

The episodic memory metaphor in text categorization with Random Indexing

Yann Vigile Hoareau & Adil El Ghali
CHArt – Lutin (Paris 8/EPHE) & INRIA Sophia Antipolis



Lutin Userlab
Cité des sciences et de l'industrie

DefT09, Paris le 22 juin 2009

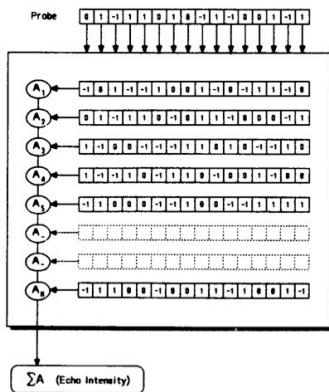


Application of an episodic memory modeling perspective on text categorization using Random Indexing

- From episodes to concepts
 - A famous model of episodic memory: MINERVA 2
 - Description
 - Simulation of the effect of frequency episodes
 - A Word Vector model: Random Indexing
 - Description
 - Text categorization and the application of the distributional hypothesis
- The DEFT09 text-mining contest
- Results
- Perspectives

2

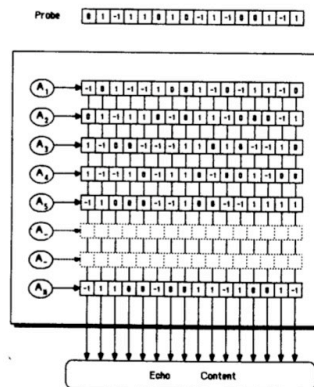
MINERVA 2 from (Hintzman, 1988)



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

3

MINERVA 2 from (Hintzman, 1988)

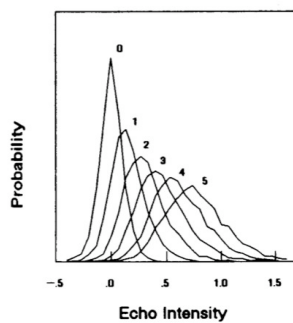


$$I = \sum_{i=1}^M A_i \text{ where } A_i = S_i^3$$

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

4

Effect of frequency of episodes on echo (Hintzman, 1988)



Mean and variance of echo increase with frequency

5

Application of an episodic memory modeling perspective on text categorization using Random Indexing

- From episodes to concepts
 - A famous model of episodic memory: MINERVA 2
 - Description
 - Simulation of the effect of frequency episodes
 - A Word Vector model: Random Indexing
 - Description
 - Text categorization and the application of the distributional hypothesis
- The DEFT09 text-mining contest
- Results
- Perspectives

6

A model of Word Vector : Random Indexing

- The common principles behind Word Vectors:
 - Implementing distributional hypothesis
 - Dealing Large corpora
 - Working on a context window
 - Building a matrix that hold the uses of words in function of their contexts
 - Reducing the matrix
 - Using vectorial methods to manipulate words or groups of words

7

A model of Word Vector : Random Indexing

- Create a matrix $A (d \times N)$, containing *index-vector*
 - d is the number of documents or contexts corresponding to the corpus
 - N , the number of dimensions ($N > 1000$)
 - Index-vector are sparse and randomly generated vectors. They consist in few numbers of +1 and -1 and **hundreds of 0**

8

A model of Word Vector : Random Indexing

- Create a matrix $B (t \times N)$ containing the *term-vectors*
 - t is the number of terms composing the corpus
 - The process is incremental. To start the matrix compilation, all cells values are initialized to 0

9

A model of Word Vector : Random Indexing

For each document of the corpus, each time a term t appears in a document d ,

- Accumulate the *index vector* corresponding to the document d to the *term vector* corresponding to the term t

10

A model of Word Vector : Random Indexing

At the end of the process, *term vectors* that appeared in similar context have accumulated similar *index vectors*.

11

A model of Word Vector : Random Indexing

- The particularities of RI :
 - As efficient as LSA (sometimes more)
 - No matrix reduction (resource demanding in LSA)

12

Application of an episodic memory modeling perspective on text categorization using Random Indexing

- From episodes to concepts
 - A famous model of episodic memory: MINERVA 2
 - Description
 - Simulation of the effect of frequency episodes
 - A Word Vector model: Random Indexing
 - Description
 - Text categorization and the application of the distributional hypothesis
- The DEFT09 text-mining contest
- Results
- Perspectives

13

DEFT'09

- A French Text Mining contest
- DEFT'09 main task was Opinion categorization:
 - Subjectivity/Objectivity detection in multi-lingual journal corpora (fr, en, it)
 - 60% of the corpora for training
 - Limited time test period (3 days)

