ExpLSA: An Approach Based on Syntactic Knowledge in Order to Improve LSA for a Conceptual Classification Task

Nicolas Béchet, Jacques Chauché, Mathieu Roche

ExpLSA: An Approach Based on Syntactic Knowledge in Order to Improve LSA for a Conceptual Classification Task

Nicolas Béchet and Mathieu Roche and Jacques Chauché

LIRMM - UMR 5506 - CNRS, Univ. Montpellier 2,
34392 Montpellier Cedex 5 - France

Abstract. Latent Semantic Analysis (LSA) is nowadays used in various thematic like cognitive models, educational applications but also in classification. We propose in this paper to study different methods of proximity of terms based on LSA. We improve this semantic analysis with additional semantic information using Tree-tagger or a syntactic analysis to expand the studied corpus. We finally apply LSA on the new expanded corpus.

1 Introduction

Classification's domain has many research fields like conceptual classification. This one consists in gathering terms in concepts defined by an expert. For example, *exhaust pipe, windshield wiper*, and *rearview mirror* terms can be associated to the *automobile* concept. Then, these terms are classified by semantic proximity with different algorithms like k nearest neighbor (KNN) or k means. The corpora have different types as the language, the syntax, the domain (biology, medicine, etc) using a specialized semantic, etc. Then these complex textual data require a specific process.

In this paper, we describe the first step of a conceptual classification, the study of proximity of the terms. First, we use the Latent Semantic Analysis (LSA) method evolved by [1][1]. LSA is a statistic method applied to high dimension corpora to gather terms (conceptual classification) or contexts (textual classification). After the latent semantic analysis application on the corpus, a semantic space associating each word to a vector is returned. Then, the proximity of the two words can be obtained by measuring the cosine between two vectors. Our aim is to improve the performance of the LSA method by an approach named *ExpLSA*.

The *ExpLSA* approach (context **Exp**ansion with **LSA**) consists in expanding the corpus before the application of a "traditional" latent semantic analysis. This context expansion uses semantic knowledge obtained by syntax, what allows to use *ExpLSA* as well specialized corpus as not. Actually, it is not necessary to use training corpus, so it is not necessary to know the general domain of the corpus.

[1] http://www.msci.memphis.edu/∼wiemerhp/trg/lsa-followup.html

In this paper, we use a Human Resources corpus of PerformanSe[2] company in French. It uses a specialized vocabulary. This corpus was submitted to a human expertise which validates our experiments.

We propose in the next section to explain theoretical characteristics of LSA method and its limits. Section 3 proposes a state-of-the-art in the field of the syntactic knowledge used with LSA. Then we present the different steps of our method (section 4). We also describe (section 5) the experimental protocol applied. Finally we will present results obtained.

2 LSA

The LSA method is based on the fact that words which appear in the same context are semantically close. Corpus is represented by a matrix. Lines are relating to words and columns are the several contexts (document, section, sentence, etc). Every cells of matrix represent the number of words in the contexts. Two semantically close words are represented by close vectors. The proximity measure is generally defined by the cosine between the two vectors.

LSA is based on the Singular Value Decomposition (SVD) theory. $A = [a_{ij}]$ where a_{ij} is the frequency of the word i in the context j, breaking down in a product of three matrices USV^T. U and V are orthogonal matrices and S a diagonal matrix.

Let us S_k where $k < r$ the matrix built by removing of S the $r - k$ columns which have the smallest singularly values. We take U_k and V_k, matrices obtained by removing corresponding columns of U and V matrices. Then, the $U_k S_k V_k^T$ can be considered like a compressed version of the original matrix A. Experiments presented in the section 5 were made with a factor k equal to 200.

Before performing a singularly value decomposition, a first step of normalization of the original matrix A is applied. This normalization consists in applying a logarithm and an entropy measure on the matrix A. This transformation allows to refer the weight of the words on their contexts. In a similar way to the work of [2], this normalization can also refer on the tf×idf method, well-known approach in the field of the Information Retrieval (IR).

