used in the corpus. No language or domain knowledge are needed. Rehder *et al.* [12] show that a poor context (less than 60 words) returns disappointing results with LSA. In addition, the efficiency of LSA is weak with a proximity of the vocabulary used. For example, a very accurate classification of texts based on very close domains can be difficult with LSA.

Our aim is to improve the LSA performances by using an approach called ExpLSA (Expansion of contexts with LSA) in a textual classification context. ExpLSA consists in expanding a corpus for finally apply a ’classic’ LSA approach. With ExpLSA, we want to satisfy the lack of feature quantity of documents (or contexts) by adding knowledge. By applying ExpLSA, we add informations (words) to have a more relevant context in order to improve the LSA reduction. The LSA and ExpLSA methods are the first step of the textual classification. Then, we apply usual classification algorithms on the LSA and ExpLSA representations of the documents: kNN, NaiveBayes, and SVM algorithms (described in section 5.1). We consider the size of corpora, the size of the contexts (documents), and the type of corpora to evaluate our approach. We experiment ExpLSA with two corpora writing in French (opinion and news corpora).
The paper is organized as follows: In the section 2, we present a state-of-the-art by adding syntactic knowledge to LSA. Then we present our ExpLSA method (section 3). Section 4 deals with the use of ExpLSA with different corpora and tasks. Finally, the experimental protocol and results will be presented in section 5.

2 Adding Syntactic Knowledge to LSA: The state-of-the-art

Landauer *et al.* [9] present the problem of the lack of syntactic knowledge with LSA method. The authors compare their methods to a human evaluation. They propose to human experts to evaluate essays of 250 words on the human heart writing by students. A semantic space have been built from 27 papers about human heart learned by LSA. Experiments show good results for the LSA method comparing to the human expertise. Bad results was the consequence of a small paucity of syntactic knowledge in the approach that has been used. Thus, the work below shows how this knowledge can be added to LSA.
The first approach of [16] uses the Brill tagger [3] to assign a Part-Of-Speech (POS) tag to every word. The tags are attached to each word with a (“_”). So LSA can consider each word/tag combination as a single term. Results of similarity computing with such method stay disappointing. The second approach of [16] is characterized by the use of a syntactic analysis in order to split text before applying the Latent Semantic Analysis. This approach is called Structured LSA (SLSA). A syntactic analysis of sentences based on different elements (subject, verb, and object) is firstly made. Then, similarity scores (obtained by a cosine computing) between the vectors of the three matrices obtained by LSA are evaluated. The average of the similarities is finally computed. This method gave satisfactory results compared to the “traditional LSA”.
The approach described in [7] proposes a model called SELSA. Instead of generating a matrix of co-occurrences word/document, a matrix where each row contains all the combinations of words_tags, and a column represents a document. The label “prefix” informs about the syntactic class of the word neighborhood. The principle of SELSA is that the sense of a word is given by the syntactic neighborhood. This approach is rather similar to the use of the Brill tagger presented in [16]. But SELSA extends and generalizes this work. A word with a syntactic context specified by its adjacent words is seen as a knowledge representation unit. The evaluation shows that SELSA makes less errors than LSA but these errors are more harmful.

The ExpLSA approach presented in this paper is based on a different context. We propose to use the regularity of some syntactic relations in order to expand the context (documents) as described in the following section.
The use of lexical and semantical resources to expand the contexts is currently used in Information Retrieval (IR) for indexing or expanding queries. The approaches use generic lexical knowledge [15], [10] by adding terms semantically close to the original terms. Other approaches use domain-knowledge to improve the quality of categorization or clustering task. For example, in the method presented in [6] the vectors representing each document are based on the concepts of ontologies. But like the query expansion task, our approach does not use domain-knowledge.

3 Our ExpLSA approach

The ExpLSA approach proposes to expand a corpus by expanding the sentences. It is based on a syntactical method which completes words of the corpus with words which are semantically close. ExpLSA is described in [2]. The following sections summarize our approach applied to a document classification task.

