Detection of opinions and facts. A cognitive approach

Yann Vigile Hoareau†
Human and Artificial Cognition Lab.
University of Paris 8 – 2 rue de la Liberté.
93526 St Denis Cedex 02 -FRANCE
hoareau@lutin-userlab.fr

Adil El-Ghali
Human and Artificial Cognition Lab.
University of Paris 8 – 2 rue de la Liberté.
93526 St Denis Cedex 02 -FRANCE
elghali@lutin-userlab.fr

Charles Tijus
Human and Artificial Cognition Lab.
University of Paris 8 – 2 rue de la Liberté.
93526 St Denis Cedex 02 -FRANCE
tijus@lutin-userlab.fr

Abstract

A model of episodic memory is derived to propose algorithms of text categorization with semantic space models. Performances of two algorithms named *Target vector* and *Sub-target vector* are contrasted using textual material of the text-mining context 'DEFT09'. The experience reported here have been realized on the english corpus which is composed of articles of the economic newspaper "*The Financial Times*". The aim of the task was to categorize texts in function of the factuality or subjectivity they expressed. Results confirm (i) that the episodic memory metaphor provides a convenient framework to propose efficient algorithm for text categorization, and (ii) that *Sub-target vector algorithm* outperforms the *Target vector algorithm*.

Keywords

Random Indexing, episodic memory, text-mining, categorization.

1. Introduction

Since its early introduction, the model that is now named Latent Semantic Analysis [14] has been proposed as a method of matrix reduction and vectorial representation of information for indexing textual documents. The model was known as Latent Semantic Indexing [3] at that time.

Originally only concerned by indexing tasks, LSA has been extended to the area of human memory simulation. Researchers in cognitive psychology got interested in it and then proposed it as a plausible model of human behavior in different tasks such as synonymy test [14] and problem solving [17]. The most famous application in cognitive psychology is the coupled CI-LSA model of text comprehension [12], which combines the previous "Construction-Integration" model of reading [11] with LSA as model of semantic memory. Whereas research involving LSA has been split in two main fields with the text-mining on the one hand and cognitive psychology on the other hand, our paper deals with both of those fields. Discussions of MINERVA 2 model of human episodic memory [6][7] allow proposing an operative algorithm for texts categorization.

LSA has been known to perform in synonymy test and other equivalent thematic classification tasks [14]. The model has been recently successfully applied on opinion judgment task [1]. There are very important differences between thematic classification, and opinion judgment classification. Firstly, thematic classification is directly connected to the distributional hypothesis, which states that "words that appear in similar contexts have similar meanings". Here is the reason why LSA is able to find words that *share the same thematic, ie "appear in equivalent contexts"*. Secondly, in opinion judgment classification, different thematics could possibly belong to the same category of opinion. For example, I have a good opinion of different movies, which do not deal with the same topic. If I write texts in which I give my opinion of each movie, those texts will be influenced by the topic of the movie for a part, as well as by my motivation to exhibit how and why I loved them for another part. In consequence, the basic application of the distributional hypothesis cannot account for judgment opinion task.

In this paper, we will explore two lines of investigation. In the first line, we will propose the paradigmatic breakthrough that has been realized to find a solution to the limitation of the basic application of the distributional hypothesis. This breakthrough consists in switching from the semantic memory research field to the episodic memory metaphor to drive the similarity comparison stage. The episodic memory metaphor has been tested with LSA [8]. The second line that will be developed in this paper will consist in testing the episodic memory metaphor with an alternative method of Words Vectors construction, named Random Indexing.

2. Abstractive versus non-abstractive models of memory

In the debate within cognitive psychology about the distinction between "abstractive" versus "non-abstractive" models of memory [18][21], LSA has been proposed as belonging to the abstractive family [2]. This proposition is congruent with the affirmation by Landauer, Foltz and Laham that "the representations of passages that LSA forms can be interpreted as abstractions of "episodes", sometimes of episodes of purely verbal content such as philosophical arguments, and sometimes episodes from real or imagined life coded into verbal descriptions" [15: 15]. Tiberghien considers that "it would be more precise and theoretically more adequate, to consider that all the models are 'abstractive' but, for some of them this abstractive process

International Conference RANLP 2009 - Borovets, Bulgaria, pages 150–154

happens during encoding and for some others it happens during retrieval" [21: 145]. Because the abstractive process occurs during encoding, LSA and other Word Vector models are categorized as belonging to the abstractive model family.

