Improving LSA by expanding the contexts

Nicolas Béchet, Mathieu Roche, and Jacques Chauché

LIRMM - UMR 5506, CNRS, Univ. Montpellier 2,
{nicolas.bechet,mroche,chauche}@lirmm.fr

Abstract. Latent Semantic Analysis is used in many research fields with several applications of classifications. We propose to improve LSA with additional semantic information found with syntactic knowledge.

1 Introduction

In this paper, we use the Latent Semantic Analysis (LSA) approach [1]. LSA is a statistic method applied to high dimension corpora to gather terms (conceptual classification) or contexts (textual classification). The proximity between terms or contexts provided by LSA represents a first step of classification tasks. Our approach, *ExpLSA* (**Exp**ansion of Contexts with **LSA**), consists in expanding the context. This expansion is based on semantic information found with syntactic knowledge. In this paper, we use a Human Resources corpus of PerformanSe company[1] (3784 KB) in French.

For the LSA method, the words that appear in the same context are semantically close. A corpus is represented by a matrix. The lines represent the words and the columns are the different contexts (document, section, sentence, etc). Each cell in the matrix stands for the number of words in a context. Two semantically close words have vectors close (lines of the matrix). The proximity measure is generally defined by the cosine between the vectors.

LSA is based on the Singular Value Decomposition (SVD) theory. $A = [a_{ij}]$ where a_{ij} is the frequency of 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 singular 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 an approximation of the version of the original matrix A. Experiments presented in section 4 are applied with a factor $k = 50$, a low value that is more suitable for small corpora.

Before the singular value decomposition, a first step of normalization of original matrix A is applied. This normalization consists in computing a logarithm and an entropy computation on matrix A. This process allows to estimate the weight of words in their contexts. This normalization can also be based on the tf×idf method, a well-known approach in the field of the Information Retrieval

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

(IR). Let us note that the punctuations and stop words (like "and", "a", "with", etc) are not taken into account to compute LSA.

LSA has many advantages like the languages and domains independence. Nevertheless, an important limit of LSA is based on the size of contexts. Rehder *et al.* showed that the contexts with less than 60 words obtain disappointing results [2].

2 State-of-the-art based on the addition of syntax to LSA

The approaches described in [3, 5] take into account the syntactical knowledge. The approach of [3] uses the Brill's tagger [4] to assign a part-of-speech tag to every word. With this method, LSA considers each word/tag as a single term. This method gives disappointing results. The second approach described in [3] is based on the syntactic analysis in order to segment a text. A syntactic analysis of sentences on three elements (subject, verb, and object) is firstly done. Then, the similarity (cosine) is calculated separately for the three elements (three LSA matrices). The average of the similarities is finally computed. This method gave satisfactory results compared to "traditional LSA".

The approach described in [5] proposes a model called SELSA. It uses part-of-speech tag and a "prefix" label. This one informs about the syntactic type of the words' neighborhood. This approach is close to [3] but SELSA extends this work by generalizing it. A word with a syntactic context specified by its adjacent words is seen as a unit representation of knowledge. SELSA makes less errors than LSA but these errors are more harmful.

In our work, the contexts are represented by sentences. They have a small size giving low results with the LSA method [2]. We propose to use the regularity of some syntactic relations in order to expand the context.

3 Our approach: *ExpLSA*

The final aim consists in automatically gathering terms (conceptual classification) extracted by a system like SYNTEX [6] or EXIT [7]. We propose to gather nominal terms extracted with EXIT from the Human Resources corpus. LSA and *ExpLSA* are the first stage for the conceptual classification task.

The first step of the *ExpLSA* approach identifies the different terms extracted by EXIT. This process consists in representing each term by only one word (for instance, the french term *attitude profondément participative* becomes *noun234* which is the 234th term of a list extracted by EXIT).

After this process, SYGMART parser [8] is applied. This one gives the syntactic relations of each sentence. In our approach, we study Verb-Object relations (Verb_Object, Verb_Preposition_Complement) of our corpus.

The next step of our approach studies semantic proximity between verbs using the Asium measure [9]. With this measure, the verbs are semantically close when they have a lot of common objects. In the next section (section 4), several Asium thresholds based on the similarity values between the verbs will be presented. When the values of the Asium threshold are high, the verbs are close.

The next step proposes to gather common objects (words) of close verbs. Words of the corpus are replaced with all the words of its same group built at the precedent step. For example, our initial lemmatized sentence in French: "*Votre* ***interlocuteur*** *être donc bien inspiré...*" becomes finally: "*Votre* ***(interlocuteur collaborateur)*** *être donc bien inspiré...*". LSA can be applied with the expanded corpus. Very general nouns are not selected to expand context (as "chose" (*thing*), "personne" (*person*), etc).