3.1 The use of syntactic parser

We firstly use the Sygfran parser [4] to extract syntactic Verb-Object relations. For instance, we extract the syntactic relation (in French) *verb: nécessiter (to need), Object: professionnels (professionals)* from the French sentence “*L'accompagnement nécessite des professionnels (the follow-up needs professionals)*”.
When all the syntactic relations are extracted, the corpus is lemmatized by the Sygmart system [4].

3.2 The objects gathering in function of verbs proximity

We evaluate the semantic proximity between verbs. To compute this proximity, we use the Asium measure [5]. This measure considers two verbs as close if they have a significant number of mutual features (objects). For example in the figure 1, the couple *requérir - nécessiter (to require - to need)* has mutual features *professionnel (professional)* and *connaissance (knowledge)*. The Asium measure is described in [5].
We keep the closest semantically couples of verbs given by the Asium measure. Thus, we gather the objects of the closest semantically verbs. We have two possibilities to expand the contexts with the objects gathering. The first method consists in completing corpus with mutual words of the two verbs (connaissance (knowledge) and professionnel (professional) in the example of the figure 1). This method is called *intersection*. The second method is to consider the mutual and the complementary objects of the two verbs as [5] (connaissance (knowledge), professionnel (professional), perptuit (perpetuity), and temps (time) in the example of the figure 1). This method is called *complementary*.
We complete the initial corpus by adding to each word the

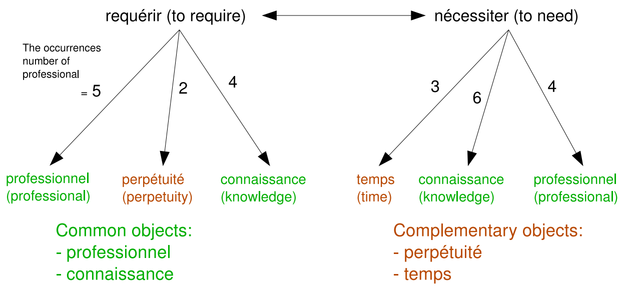


Figure 1. The gathering methods

mutual features given by the Asium measure. We give below an example of expansion of contexts:
The French sentence:
- *L'accompagnement nécessite des professionnels.*
becomes the lemmatized sentence:
- *L'accompagnement nécessiter de le professionnel.*
and becomes with the intersection gathering method:
- *L'accompagnement nécessiter de le (professionnel connaissance).*
and finally becomes with the complementary gathering method:
- *L'accompagnement nécessiter de le (professionnel connaissance perpétuité temps).*
We call ”*ExpLSA int*“ and ”*ExpLSA comp*“ the intersection and the complementary approaches.
The final step of ExpLSA is to apply LSA on the expanded corpus.
ExpLSA relevance is depending of the textual data used as document and corpora size, and type of studied corpora. This point will be described in the next section.

4 Using ExpLSA for Different Textual Data

4.1 Size of documents

Context size is an important aspect for the text classification tasks (in our case, the context is a document).
The training set resulting of small documents is very poor (because there are few features) to obtain a correct classification. A solution to improve the categorization task is to expand the training set by adding knowledge as the studies of *Zelikovitz* [18], [17].
The approach of [18] deals with a combination of labeled training data (of short string) and a second corpus of longer documents (unlabeled) to assist the classification task. Our approach is based on a different context because it does not need labeled dataset. The second approach proposes to add in the original Salton matrix, the test data to improve the quality of the training set. With the application of SVD[1] algorithm, this attractive approach gives good results. This second approach does not need external knowledge as our approach.
Actually, ExpLSA proposes to use the corpus resources to expand contexts. Thus, we can solve the lack of features problem.