Also let us specify that we do not consider punctuations and some words not relevant in a semantical point of view like: “and”, “a”, “with”, etc. (stop words).

LSA gives many advantages like notions of independence of the language of the corpus. It needs no language or domain knowledge. However, LSA has limits. Firstly, we consider size of chosen contexts. [3] showed during its experiments that if contexts have less than 60 words, the results can be disappointing. In addition, the efficiency of LSA is weak with a proximity of the vocabulary used. For example, a very precise classification of texts based on very close domains can be difficult with LSA.

[2] http://www.performanse.fr/

3 State-of-the-art on the addition of syntax to LSA

[4] presents the problem of the lack of syntactic knowledge with LSA method. They 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. Tests performed give 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 used. Thus, the work below demonstrates how these knowledge can be added to LSA.

The first approach of [5] uses the Brill tagger [6] to assign a part-of-speech tag to every word. The tags are attached to each word with an underscore. So LSA can consider each word/tag combination as a single term. Results of similarity calculation with such method stay disappointing. The second approach of [5] is characterized by the use of a syntactic analysis in order to segment 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 "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 line contains all the combinations of words_tags, and a column represents a document. The label "prefix" informs about the syntactic type of the word neighborhood. The principle of SELSA is that the sense of a word is given by the syntactic neighborhood from which it results. This approach is rather similar to the use of the Brill tagger presented in [5]. But SELSA extends and generalizes this work. A word with a syntactic context specified by its adjacent words is seen as a unit knowledge representation. The evaluation shows that SELSA makes less errors than LSA but these errors are more harmful.

The *ExpLSA* approach presented in this paper is placed in a different context. In fact, in our studies, the contexts are represented by sentences. These ones have a reduced size which tends to give low results with LSA [3]. In our approach, we propose to use the regularity of some syntactic relations in order to expand the context as described in the next section.

4 Our approach: *ExpLSA*

The final aim consists in automatically gathering terms extracted by systems adapted to French corpora such as Acabit [8], Syntex [9], Exit [10]. In our case, we propose to gather nominal terms extracted with Exit from the Human Resources corpus (written in French). The extracted terms with this system are phrases respecting the syntactical patterns (noun-prep-noun, adj-noun, noun-adj, etc). In addition, Exit is based on a statistical method to rank terms extracted. It uses an iterative approach to build complex terms.

The first step of the conceptual classification can be done by *ExpLSA*. Its principle is described in the following sections.

4.1 General principle of ExpLSA

Our approach makes an expansion of lemmatized sentences based on a syntactic method. It comes out a richer context by completing words of the corpus by words considered semantically similar. We have for example the sentence (in French): *"Vos interlocuteurs seront donc bien inspirés..."*. We transform it firstly in a lemmatized sentence: *"Votre interlocuteur être donc bien inspiré..."*. Then, it will be expanded by other terms. This sentence becomes *"Votre (interlocuteur collaborateur) être donc bien inspiré..."*. This set of terms semantically close are selected with a measure presented in the next section.

4.2 The syntactic analysis to evaluate the semantic proximity

In order to improve initial corpus, we propose firstly to apply a syntactic analysis with the SYGMART French parser [12]. This one gives the existing syntactic relations in each sentence. In our approach, we only used the Verb-Object relations (Verb_DO, Verb_Preposition_Complement) of the corpus.

When all Verb-Object relations are extracted, we use a measure to evaluate the semantic proximity of words, the Asium measure [13]. This one proposes to evaluate verbs considered as close if they have a significant number of common features. The principle of the approach is similar to the method presented in [11].

We consider verbs, the associated prepositions and features after a syntactic parsing. The Asium measure consists in computing a similarity measure between verbs.