A model like MINERVA 2 or other Multiple-Trace models are considered as "non-abstractive" because the abstractive process occurs during retrieval. According to MINERVA 2, memory consists of events or episodes that are represented and stored as vectors. The activation value of each coordinate stores features of episodes. Each vector corresponds to an episode in the system's life. Retrieval consists of a two stage calculation. First, a similarity calculation is carried out between the probe-vector and all the episode-vectors in memory (see ***Eq 1***). Episodes that are most similar will be affected by a higher level of activation than episodes that are least similar. Second, a calculation is made to compare the level of activation of each feature and this corresponds to the "echo" phenomena of memory. The "echo" calculation produces a new vector that inherits the features of the most activated vectors, even those parts that did not actually exist in the probes. The "echo" has two components: intensity which is denoted *I* (see ***Eq 2***), and content which corresponds to the sum of the content of all traces in memory, weighted by their activation level (see ***Eq 3***). "Echo" constitutes the process of abstraction that Rousset (2000) qualified as "re-creation.

$$S_i = \sum_{j=1}^{N} \frac{P_j T_{i,j}}{N_i}$$

Eq 1 Similarity of a trace i, where P_j is the value of feature j in the probe, and $T_{i,j}$ the value of feature j in trace I

$$\mathbf{I} = \sum_{i=1}^{M} A_i, \text{ where } A_i = S_i^3$$

Eq 2 Intensity of the « echo »

$$C_j = \sum_{i=1}^{M} A_i T_{i,j}$$

Eq 3 The content of the « echo »

3. The episodic memory metaphor in opinion judgment classification task

LSA has been successfully applied in tasks of text classification with texts expressing subjective opinion in the DEFT07 contest [1]. The Multiple-Trace approach has been proposed to account for semantic space performance when modifying factors like generality/specificity of episodes that compose the space [8]. Two predictions of MINERVA 2 model has been tested and confirmed. First, two methods of semantic space construction are compared.

In one method, different categories of episodes are blended in the same *global* semantic space. In the other method, each semantic space is built from a single category of episodes. These spaces are named *specific*. For each method of semantic space construction (*global vs specific*), two experimental conditions are compared. In the first condition, the number of episodes corresponding to each category is equalized. In the other condition, the number of episodes corresponding to each category is not controlled.

For the global space condition, correlation analysis showed that the relationship between relative amount of episodes and F-score was more important than the relationship between absolute amount of episode and F-score (r = .96, α > .001 *versus* r = .74, α > .05). For the specific space condition, the relationship between F-score and relative amount of data was almost the same as the relationship between F-score and absolute amount of data (r = .84, α > .01 *versus* r = .87, α > .01).

As predicted by MINERVA 2, modifying the relative amount of episodes or the absolute amount of episodes has an almost equivalent effect on performance for specialized spaces, whereas modifying relative amount of episodes has a more important effect on performance than modifying absolute amount of episode for general spaces.

4. The episodic memory metaphor for similarity judgment algorithm

The algorithm used in Deft07 to identify opinion judgment expressed by unknown texts, consisted in creating a target vector for each type of opinion that should be identified. These target vectors are created by the sum of vectors of all documents that belong to a given category of opinion[1]. For example, the target vector that was used to identify "good critics of movies" was a summed vector of all documents known to be a "good critic of movie". In-comings "text-to-be-indexed" were compared to the target vectors of each category of opinion. Then, the text was categorized with the opinion of the target vector to which it was the more similar. The comparison of similarity used the calculation of the cosine of the angle between the vector of the "text-to-be-indexed" and the target vector. The use of cosine calculation makes it possible to compare the very large target-vectors (hundreds of documents) to the very small text-to-be-indexed vector (one document).