4.2 Size of corpora

In this section, we focus on the impact of ExpLSA applied on different size of corpora. The use of ExpLSA with large corpora (more than one million of words) allows the extraction of a lot of syntactical relations. Then, ExpLSA adds a large amount of textual data with several irrelevant expansions. Then, we propose to select terms to add by a new parameter called $nbMin$. $nbMin$ is the minimum occurrence number of an object in Verb-Object syntactic relations used for the expansion. For instance, we consider an object *consul* and $nbMin$ at five. If *consul* appears only four times in the syntactic relations which have been extracted, this word can not be used in the expansion process.

4.3 Corpus and task

The type of the studied text impacts the classification task. The nature of corpus has consequences on the number of classes. For instance, a corpus of news contains a significant number of classes which can be more difficult to categorize. Also, the nature of classes could be semantically close, as political and society topics, for example. By opposition, other corpora as opinion corpora are more suitable for a syntactical relation extraction. The homogeneity of topics in these corpora allows to obtain very relevant relations and can improve the quality of corpus expansion. However, opinion corpora are more difficult to categorize because they require an accurate classification (polarity classification: Positive/Negative opinion) based on a same domain. We will experiment these two types of corpora (news and opinion corpora) in section 5.

5 Experiments

To discuss the quality of the results given by our approach, we describe the experimental protocol in the following section.

5.1 Experimental protocol

In our experiments, we use two corpora written in French:

[1] The Latent Semantic Indexing (LSI) which is LSA applied to indexation tasks was performed

(1) The first is a corpus of news extracted from Yahoo site (http://fr.news.yahoo.com/). It contains 2828 news (5.3 MB) categorized in eleven classes: France, economy, unusual, health, world, politics, culture, science, people, technology, and sport.

(2) The second is an opinion corpus of the DEFT challenge (http://deft07.limsi.fr) with movies, books, comics, CD reviews. It contains 2074 reviews (4.5 MB) categorized in three opinion classes: agree, disagree, and neutral.

These experiments aim at comparing LSA and ExpLSA by performing an automatic classification of documents in a supervised machine learning context. We experiment three algorithms sumarized below. They are precisely described in [14] and [1]. Thus, the vectors used by these algorithms (Input) are given by LSA and ExpLSA approaches.

- **kNN (The k nearest neighbor)**.
  The kNN algorithm specifies the class of a new document by selecting the majority class of the k nearest documents from the learning data. The similarity measure between two vectors representing a document is the cosine. For each classification of a new document, it is necessary to evaluate this new document with all the vectors of the learning data. Therefore, this algorithm is time comsuming.

- **The NaiveBayes**.
  For a NaiveBayes approach, a class is determined as follow. Let C be a group of classes and a specific instance (set of attributes A). A NaiveBayes classification value c is defined by:

  $$c = argmaxP(c_j)_{c_j \in C} \prod_{a_i \in A} P(a_i|c_j) \text{ with}$$
  $$P(a_i|c_j) = \frac{P(c_j|a_i) \times P(a_i)}{P(c_j)} \text{ where:}$$

  - $P(a_i)$ is the probability that the a_i hypothesis was verified independently of other data c_j.
  - $P(a_i|c_j)$ is the probability to observe a_i data for a verification of c_j hypothesis.

  This algorithm is particularly fast with a good compromise speed-quality.

- **The SVM (Support Vector Machines)**.
  The SVM algorithm consists in finding a separator hyperplane between classes. We use in our experiments the SMO algorithm which allows to use the SVM algorithm for a multi-classes problem [11]. This algorithm is more slow than the NaiveBayes model, but it generally obtains better results.