14

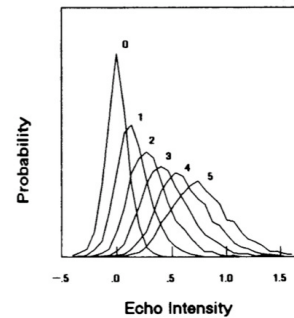
Principles

- Build a semantic memory from all the available episodes
- Organize episodes in categories following principles of episodic memory models
 - Splitting the categories into homogeneous sub-categories regarding their *typicality*

The created sub-categories are considered as a local episodic memories

15

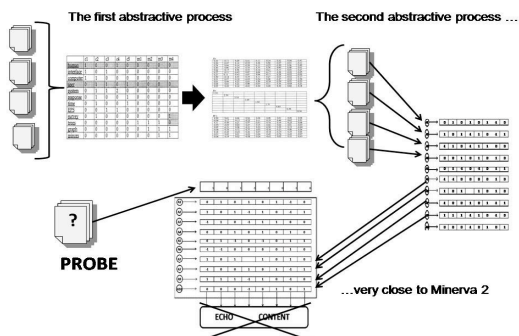
Effect of frequency of episodes on echo (Hinztman, 1988)



Mean and variance of echo increase with frequency

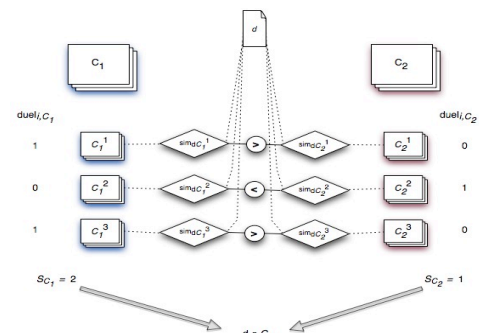
16

Principles



17

Assigning a category



18

Application of an episodic memory modeling perspective on text categorization using Random Indexing

- From episodes to concepts
 - A famous model of episodic memory: MINERVA 2
 - Description
 - Simulation of the effect of frequency episodes
 - A Word Vector model: Random Indexing
 - Description
 - Text categorization and the application of the distributional hypothesis
- The DEFT09 text-mining contest
- Results
- Perspectives

19

Results

		French	English	Italian			
Number of Documents	Appr.	25176	7866	1496			
	Test	16788	5245	999			
Size (Kb)	~						
	Appr.	80000	25000	6000			
	Test	51000	16000	3500			
Number of Dimensions		5120	1500	4096			
Number of Cycles		50	20	40			
Number of Sub-categories		5	9	5			
Precision	Obj	0.740	0.941	0.746	0.59	0.710	0.828
	Subj		0.540		0.901		0.591
Recall	Obj	0.803	0.869	0.719	0.927	0.723	0.681
	Subj		0.738		0.510		0.765
F-score		0.771		0.732		0.716	

20

Application of an episodic memory modeling perspective on text categorization using Random Indexing

- From episodes to concepts
 - A famous model of episodic memory: MINERVA 2
 - Description
 - Simulation of the effect of frequency episodes
 - A Word Vector model: Random Indexing
 - Description
 - Text categorization and the application of the distributional hypothesis
- The DEFT09 text-mining contest
- Results
- Perspectives

21

Perspectives

- A cognitive model of Text Categorization: Alida
- Application to a larger opinion categorization tasks: TREC09-Blog track
- Interfacing with other Word-vector methods (LSA, HAL, ...)

22

Possible applications in Education

- Educational resources management:
 - Resource retrieval: help users to determine the value of an education resource (factual vs. opinion)
 - Resource classification in both thematic and opinion dimensions
- Assessment or essay-scoring:
 - Assigning a value of objectivity to an essay

23

Alida

- Yann Vigile Hoareau, Adil El Ghali, and Denis Legros **Approche Multi-traces et catégorisation de textes avec Random Indexing**. Dans les Actes de atelier de clôture de l'édition 2009 du défi fouille de texte DEFT'09, 22 Juin 2009, Paris, France.
- Yann Vigile Hoareau and Adil El Ghali **Typicality modeling and application to text categorization**. In *Proceeding of Recent Advances in Natural Language Processing RANLP'09*, September 14-16, 2009, Borovets, Bulgaria.
- Yann Vigile Hoareau and Adil El Ghali **Random Indexing and the episodic memory metaphor. Application to text categorization**. In *Proceeding of the 14th annual International CSI Computer Conference CSICC'09*, October 20-21, 2009, Tehran, Iran.

24

Thank you

25

Special thanks to

- Alice
- Axel
- Charles
- Denis
- Rose
- Sandra
- ...

Alida

26