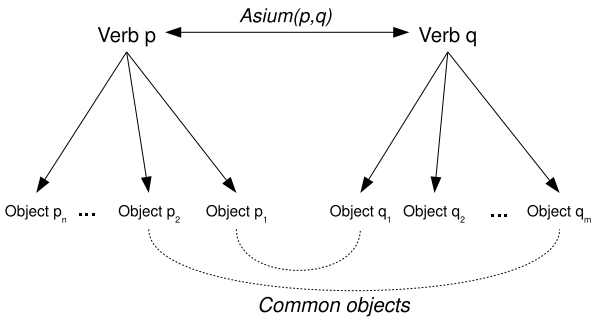


Fig. 1. The Asium measure between p and q verbs.

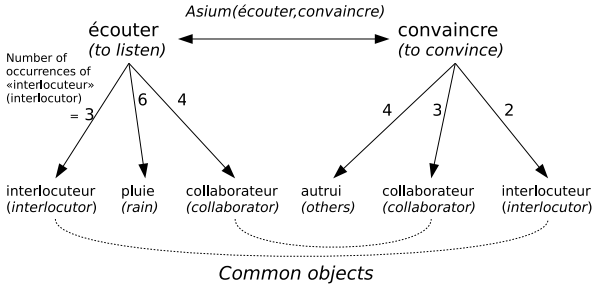
We consider the p and q verbs with their respective p_1,...,p_n and q_1,...,q_n objects illustrated in the figure 1. $NbOC_p(q_i)$ is the occurrences number of q_i

object of q and p verbs (common objects). $NbO(q_i)$ is the occurrences number of q_i objects of q verb. Then, the Asium measure is:

$$Asium(p,q) = \frac{log_{Asium}(\sum NbOC_q(p_i)) + log_{Asium}(\sum NbOC_p(q_i))}{log_{Asium}(\sum NbO(p_i)) + log_{Asium}(\sum NbO(q_i))}$$

Where $log_{Asium}(x)$ equal to :
- for $x = 0$, $log_{Asium}(x) = 0$
- else $log_{Asium}(x) = log(x) + 1$

The example of the figure 2 gives an application of the Asium measure for the *écouter (to listen)* and *convaincre (to convince)* verbs. Next, we consider different similarity threshold values of the Asium measure. Over this one, verbs are considered as close by the Asium measure. The expansion method is described in the next section.



$$Asium(\text{écouter,convaincre}) = \frac{\log_{Asium}(4+3)+\log_{Asium}(3+2)}{\log_{Asium}(3+6+4)+\log_{Asium}(4+3+2)} = 0{,}87$$

Fig. 2. The Asium measure between the verbs *écouter (to listen)* and *convaincre (to convince)*

4.3 The steps of ExpLSA

After the description of the Asium measure, we propose to clarify the different steps of *ExpLSA* to expand the contexts.

The first step of the *ExpLSA* approach identifies the different terms extracted by Exit. Then a term is represented like a single word (for instance, the *attitude profondément participative* term becomes *noun234* which is the 234th term of a list extracted by Exit).

After the extraction of the syntactic Verb-Object relations by using a syntactic analysis, the next step of our approach is to study the semantic proximity between verbs using the Asium measure. We deduced for example that verbs *écouter (to listen)* and *convaincre (to convince)* are semantically close by using the Asium measure because they have the common objects *interlocuteur (interlocutor) and collaborateur (collaborator)* (See figure 2).

The next step proposes to gather the objects of the closest semantically verbs. Then, we consider two gathering methods. The first method consists in completing corpus with common words of the two verbs (*interlocuteur* and *collaborateur* in the example of the figure 2). The second method is to consider the common and the complementary objects of the two verbs to expand corpus like the work of Faure and Nedellec [17] (*entourage, autrui, pluie, collaborateur* in the example of the figure 2).

Then we propose to expand initial corpus by catching to each word, the other words judged close. Then, our initial French sentence:

– *Votre* ***interlocuteur*** *être donc bien inspiré...*

becomes with the first gathering method:

– *Votre* ***(interlocuteur collaborateur)*** *être donc bien inspiré...*

and becomes with the second gathering method:

– *Votre* ***(autrui pluie interlocuteur collaborateur)*** *être donc bien inspiré...*

Latent Semantic Analysis can be applied with the expanded corpus.