To evaluate the performances of these algorithms, we apply a n-fold cross-validation method (10-fold CV in our experiments). We use the Weka[2] application to perform our experiments by keeping default features of algorithms that we used. This approach considers, alternately, data as a test and have a learning data. The test data is evaluated by using the recall and precision measures:

$$\text{precision}_i = \frac{\text{Number of items correctly assigned to the class } i}{\text{Number of items assigned to the class } i} \quad (1)$$

$$\text{recall}_i = \frac{\text{Number of items correctly assigned to the class } i}{\text{Number of items of the class } i} \quad (2)$$

It is generally important to determine a compromise between recall and precision. We use the F-measure for each class i:

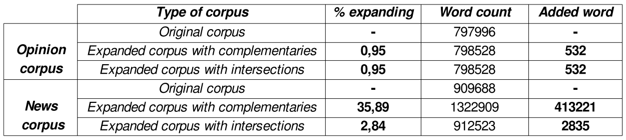
$$fscore_i(\beta) = \frac{(\beta^2 + 1) \times p_i \times r_i}{\beta^2 \times p_i + r_i} \quad (3)$$
$$r_i = \text{recall}_i \text{ and } p_i = \text{precision}_i$$

$$Fscore(\beta) = \frac{\sum_{i=1}^{k} fscore_i(\beta)}{k} \quad (4)$$
$$k = \text{ the number of classes}$$

The factor β of the formula (3) is used to provide the variations between recall and precision. In order to consider identically the recall and the precision, we generally state the value of β at 1. In our experiments, we apply this value ($\beta = 1$).

Note that the F-measure is the macro-average of the F-measure for each class. In further experiments, we plan to use the micro-average of the F-measure.

5.2 Results with small corpora



	Type of corpus	% expanding	Word count	Added word
Opinion corpus	Original corpus	-	797996	-
	Expanded corpus with complementaries	0,95	798528	532
	Expanded corpus with intersections	0,95	798528	532
News corpus	Original corpus	-	909688	-
	Expanded corpus with complementaries	35,89	1322909	413221
	Expanded corpus with intersections	2,84	912523	2835

Table 1. The expanding rate for the two short corpora

We firstly propose to compare LSA and ExpLSA on small corpora (approximatively 5 MB) in order to measure the impact of our expansion approaches. The table 1 compares the expanding rate for the two corpora. The opinion corpus obtains the same expanding ratio for both methods of ExpLSA (intersection and complementary). Thus, verbs that are used for the expansion have the same mutual objects (i.e., this corpus is very homogeneous). The corpus of

[2] $http : //www.cs.waikato.ac.nz/\tilde{~}ml/$

news has an important expansion of contexts with the complementary method compared to the intersection method. Table 2 shows that the F-measure scores are degraded with the corpus of news. The opinion corpus gives better results[3] with all algorithms based on ExpLSA.

	Corpus :	Opinion	News
algorithm	method	F-measure	F-measure
kNN	LSA	48,5%	**59,4%**
	ExpLSA comp	**48,9%**	54,5%
	ExpLSA int	**48,9%**	57,4%
Naives Bayes	LSA	49,8%	**65,5%**
	ExpLSA comp	**52,1%**	55,7%
	ExpLSA int	**52,1%**	64,8%
SVM	LSA	43,0%	**66,7%**
	ExpLSA comp	**43,8%**	61,6%
	ExpLSA int	**43,8%**	65,2%

Table 2. F-measure scores comparing the LSA and both ExpLSA approaches for the news and opinion corpora

We propose now to compare the expansion quality by considering the size of corpora. We have built three sizes of corpora: short, medium, and large. Table 3 shows the number of words in function of the document sizes for both corpora. These sizes were experimentally obtained in order to reach uniform repartition for each size of documents.

Table 4 shows results obtained for the opinion corpus and the corpus of news in function of the size of the documents. The improvement of the results of ExpLSA is confirmed for all the sizes of the documents (except for the kNN algorithm).

	Document size (number of words)		
Corpus	short	medium	long
Opinion	< 300	between 300 and 450	> 450
News	< 250	between 250 and 500	> 500

Table 3. Number of words in function of the document size

For the news corpus, results are improved for the kNN and SVM algorithms with ExpLSA for medium and large documents using the intersection method. Overall, the intersection approach gives better results than the complementary ExpLSA method.