The intuition that was underlying the construction of these very large target-vectors was that the classical distributional hypothesis approach has to be derived to perform in opinion judgment task. The idea was to sum vectors of all documents corresponding to a given opinion

[1]In data-mining contests, a classified corpus is given in what is called a *learning stage* to make it possible to implement algorithms that will be used to categorize un-classified documents in the *test stage*.

category to take advantage of the great number of documents to draw a vector that (i) would not correspond to any topic in particular, and in contrast, (ii) would hold information that would correspond to the linguistic way a given opinion is statistically expressed in numbers of texts. Applying the Multiple-Trace approach specifically to the stage of similarity comparison makes it possible to consider a target vector as an episodic memory that should behave like MINERVA 2 model predicts. Indeed, in considering each document as a specific episode, target-vectors become episodic memories, which are constituted of different episodes of the same category of opinion. As described above, the calculus of "echo" of MINERVA 2 predicts that the more a probe is similar to great numbers of episodes, the more the memory system would react by a strong value of "echo". It is neither mathematically nor psychologically wrong to consider that the value of "echo" in MINERVA 2 and the value of the cosine in LSA behave and can be interpreted in the same way. In consequence, MINERVA 2 gives a theoretical basement to our first intuitive method of vector target construction. The large size target vector method functioned pretty well and contributed to rank second in the Deft07.

5. Target-vectors as homogeneous episodic memory

The use of the episodic memory metaphor accounts for the limitation of the basic application of the distributional hypothesis for opinion judgment task. In creating these large target vectors, we are creating episodic memories, which behaviors became understandable with the MINERVA 2 model. Predictions concerning "echo" involve that the episodic memories will be more sensitive to probe episodes that are well represented in the memory and less sensible to probe episodes that are less represented. In other words, target vectors will be more sensible to typical documents and less sensible to non-typical documents. Theories of categorization showed that some items are typical of the category they belong, others are not. The typicality of an item is generally defined as (i) a high similarity with items of a given category and (ii) a low similarity with items of other categories.

Target vectors have been produced with the aim of creating episodic memories, which would hold the statistical linguistic signature of a given category of opinion. "Echo" predicts that target vectors will not identify non-typical documents as well as typical documents. We assume that a homogeneous episodic memory, which holds non-typical documents of a given category will be more sensitive to non-typical documents than a heterogeneous episodic memory, which holds typical and non-typical documents, all blended.

Our hypothesis has been implemented for the DEFT09. The aim of the task 1 was to classify texts that express facts or opinions, respectively corresponding "objective" *versus* "subjective" categories. First, we created a target vector in summing all vectors of all documents for each category. These target vectors had the same properties than those of the Deft07. Second, to be able to identify and regroup typical *versus* non-typical documents, a calculation of similarity is realized between (i) each document that composes the target vector, and (ii) the target vector. Documents that compose the target vector are ordered in function of their similarity with the target vector. Third, documents are regrouped in *n* sub-target vectors in a way that (i) each sub-target vector has the same amount of documents, and (ii) documents of the same degree of similarity with the target vector are regrouped in the same sub-target vector. The number of sub-target vectors for each category is a parameter of the algorithm we developed. Whereas the target vector algorithm has been tested with LSA, we propose to compare the *target-vector algorithm* with the *sub-target vector algorithm* using an alternative method of Word vector named Random Indexing.

6. Random Indexing

Word vectors correspond to a family of models in which LSA is the most known. Several principles are common to all of these models (see [18]):

- They are based on the distributional hypothesis
- They involve a method of counting words in a given unit of context
- They have a statistical method, which abstracts the meaning of concepts from large distributions of words in context
- They use a vectorial representation of word meaning.

As we will see, Random Indexing is not a typical item of its category. In the other models, the list of principles enounced above is also the stages of a semantic space construction. Particularities of the Random Indexing (RI) model are that (i) it does not create co-occurrence matrix (but it is possible if needed) and (ii) it does not need heavy statistical treatments like SVD for LSA. Contrary to the other Word Vector models, RI is not based on statistics but on random projections. The construction of a semantic space with RI is as follows:

- Create a matrix A (d x N), containing *Index vectors*, where d is the number of documents or contexts and N, the number of dimensions ($N >$ 1000) decided by the experimenter. *Index vectors* are sparse and randomly generated. They consist in small numbers +1 and -1 and thousands of 0.
- Create a matrix B (t x N), containing *term vectors*, where t is the number of different terms in the

corpus. Set all vectors with null values to start the semantic space construction.

- Scan each document of the corpus. Each time a term *t* appears in a document *d*, accumulate the randomly generated *d-index vector* to the *t-term vector*.