A list of general words with a poor semantic content are not selected to expand the context (for instance, general words like "chose" (*thing*), "personne" (*person*), etc). This list is manually made.

The evaluation presented in the next section compares results obtained with *ExpLSA* with results of the experts who had manually associated relevant terms to the concepts.

5 Experiments

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

5.1 Experimental protocol

In our experiments, we use a Human Resources corpus manually evaluated. This expertise gives a conceptual classification of all terms extracted by Exit. The concepts are defined by the expert. For instance, with our corpus, the expert defined "Relationnel" (*relational*) concept. The *confrontation ouverte (open adversarial)*, *contact superficiel (superficial contact)*, and *entourage compréhensif (understanding circle)* terms are instances of this concept. Our aim is to evaluate similarities obtained with the following methods:

– *M1:* LSA
– *M2:* the intersections method of *ExpLSA*
– *M3:* the complementaries method of *ExpLSA*
– *M4:* LSA + Tree-tagger

The LSA + Tree-tagger method consists in using the Tree-tagger [18]. We use a part-of-speech tagger like the approach of [5] developed in the section 3.

To compare these methods, we select the most representative concepts, *i.e.* concepts gathering a minimum of 200 distinct terms given by the expertise. We obtain four concepts ($C1$, $C2$, $C3$, and $C4$). Twenty terms of each concept which have the most number of occurrences are selected. Then we measure the similarity (with the cosine) of the couples of terms of the concepts 2-2 (terms of $C1$ and $C2$ concepts, $C1$ and $C3$ concepts, etc). For example, the similarity measure is computed between the terms of the $C1$ concept with the terms of the $C1 + C2$, $C1 + C3$, and $C1 + C4$ concepts. A couple of terms is called relevant if they are an instance of the same concept.

The experiments consist in ranking these couples of terms based on the similarity measure using the M1, M2, M3 (section 5.2), and M4 (section 5.3) methods. Then we compute the precision for the n first couples of terms. Precision gives the proportion of relevant couples returned by the method. The recall has not judged adapted[3].

5.2 The *ExpLSA* methods comparison

We propose a first experiment comparing LSA with both methods of *ExpLSA* used to expand the corpus: The complementaries and the intersections approaches. The table 1 gives average precision of the 100 first couples of terms (for each couple of concepts evaluated) with a threshold of the Asium measure at 0.6. This value of threshold makes a large expansion of the corpus. This table indicates better results for the *ExpLSA* with the intersections method (M2 method) except for the "activity-environment" couple of concepts. But these improvements are mitigate; we have a large majority of cases where the LSA method is not improved because an Asium threshold to 0.6 makes a large expansion with a lot of noise added. We conclude experimentally that the better threshold value of *ExpLSA* is 0.8. It is a good compromise between quantity and quality of expanded knowledge. The table 2 is based on an Asium threshold to 0.8. This one gives better results for the intersections approach (M2 method) in comparison with LSA. The figure 3 confirms the weak results for the complementaries approach. We only conserve M2 method in the next section.

[3] Our approach allows to obtain relevant couples at the beginning of the list. Nevertheless, in a case of few couples are considered ($n < 100$), recall is naturally week and not adapted to evaluate the performance of our approach.