		Short		Medium		Large	
		Opinion	News	Opinion	News	Opinion	News
kNN	LSA	**53,5%**	**50,0%**	51,7%	55,7%	**45,8%**	57,1%
	ExpLSA comp	48,0%	46,5%	**52,1%**	49,4%	41,6%	52,8%
	ExpLSA int	48,0%	48,1%	**52,1%**	**56,1%**	41,6%	**58,5%**
Naive Bayes	LSA	53,0%	61,1%	55,8%	**62,9%**	50,2%	**64,1%**
	ExpLSA comp	**54,4%**	48,1%	**57,4%**	56,3%	**50,4%**	58,9%
	ExpLSA int	**54,4%**	**63,6%**	**57,4%**	61,7%	**50,4%**	63,6%
SVM	LSA	50,4%	**58,6%**	51,6%	61,3%	39,4%	62,0%
	ExpLSA comp	**54,3%**	51,7%	**53,9%**	58,2%	**40,6%**	59,3%
	ExpLSA int	**54,3%**	57,3%	**53,9%**	**61,6%**	**40,6%**	**62,2%**

Table 4. LSA and ExpLSA F-measures (with different sizes of corpora)

5.3 Results with an important corpus

Now we propose to evaluate our approach with a biggest news corpus. This one contains 8,948 news (16.5 MB). The better values for the $NbOcc$ parameter presented in section 4, were experimentally obtained: 8 for the complementary method and 5 for the intersection method. The expanded corpora have respectively an expanding ratio of 29.76% and 26.19%.

The intersection method gives better results than the complementary one like the results developped in the section 5.2.

		F-measure		
algorithm	method	short	medium	large
kNN	LSA	**66,2%**	**68,0%**	68,7%
	ExpLSA comp	64,2%	64,0%	67,8%
	ExpLSA int	62,1%	65,0%	**69,1%**
NaiveBayes	LSA	**66,4%**	**66,2%**	64,5%
	ExpLSA comp	60,5%	63,1%	63,3%
	ExpLSA int	62,0%	64,2%	**65,3%**
SVM	LSA	**65,6%**	**69,0%**	**65,1%**
	ExpLSA comp	61,4%	67,0%	63,6%
	ExpLSA int	62,7%	67,9%	64,8%

Table 5. F-measure scores comparing the LSA and both ExpLSA approaches for the large corpus of news

The table 5 presents results with this news corpus. We consider small, medium, and large documents. These results suggest that ExpLSA which uses the intersection method gives better results with the kNN and NaiveBayes algorithms for the large documents. In the other cases, our approach decreases the results by comparison with LSA.

6 Discussions

Now, we propose to discuss these results for the different sizes and types of corpora. Mono-thematic corpora contain

[3] The small variance in results obtained with SVM and particularly kNN may be due to the statistical error of cross-validation and can also explain these results.

more homogeneous syntactic relations (*i.e.* better quality of the relations) as explained in section 4.3. Then, this characteristic could explain that ExpLSA gives better results for the opinion classification task, and the decreasing results of the corpus of news, which are multi-thematic (table 2).

ExpLSA consists in expanding the corpus before the application of LSA. This expansion with relevant words can also occasionally give some noise. But this one will be reduced by the LSA approach on large documents. By opposition, the LSA application with small documents can increase the noise caused by the small quantity of features. This situation could explain the results obtained by the kNN and SVM algorithms when we focus on the classification of small documents (table 4). The NaiveBayes algorithm gives opposed results. The efficiency of NaiveBayes algorithm to categorize the partial data could explain this result.

The large corpus contains an important amount of features. For these large corpora, we can suppose (table 5) that expanded contexts are not necessary (due to the amount of features used for the machine learning process).

7 Conclusion and perspectives

We propose in this paper to improve text classification by applying a context expansion.

LSA is a statistical method which can be used to gather contexts (document classification). We propose an approach called ExpLSA providing a context expansion in order to categorize documents. We use syntactical resources to perform these context expansions.

We proposed to use three algorithms to evaluate the results: kNN, NaiveBayes, and SVM algorithms.

