Automatic Categorization of Job Postings

Julie Séguéla, Gilbert Saporta

Automatic Categorization of Job Postings

Julie Séguéla[1,2] and Gilbert Saporta[1]

[1]Laboratoire Cédric, Conservatoire National des Arts et Métiers, Paris, France

[2]Multiposting.fr, 33 rue Réaumur, Paris, France, *jseguela@multiposting.fr*

Context & Objectives

• **Context :**

-Recruiters need to assess job posting performance on job search web sites

-But all sites have different ways to reference job postings

• **Objectives :**

-Be able to identify the job occupation whatever the posting

-Improve the categorization efficiency thanks to a processing adapted to the particularity of job postings

Corpus

• French corpus of 700 job postings (baseline method : 1060 terms)

• Two-level classification : focus on 4 categories / 14 occupations

Sales / Business Development		Marketing / Product	
Business Development / New Accounts	Telesales	Marketing Communications	Marketing Production
Sales Support / Assistance	Field Sales	Product Management	
IT / Software Development		**Accounting / Finance**	
Network and Server Administration	IT Project Management	Bookkeeping / General Ledger	Corporate Accounting
Software / System Architecture	Systems Analysis	Claims Review and Adjusting	

Method : Posting Representation & Classification

Method : Multi-term word identification

• Identify repeated segments to take into account multi-term words

Method : Title term weighting

• Title terms are weighting twice

Method : χ^2 statistic tresholding (650 terms)

• Selection of occupation most specific terms

Method : Baseline (preprocessing and first selection)

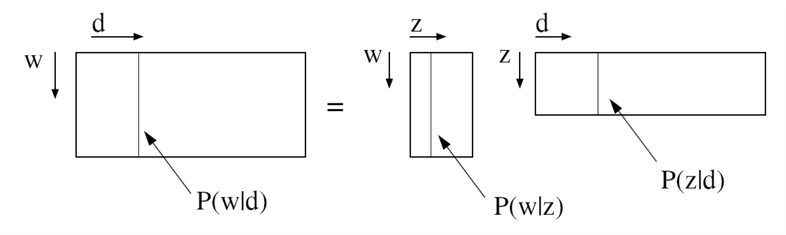
• Filter out duplicates
• Lemmatisation (with TreeTagger)
• Filter out prepositions, adverbs, pronouns…
• Document frequency tresholding (df > 10)
• Term frequency representation

Method : Probabilistic Latent Semantic Analysis

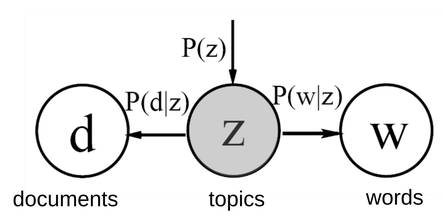
• **Objectives :** Adress the issue of polysemy and synonymy
Representation of documents in the latent semantic space

• **Statistical model :** Latent variable model for co-occurrence data (« Aspect model »)

Aspect model representation

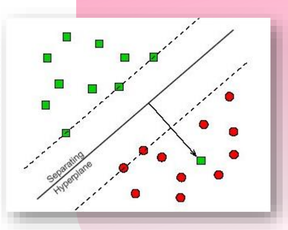


Aspect model in the symmetric parameterization



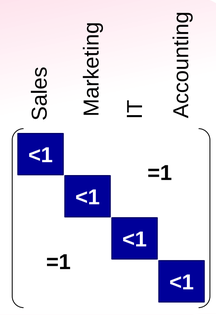
Topic features $P(z|d)$

Classification : SVM (linear kernel and error cost parameter tuning)



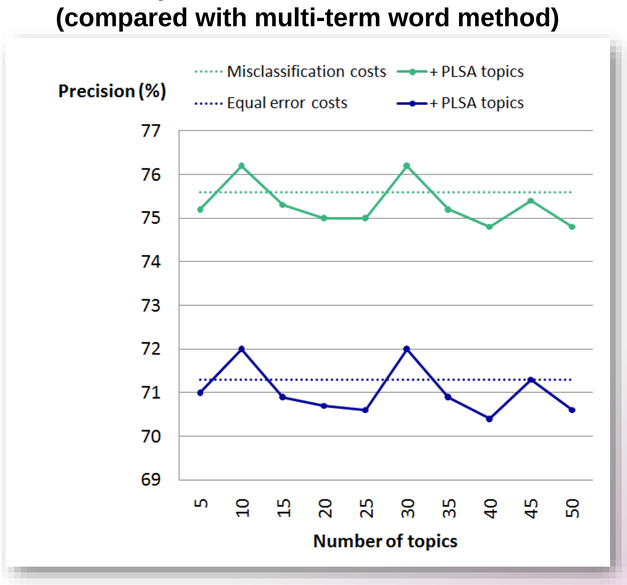
Classification : performance evaluation

• Average precision on occupations
• Deducting precision on categories
• Implementation of a *misclassification cost matrix*



Results

Classification precision using PLSA latent topics as additional features (compared with multi-term word method)



• It is necessary to choose carefully the number of topics

Two PLSA topic factors (from a 30 factor decomposition) with high probability to generate the word "development", represented by their most probable words

« IT »	« Sales »
development	commercial
study	sale
project	sector
java	develop
implementation	customer
computing	development
technique	prospect

• The word "development" occurs in two different contexts

Comparison of method average precisions
(without and with implementation of a misclassification error cost)

Method	Job occupations (equal error costs)	Job occupations (misclassification costs)	Job categories
Baseline	64,4	69,4	91,6
χ^2 statistic tresholding	68,3	72,8	92,2
Title term weighting	69,2	73,6	92,5
Multi-term word identification	**71,3**	**75,6**	91,4
PLSA	**72**	**76,2**	92,8

Conclusions & Future Work

• Our methodology allows to :

-Standardize the job category on all our posting database

-Automate job posting labeling according to each job board referencing system

• **Future work :**

-Use the job category to recommend the use of job search web sites in order to optimize recruitment performance

References

Hofmann T. (1999) Probabilistic Latent Semantic Analysis. In: *Proceedings of UAI'99, Uncertainty in Artificial Intelligence*. Stockholm.

Joachims T. (1997) Text Categorization with Support Vector Machines: Learning with Many Relevant Features. In: *Machine Learning: ECML-98*. Springer, Berlin, 137-142.

Yang Y. and Pedersen J. P. (1997) A comparative study on feature selection in text categorization. In: *Proceedings of ICML'97, International Conference on Machine Learning.* Nashville, US, 412-420.