		The n first terms								
Pairs of concepts	*Method*	20	30	40	50	60	70	80	90	100
Activity	M1	**0,551**	0,543	0,536	0,527	0,520	0,518	0,520	0,524	0,521
Behaviour	M2	0,539	**0,553**	**0,545**	**0,541**	**0,545**	**0,546**	**0,537**	**0,534**	**0,528**
	M3	0,533	0,519	0,509	0,521	0,524	0,520	0,516	0,512	0,511
Activity	M1	0,520	**0,528**	0,515	0,518	0,514	0,513	0,513	0,512	0,514
Environment	M2	**0,548**	0,519	**0,524**	0,519	0,513	0,512	0,511	0,508	0,510
	M3	0,538	0,526	0,521	**0,524**	**0,524**	**0,524**	**0,521**	**0,518**	**0,520**
Activity	M1	**0,580**	**0,567**	0,559	0,546	**0,546**	0,536	**0,540**	**0,545**	**0,543**
Relational	M2	0,566	0,563	**0,561**	**0,548**	0,545	**0,539**	**0,540**	0,541	0,538
	M3	0,549	0,538	0,529	0,526	0,528	0,519	0,520	0,519	0,522
Behaviour	M1	**0,586**	**0,566**	**0,556**	**0,546**	**0,542**	**0,538**	**0,533**	**0,533**	**0,531**
Environment	M2	0,526	0,522	0,513	0,519	0,521	0,518	0,513	0,513	0,512
	M3	0,539	0,522	0,523	0,525	0,521	0,524	0,523	0,525	0,523
Behaviour	M1	**0,555**	**0,548**	**0,539**	0,528	0,523	**0,523**	**0,524**	**0,525**	**0,526**
Relational	M2	0,540	0,530	0,527	**0,531**	**0,524**	**0,523**	0,523	0,517	0,519
	M3	0,523	0,516	0,509	0,502	0,507	0,506	0,512	0,516	0,519
Environment	M1	0,559	0,550	0,548	**0,560**	**0,544**	**0,539**	**0,539**	**0,540**	**0,537**
Relational	M2	**0,560**	**0,558**	**0,551**	0,538	0,532	0,529	0,526	0,526	0,526
	M3	0,538	0,506	0,506	0,519	0,520	0,521	0,516	0,518	0,515

Table 1. Precision of the 100 first couples. Method M1 = LSA, Method M2 = ExpLSA with the intersections method, M3 = ExpLSA with the complementaries method. Asium threshold = 0.6.

		The n first terms								
Pairs of concepts	*Method*	20	30	40	50	60	70	80	90	100
Activity	M1	0,551	0,543	0,536	0,527	0,520	0,518	0,520	0,524	0,521
Behaviour	M2	**0,558**	**0,564**	**0,558**	**0,551**	**0,542**	**0,541**	**0,534**	**0,531**	**0,530**
	M3	0,514	0,522	0,524	0,529	0,529	0,518	0,523	0,520	0,516
Activity	M1	0,520	0,528	0,515	0,518	0,514	0,513	0,513	0,512	0,514
Environment	M2	**0,531**	**0,536**	**0,537**	**0,535**	**0,534**	**0,530**	**0,525**	**0,523**	**0,518**
	M3	0,519	0,506	0,506	0,504	0,501	0,500	0,496	0,500	0,500
Activity	M1	**0,580**	0,567	**0,559**	0,546	**0,546**	0,536	**0,540**	**0,545**	**0,543**
Relational	M2	0,576	**0,572**	0,555	**0,549**	0,545	**0,546**	0,539	0,534	0,534
	M3	0,546	0,549	0,548	0,536	0,530	0,523	0,527	0,522	0,516
Behaviour	M1	0,586	0,566	0,556	0,546	0,542	0,538	0,533	0,533	0,531
Environment	M2	**0,600**	**0,583**	**0,573**	**0,572**	**0,560**	**0,551**	**0,545**	**0,547**	**0,543**
	M3	0,526	0,520	0,511	0,517	0,513	0,514	0,511	0,513	0,511
Behaviour	M1	**0,555**	**0,548**	**0,539**	0,528	0,523	0,523	0,524	0,525	0,526
Relational	M2	0,545	0,528	0,536	**0,539**	**0,537**	**0,534**	**0,531**	**0,531**	**0,527**
	M3	0,509	0,515	0,522	0,522	0,519	0,518	0,514	0,511	0,508
Environment	M1	0,559	0,550	0,548	**0,560**	0,544	0,539	0,539	**0,540**	0,537
Relational	M2	**0,579**	**0,563**	**0,559**	0,555	**0,548**	**0,546**	**0,543**	0,539	**0,539**
	M3	0,488	0,486	0,493	0,498	0,497	0,498	0,496	0,494	0,497

Table 2. Precision of the 100 first couples. Method M1 = LSA, Method M2 = ExpLSA with the intersections method, M3 = ExpLSA with the complementaries method. Asium threshold = 0.8.