At the end of the process, *term vectors* that appeared in similar contexts have accumulated similar *index vectors*. There is a training cycle option in the model. When the scan has been computed for all documents, the matrix B is charged for all *term vectors*. Then a matrix A' (*d'* x *N*), with *d' = d* can be computed with the output of *term vectors*. The number of training cycle is a parameter in the model. The training process output is consistent with what has been described for neural network learning. The RI model has performed in TOEFL synonymy test [9][10] as well as in text categorization [18].

7. Experiment

7.1 Method and material

The experiment reported here has been realized the task 1 of the DEFT09 using the english corpus. The purpose of the task 1 was the detection of the subjectivity or objectivity character of a text. As described by the committee, "the reference is established by projecting each section on both the subjective and the objective dimension. For instance, the Letter from the editor, which usually states an opinion, has the type subjective, while the News, describing actual facts, have the type objective"[2]. The english corpus was composed of articles of the economic newspaper "*The Financial Time*". In the learning stage, 60% of the total corpus is given to each team engaged to allow them to implement algorithms that will then be applied on the 40% of uncategorized documents during the test stage. We realized our learning session using the "ten cross-folder" method. Table 1 give a description of the corpus.

Table 1. Description of the corpus of learning and test

	Learning		**Test**	
	Number of documents	Size (Kb)	Number of documents	Size (Kb)
Objective	3440	15840	5245	27996
Subjective	4426	26016		

7.2 Results

Precision and recall performances are reported for the *Target vector algorithm* and the *Sub-target vector algorithm*. Taking account that the value of 1 for recall means that all documents have been categorized in the same category. Hence those scores should be considered as aberrant.

This is the case for two reported results: the *Target vector algorithm*, which have 1 target and the *Sub-target vector algorithm* using 11 targets (indicated by the double slash in Table 2) both have 0.432 for Precision and 1 in Recall. Those results demonstrate that the *Target vector algorithm* was not able to perform in the considered task.

Concerning the *Sub-target vector algorithm,* the systems performs better using 9 sub-targets (Precision of 0.740 and Recall of 0.708 using 1000 dimensions and respectively 0.746 and 0.718 using 2000 dimensions). This result involves that there is an optimum threshold for the number of sub-target vectors. Considering Multiple-Trace approach, this threshold corresponds to the moment where episodic memories or sub-targets are the most homogeneous.

Runs realized changing the specific parameters of Random Indexing as the number of dimensions and the number of training cycles show that the optimum partition realized with the *Sub-target vector algorithm* using 9 sub-targets doest not change significantly (between 0.740 and 0.746 for the Precision and between 0.708 and 0.718 for Recall). Those results show that performance of the *Sub-target vector algorithm* is more dependent of the number of sub-target used and less dependent of the parameters of Random Indexing.

Table 2. Parameters and scores

Parameters			**Score**	
Dimensions	Cycles	Sub-target	Precision	Recall
Target-vector algorithm				
1500	10	1	0.432	1//
Sub-target vector algorithm				
1500	10	3	0.648	0.508
1500	10	5	0.688	0.530
1500	10	7	0.652	0.503
1000	10	9	0.740	0.708
1500	10	9	0.738	0.704
2000	10	9	0.746	0.718
1500	10	11	0.432	1//

8. Conclusions

Target vector algorithm consisted in creating a very large vector composed of each and every documents of a given category as target vector used to identify the category a document belongs to. The proposed theoretical switching from abstractive to non-abstractive model of memory has

[2] DEFT09 website: http://deft09.limsi.fr/index.php?id=1&lang=en

been described and tested to account for the *Target-vector algorithm*. Those large target vectors have been considered as episodic memories and MINERVA 2 has been used as a metaphor to predict and interpret behaviors of such episodic memories. The *Target-vector algorithm,* which has been developped for the DEFT07, has been applied on the DEFT09 corpus. Results reported demonstrate very bad performance.

Computing the *Sub-target algorithm* with different numbers of homogeneous sub-targets was congruent with predictions derived from the "echo" calculation of Minerva 2. Indeed, performance reported for the *Sub-target algorithm* using 9 sub-targets demonstrated that there is an optimal partition of similar episodes in sub-target that upgrades the system's performance.

The work presented here is in the line of researches that study the effect of typicality or the effect of frequency of episodes on the capability of the memory system to recognize or to recall a particular event or item. Future developments of our research should highlight, in a more reliable way, how mathematical description of the human cognitive system could upgrade artificial computing systems, particularly in Natural Language Processing applications.