ExpLSA approach improves results of LSA for: (i) Small corpora (better results with opinion corpus and medium and long documents of news corpus), (ii) Difficult tasks of classification (for instance, opinion classification)

In a future work, we firstly propose to adapt a hybrid approach combining the ExpLSA method and the 'classic' LSA approach. We plan to validate our expansion by quering the web using search engines. We will also add other sets of syntactic knowledge to improve LSA. Finally, we will use a semantical vector representation of data by using the Sygmart system [4].

References

[1] K. Aas and L. Eikvil. Text categorisation: A survey. Technical report, Norwegian Computing Center, June 1999. Norsk Regnesentral (Norwegian Computing Center, NR), 1999.

[2] N. Béchet, M. Roche, and J. Chauché. ExpLSA: An approach based on syntactic knowledge in order to improve LSA for a conceptual classification task. In *RCS volume (posters proceedings), CICLing 2008*, pages 213–224, 2008.

[3] E. Brill. Some advances in transformation-based part of speech tagging. In *AAAI, Vol. 1*, pages 722–727, 1994.

[4] J. Chauché. Un outil multidimensionnel de l'analyse du discours. In *Proceedings of Coling, Standford University, California*, pages 11–15, 1984.

[5] D. Faure and C. Nedellec. Knowledge acquisition of predicate argument structures from technical texts using machine learning: The system ASIUM. In *Proceedings of the 11th European Workshop, Knowledge Acquisition, Modelling and Management, number 1937 in LNAI*, pages 329–334, 1999.

[6] A. Hotho, A. Maedche, and S. Staab. Ontology-based text document clustering, 2002.

[7] D. Kanejiya, A. Kumar, and S. Prasad. Automatic evaluation of students' answers using syntactically enhanced lsa. In *Proceedings of the Human Language Technology Conference (HLT-NAACL 2003) Workshop on Building Educational Applications using NLP*, 2003.

[8] T. Landauer and S. Dumais. A solution to plato's problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge. *Psychological Review*, 104(2):211–240, 1997.

[9] T. Landauer, D. Laham, B. Rehder, and M. E. Schreiner. How well can passage meaning be derived without using word order? A comparison of latent semantic analysis and humans. In *Proceedings of the 19th annual meeting of the Cognitive Science Society*, pages 412–417, 1997.

[10] D. Moldovan and R. Mihalcea. Improving the search on the internet by using wordnet and lexical operators. In *IEEE Internet Computing 4(1)*, pages 34–43, 2000.

[11] J. Platt. Fast training of support vector machines using sequential minimal optimization. In *B. Scholkopf, C. J. C. Burges, and A. J. Smola, editors, Advances in Kernel Methods – Support Vector Learning, pages 185-208. MIT Press, Cambridge, MA*, 1999.

[12] B. Rehder, M. Schreiner, M. Wolfe, D. Laham, T. Landauer, and W. Kintsch. Using latent semantic analysis to assess knowledge: Some technical considerations. In *Discourse Processes*, volume 25, pages 337–354, 1998.

[13] G. Salton. *Automatic Information Organization and Retrieval.* McGraw Hill Text, 1968.

[14] F. Sebastiani. Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1):1–47, 2002.

[15] E. Voorhees. Query expansion using lexical-semantic relations. In *ACM SIGIR94, Dublin*, 1994.

[16] P. Wiemer-Hastings and I. Zipitria. Rules for syntax, vectors for semantics. In *Proceedings of the Twenty-third Annual Conference of the Cognitive Science Society*, 2001.

[17] S. Zelikovitz. Transductive lsi for short text classification problems. In *Proceedings of the 17th International FLAIRS Conference*, 2004.

[18] S. Zelikovitz and H. Hirsh. Improving short text classification using unlabeled background knowledge. In P. Langley, editor, *Proceedings of ICML-00, 17th International Conference on Machine Learning*, pages 1183–1190, Stanford, US, 2000. Morgan Kaufmann Publishers, San Francisco, US.