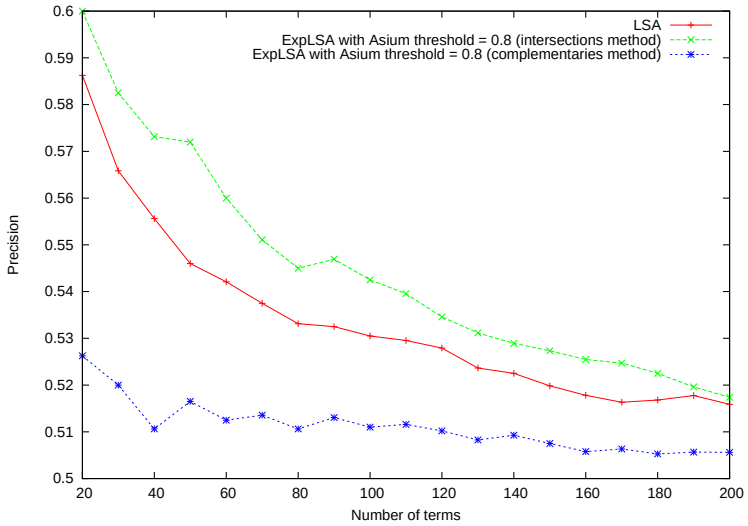


Fig. 3. Precision of the 200 first couples comparing the *ExpLSA* methods between concepts behavior and environment

5.3 ExpLSA compared to the LSA + Tree-tagger approach

In this section we use the syntactic knowledge in two ways: *ExpLSA* and LSA + Tree-tagger. LSA + Tree-tagger adds grammatical knowledge (part-of-speech category) to the corpus by assign a part-of-speech tag to the words. LSA is applied with this tagged corpus. We see in the table 3 and in the figure 4 that the precision of *ExpLSA* is better for the first couples. This is the researched result because that means the first couples provided at the expert are relevant. These encouraging results[4] of *ExpLSA* and LSA + Tree-tagger methods allow to consider a hybrid approach by combine both.

5.4 ExpLSA applied to Text Categorization

In recent works, we proposed to evaluate the *ExpLSA* method for a task of Text Categorization (TC). The aim of these works was to improve TC tasks: SVM (Support Vector Machines), k-NN (k nearest neighbor), Naive Bayes, and Decision Tree (C4.5) algorithms [19] using LSA and *ExpLSA* representations. We use a French news corpus which contains 2828 articles (5,3 MB) provided by yahoo's French web site (http://fr.news.yahoo.com/). In these experiments we calculate the average F-measure by performing a ten-fold cross-validation. We obtained good results of *ExpLSA* representation with the application of SVM,

[4] These results are a little different comparing to results presented in [16] because we proposed a different experimental protocol by considering every words of concept and not only the twenty most common words.