9. References

[1] M. Ahat, W. Lenhart, H. Baier, Y. V. Hoareau, S. Jhean-Larose, & G. Denhière,(2007) "Le concours DEFT'07 envisagé du point de vue de l'Analyse de la Sémantique Latente (LSA)". In Proceedings of the conference DEFT'07, Grenoble, 2007.

[2] C. Bellissens, P. Thérouane, & G. Denhière, Les modèles vectoriels de la mémoire sémantique : description, validation et perspectives, Le Langage et L'Homme, 34 (2004), 101-122.

[3] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, & R. Harshman, Indexing By Latent Semantic Analysis, Journal of the American Society For Information Science, 41(1990), 391-407.

[4] G. Denhière, B. Lemaire, C. Bellissens, & S. Jhean-Larose, A semantic space for modeling children's semantic memory, In T. K. Landauer, D. McNamara, S. Dennis, W. Kintsch (Eds) The Handbook of Latent Semantic Analysis, Lawrence Erlbaum Associates, Mahwah, 2007.

[5] S. T. Dumais, Improving the retrieval of information from external resources, Behavior Research Methods, Instruments, & Computers, 23 (1991), 229-236.

[6] D. L. Hintzman, MINERVA 2: A simulation of human memory, Behavior Research Methods, Intruments, & Computers, 16 (1984), 96-101.

[7] D. L. Hintzman, Judgments of frequency and recognition memory in a Multiple-Trace Model, Psychological Review, 95 (1988), 528-551.

[8] Y. V. Hoareau, M. Ahat, G. Denhière, S. Jean-Lharose, W. Lenhard & H. Baier & L. Legros, Indexing with LSA : the Multiple Trace approach, (submitted).

[9] P. Kanerva, J. Kristoferson, & A. Holst, Random Indexing of Text Samples for Latent Semantic Analysis, In L.R. Gleitman, and A.K. Josh (Eds.), Proceedings of the 22nd Annual Conference of the Cognitive Science Society, Lawrence Erlbaum Associates, Mahwah, 2000.

[10] J. Karlgren, & M. Sahlgren, From Words to Understanding, In Y. Uesaka, P. Kanerva, & H. Asoh (Eds.) Foundations of Real-World Intelligence, CSLI Publications, Stanford, 2001.

[11] W. Kintsch, The role of knowledge in discourse comprehension : a construction-integration model, Psychological Review, 95 (1988), 163-182

[12] W. Kintsch, Comprehension: a paradigm for cognition, Cambridge University Press, Cambridge, 1998.

[13] T. K. Landauer, LSA as a Theory of Meaning. In T. K. Landauer, D. McNamara, S. Dennis, W. Kintsch (Eds). The Handbook of Latent Semantic Analysis, Lawrence Erlbaum Associates, Mahwah, 2007.

[14] T. K. Landauer, & S. T.Dumais, A solution to Plato's problem: The Latent Semantic Analysis theory of the acquisition, induction, and representation of knowledge, Psychological Review, 104 (1997), 211-240.

[15] T. K. Landauer, P. W. Foltz, & D. Laham, Introduction to Latent Semantic Analysis, Discourse Processes, 25 (1998), 259-284.

[16] K. Lund, C. Burgess. Producing high-dimensional semantic spaces from lexical co-occurrence, Behavior Research Methods, Instrumentation, and Computers, 28 (1996), 203-208.

[17] J.F. Quesada, W. Kintsch, & E. Gomez, A theory of Complex Problem Solving using Latent Semantic Analysis, In W. D. Gray & C. D. Schunn (Eds.) Proceedings of the 24th Annual Conference of the Cognitive Science Society, Lawrence Erlbaum Associates, Mahwah, NJ, 2002.

[18] S. Rousset, Les conceptions "système unique" de la mémoire : aspect théorique, Revue de neuropsychologie, 10 (2000), 30-56.

[19] M. Sahlgren, The Word-Space Model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces. Ph.D. dissertation, Department of Linguistics, Stockholm University, (2006).

[20] M. Sahlgren, & R. Cöster, Using Bag-of-Concepts to Improve the Performance of Support Vector Machines in Text Categorization, Proceedings of the 20th International Conference on Computational Linguistics, Geneva, 2004.

[21] G. Tiberghien, La mémoire oubliée, Mardaga, Liège, 1994