4 Experiments

In these experiments, we compare similarities given by LSA/*ExpLSA* with a manual expertise. The experts have manually associated terms to 17 concepts. For instance, with our corpus, the expert defined "Relationnel" (*relational*) concept where the term *contact superficiel (superficial contact)* is an instance.

The five most representative terms (the most frequent) which are instances of concepts are used in our experiments. The similarity (cosine) for all representative terms of two concepts is computed. We can verify that the most close pairs of terms given by LSA and *ExpLSA* are instances of the same concept (i.e. these pairs are called *relevant*). In order to compare the results of similarity returned by LSA and *ExpLSA*[2], we propose to calculate the ranking sum of relevant pairs of terms. Then, in our experiments, with the lower sum, we obtain the better results. This evaluation measure is an approach based on ROC curves (Receiver Operating Characteristic) and Area Under these Curves [10]. This feature is mostly used to compare ranking functions [11]. The Area Under ROC Curves is equivalent to calculate the sum of the relevant elements [12].

Pairs of concepts	**LSA**	**ExpLSA**	
		0.6 *threshold*	0.9 *threshold*
Influence / Indépendance *(Impact / Independency)*	496	530	532
Relationnel / Environnement *(Relational / Environment)*	420	468	492
Relationnel / Rôle *(Relational / Role)*	384	**359**	**355**
Rôle / Comportement-Attitude *(Role / Behaviour)*	344	389	**325**
Stress / Indépendance *(Anxiety / Independency)*	481	**392**	**401**
Stress / Vous-même *(Anxiety / Yourself)*	494	**442**	**446**
Vous-même / Comportement-Attitude *(Yourself / Behaviour)*	422	423	**407**

Table 1. LSA and *ExpLSA* with different Asium thresholds (0.6 and 0.9).

Table 1 shows the evaluations obtained on randomly selected concepts for LSA, *ExpLSA* with 0.6 threshold, and *ExpLSA* with 0.9 threshold. We compare the results with a corpus using an Asium threshold of 0.9 versus a large (but less relevant) expansion corpus using a threshold to 0.6. The ranking sums of relevant pairs of terms are compared with LSA. Our *ExpLSA* approach with 0.6 Asium threshold improves the LSA results only 3 times on 7. But when we use a 0.9 threshold, *ExpLSA* improves results 5 times on 7. Thus we achieve better

[2] only the sentences with the instances of concepts are used to compute LSA and *ExpLSA*.

quality results with a 0.9 threshold. However, there are two cases where *ExpLSA* performed badly. They could be studied in a future work.

5 Conclusion and discussion

LSA is a method applied to large corpora. Actually, this analysis is less efficient with small corpora. We study in this paper a corpus to build a conceptual classification. We complete a corpus with our *ExpLSA* approach using syntactic knowledge. Our approach does not improve results for all experiments. However, the results obtained are hopeful. Our experiments have been performed on a small number of concepts. We intend to perform *ExpLSA* with every concepts combination. Moreover, we will estimate more precisely the most appropriate Asium threshold with new experiments.

Acknowledgment: We thank Yves Kodratoff (LRI, France) and Serge Baquedano (PerformanSe company) for their expertise on the Human Resources corpus.

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. 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
3. Wiemer-Hastings P., Zipitria I.: Rules for syntax, vectors for semantics. Proceedings of the Annual Conference of the Cognitive Science Society. (2001)
4. Brill E.: Some Advances in Transformation-Based Part of Speech Tagging. AAAI, Vol. 1. (1994) 722–727
5. Kanejiya D., Kumar A., Prasad S.: Automatic Evaluation of Students' Answers using Syntactically Enhanced LSA. Proceedings of the Workshop on Building Educational Applications using NLP (HLT-NAACL conference). (2003)
6. Bourigault D., Fabre C.: Approche linguistique pour l'analyse syntaxique de corpus. Cahiers de Grammaires, Vol. 25. (2000) 131–151
7. 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
8. Chauché J.: Un outil multidimensionnel de l'analyse du discours. Proceedings of Coling, Standford University, California. (1984) 11–15
9. 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, Univ. de Paris 11. (2000)
10. Ferri C., Flach P., Hernandez-Orallo J.: Learning decision trees using the area under the ROC curve. Proceedings of ICML'02. (2002) 139–146
11. Roche M., Kodratoff Y.: Pruning Terminology Extracted from a Specialized Corpus for CV Ontology Acquisition. Proceedings of onToContent'06 workshop - OTM'06, Springer Verlag, LNCS. (2006) 1107-1116
12. Mary J.: Étude de l'Apprentissage Actif : Application à la conduite d'expériences. Phd, Univ. Paris 11. (2005)