		The n first terms								
Pairs of concepts	*Method*	20	30	40	50	60	70	80	90	100
Activity	M1	0,551	0,543	0,536	0,527	0,520	0,518	0,520	0,524	0,521
Behaviour	M2	**0,558**	**0,564**	**0,558**	**0,551**	**0,542**	**0,541**	0,534	0,531	0,530
	M4	0,551	0,535	0,542	0,547	0,541	0,540	**0,538**	**0,535**	**0,533**
Activity	M1	0,520	0,528	0,515	0,518	0,514	0,513	0,513	0,512	0,514
Environment	M2	**0,531**	**0,536**	**0,537**	**0,535**	**0,534**	**0,530**	**0,525**	**0,523**	0,518
	M4	0,505	0,508	0,505	0,505	0,509	0,513	0,518	0,522	**0,525**
Activity	M1	**0,580**	0,567	0,559	0,546	0,546	0,536	0,540	0,545	0,543
Relational	M2	0,576	**0,572**	0,555	0,549	0,545	0,546	0,539	0,534	0,534
	M4	0,559	**0,572**	**0,572**	**0,567**	**0,560**	**0,559**	**0,553**	**0,551**	**0,550**
Behaviour	M1	0,586	0,566	0,556	0,546	0,542	0,538	0,533	0,533	0,531
Environment	M2	**0,600**	**0,583**	**0,573**	**0,572**	**0,560**	**0,551**	0,545	0,547	0,543
	M4	0,563	0,561	0,561	0,565	0,553	0,549	**0,552**	**0,550**	**0,550**
Behaviour	M1	0,555	**0,548**	0,539	0,528	0,523	0,523	0,524	0,525	0,526
Relational	M2	0,545	0,528	0,536	0,539	**0,537**	**0,534**	**0,531**	**0,531**	**0,527**
	M4	**0,574**	0,545	**0,542**	**0,542**	0,536	0,530	0,529	0,525	0,523
Environment	M1	0,559	0,550	0,548	**0,560**	0,544	0,539	0,539	**0,540**	0,537
Relational	M2	**0,579**	0,563	**0,559**	0,555	**0,548**	**0,546**	**0,543**	0,539	**0,539**
	M4	0,573	**0,565**	0,554	0,550	0,544	0,536	0,535	0,531	0,530

Table 3. Precision of the 100 first couples comparing the LSA (M1), the *ExpLSA* with Asium threshold = 0.8 (M2) and LSA + Tree-tagger approach (M4)

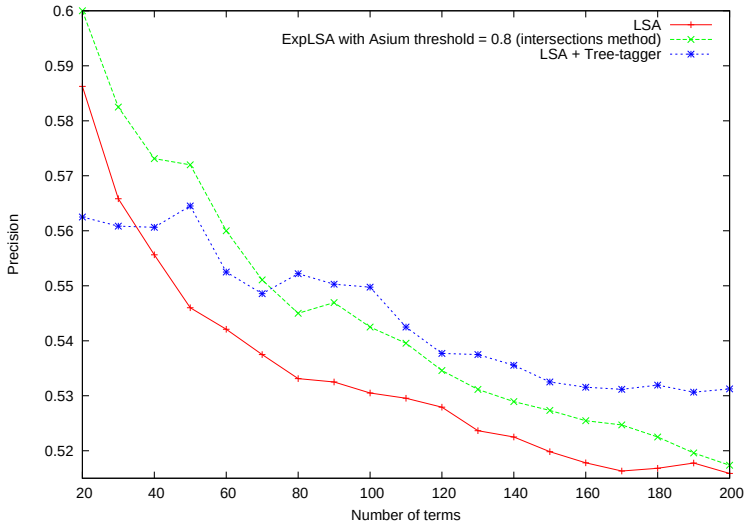


Fig. 4. Precision of the 200 first couples comparing the ExpLSA and LSA + Tree-tagger methods between concepts behavior and environment

k-NN, and C4.5 on medium and long articles of the corpus. In order to improve the classification task of short documents, we will improve quality of expansion using a validation based on the web knowledge.

6 Conclusion and discussion

LSA is a statistical method which can be used to gather terms to build a conceptual classification. However, this method gives medium results in this domain. We can explain these results by the absence of the syntactic knowledge. Quality of results can also be influenced by the size of contexts used. The LSA method gives better results with an important size of contexts.

Then, we propose to improve the LSA performances with small contexts (sentences) by an approach called *ExpLSA*. This one consists in applying a context expansion with LSA by expand the original corpus. We use syntactical resources to make these contexts expansion.

We presented two experiments to compare the *ExpLSA* approach. First, we experiment the *ExpLSA* approach based on the complementaries approach. This expansion is not relevant because this method returns a lot of noise. We consider the Asium threshold value to 0.8, which is the best value experimentally found. The intersections method of *ExpLSA* gives good results. We compare it with another method in the second experiment, the LSA + Tree-tagger approach.

In a future work, we propose firstly to adapt a hybrid approach combining the *ExpLSA* (with the intersections approach) and the LSA + Tree-tagger methods. We envisage to add other set of syntactic knowledge to improve LSA. Finally, we would want to apply *ExpLSA* in order to use other classification tasks (for instance, the text classification task).

References

1. Landauer, T., Dumais, S.: A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction and Representation of Knowledge. Psychological Review, Vol. 104 (1997) 211–240
2. Turney P.D.: Mining the Web for Synonyms: PMI–IR versus LSA on TOEFL. Proceedings of ECML'01, Lecture Notes in Computer Science (2001) 491–502
3. Rehder B., Schreiner M., Wolfe M., Laham D., Landauer T., Kintsch W.: Using Latent Semantic Analysis to assess knowledge: Some technical considerations. Discourse Processes, Vol. 25 (1998) 337–354
4. Landauer T., Laham D., Rehder B., Schreiner M. E.: How well can passage meaning be derived without using word order? A comparison of Latent Semantic Analysis and humans. Proceedings of the 19th annual meeting of the Cognitive Science Society (1997)
5. Wiemer-Hastings P., Zipitria I.: Rules for syntax, vectors for semantics. Proceedings of the Twenty-third Annual Conference of the Cognitive Science Society (2001)
6. Brill E.: Some Advances in Transformation-Based Part of Speech Tagging AAAI, Vol. 1 (1994) 722–727

7. Kanejiya D., Kumar A., Prasad S.: Automatic Evaluation of Students' Answers using Syntactically Enhanced LSA. Proceedings of the Human Language Technology Conference (HLT-NAACL 2003) Workshop on Building Educational Applications using NLP (2003)
8. Daille, B.: Study and Implementation of Combined Techniques for Automatic Extraction of Terminology. In: Resnik, P.; Klavans, J. (eds.): The Balancing Act: Combining Symbolic and Statistical Approaches to Language, MIT Press, Cambridge, MA, USA, (1996) 49–66.
9. Bourigault D., Fabre C.: Approche linguistique pour l'analyse syntaxique de corpus. Cahiers de Grammaires, Vol. 25 (2000) 131–151
10. Roche M., Heitz T., O. Matte-Tailliez O., Kodratoff Y.: EXIT: Un système itératif pour l'extraction de la terminologie du domaine à partir de corpus spécialisés. Proceedings of JADT'04, Vol. 2 (2004) 946–956
11. Bourigault D.: UPERY : un outil d'analyse distributionnelle étendue pour la construction d'ontologies à partir de corpus. Actes de TALN, Nancy (2002) 75–84
12. Chauché J.: Un outil multidimensionnel de l'analyse du discours. Proceedings of Coling, Standford University, California (1984) 11–15
13. Faure D.: Conception de méthode d'apprentissage symbolique et automatique pour l'acquisition de cadres de sous-catégorisation de verbes et de connaissances sémantiques à partir de textes : le système ASIUM. Phd, Université de Paris Sud (2000)
14. Ferri C., Flach P., Hernandez-Orallo J.: Learning decision trees using the area under the ROC curve Proceedings of ICML'02 (2002) 139–146
15. Mary J.: Étude de l'Apprentissage Actif : Application à la conduite d'expériences. Phd, Université Paris Sud (2005)
16. Béchet N., Roche M., Chauché J.: *ExpLSA* : utilisation d'informations syntaxico-sémantiques associées à LSA pour améliorer les méthodes de classification conceptuelle. Proceedings of EGC'08.
17. Faure D., Nedellec C.: Knowledge acquisition of predicate argument structures from technical texts using machine learning: The system ASIUM. Proceedings of the 11th European Workshop, Knowledge Acquisition, Modelling and Management, number 1937 in LNAI (1999) 329-334
18. Schmid H. Improvements in part-of-speech tagging with an application to German. Technical report, Universitat Stuttgart. Institut fur maschinelle Sprachverarbeitung. (Revised version of a paper presented at EACL SIGDAT, Dublin 1995).
19. Y. Yang An Evaluation of Statistical Approaches to Text Categorization. Information Retrieval, number 1-2, volume 1 (1999) 69–90