Bayesian Discriminant Analysis for Lexical Semantic Tagging

Nicolas Turenne

European Meeting on Cybernetics and Systems Research (EMCSR), Vienne (Autriche), 2002

Bayesian Discriminant Analysis for Lexical Semantic Tagging

Nicolas Turenne
UMR INRA-INAPG
Biométrie et Intelligence Artificielle (BIA)
16 rue Claude Bernard,
75231 Paris cedex 05 France
email: turenne@inapg.inra.fr

Abstract

Structuring of terminology automatically is interesting for information extraction or indexing. It remains to evaluate the results obtained. We use a clustering method to build term classes assuming that (even) an incomplete thesaurus could inform the user about the semantic interpretation of classes. A variant naive Bayesian analysis is used to model an association network between terms and categories. Conditional relations express a causal membership of a term to several categories. We show that a lexical tagging of terms classes, obtained from corpora processing, performs with a high degree of probability

1 Introduction

To describe semantics of words, word distribution in texts have been studied since early nineties [Yarowsky, 1992; Zernik,1991]. The goal was stimulated by: application of pure classification methods, bibliographic data processing and lexicon building for linguists. Nowadays, however, this research is characterized by a mixture of models (graphs, statistics, learning, linguistics) and is applied with a view to meeting real-world needs [Hearst, 1999; Frank et al., 1999; Grefenstette, 1994].

The goal of this paper is to propose a method to evaluate term classes extracted from texts by merging classes to prior categories.

Firstly, we use a term classifier (Galex, Graph Analyzer for Lexicometry) [Turenne, 2000] to obtain term classes. Secondly, to evaluate the relevance of these classes without human intervention we decide to link a thesaurus category to each class. By doing this, we focus this study to the generalization relation.

Semantic ambiguities appear in a natural language as one of the main linguistic phenomena. At the lexical level, a given word can belong to several categories (example: *table* means a furniture o and also a list).

A Bayesian approach can describe this phenomena with an association network and causal relations [Bernardo and Smith, 2000; Mladenic and Grobelnik, 1999] developed a naive Bayesian classifier applied on text data to feature subset selection and for document categorization. Their purpose was to integrate unbalanced class distributions and study misclassification costs. Our purpose is not classification learning but modeling relations between a term and its semantic fields with a belief network. The Naive Bayes method belongs to state of the art for sense disambiguisation. According to a variant of a naive Bayes method, we maximize the probability $P(C/T) = P(C/T_1, T_2, ..., T_n)$, to observe the category C knowing the set of terms T= $\{T_1, T_2, \ldots, T_i\}$.

In section 2, we propose the framework of our discriminant analysis model. In section 3, we present this model for semantic tagging. The model is essentially used to assess the coverage of a term cluster by a significant category. Finally, we present other approaches of information extraction from texts and the mean to qualify results.

2 Framework

We can define a process of tagging as a relational problem between two sets. Let T=$\{T_1,\ldots,T_n\}$ be the set of terms and N a network linking terms of T'=$\{T'_1,\ldots,T'_m\}$ and categories of C=$\{C_1,\ldots,C_p\}$. We want to know which item of C to choose for tagging the set T. The method consists in evaluating the overlapping of T and T' and get the item associated to the T' elements.

Each term of N defines a relation with a category so overlapping needs the knowledge of this conditional dependence. Moreover N admits two principal properties:

- Relation of generalization between terms and categories
- Ambiguity of a term which can belong to several categories.

The specific relational dependence and its uncertainty can be performed by a Bayesian approach. A Bayesian discriminant analysis computes the maximum likelihood to observe a category knowing a series of terms.

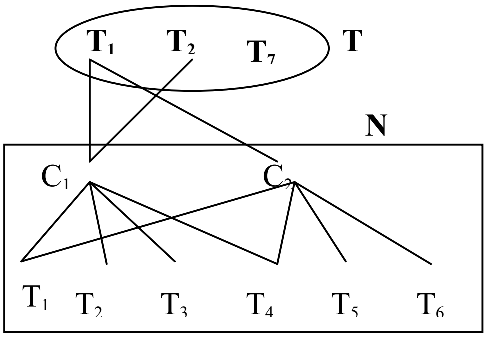

Figure 1. Network between categories and terms.

The analysis consists in determining $P(C_j/T)$ as follows in a Bayesian formulation:

$$P(C_j/T) = \frac{P(T/C_j)P(C_j)}{P(T)} \cdot \qquad (1)$$

In the following of the study we consider C_j as a numerical code designing a category. As terms are uniformly distributed in N we assume that $P(C_j)$ is invariant. We suppose that T is a set given without uncertainty. Though a classical naive Bayesian method assumes that all individual T_i are non exclusive, we do not guess any hypothesis about joint probabilities occurring in C_j, we obtain:

$$P(T/C_j) = \sum_i P(T_i/C_j) = \sum_i f_j(T_i) \cdot \qquad (2)$$

$$f_j(T) = \sum_i f_j(T_i) \cdot \qquad (3)$$

$$P(C_j/T) = A.f_j(T) \cdot \qquad (4)$$

Finally we have to maximize $f_j(T)$, the probability to observe T knowing the jth category. We do not make any hypothesis about the distribution of T over categories and we use a non-parametric method to compute $f_j(T)$. We consider the number of observation k_{ij} of any term i of T in the jth category:

$$f_j(T) = \frac{\sum_i k_{ij}}{\sum_j \sum_i k_{ij}}, \text{ where } T_i \in T \text{ and } k_{ij} = \begin{cases} 1, \text{ if } T_i \in C_j \\ 0, \text{ else} \end{cases} \qquad (5)$$

So the final set of results is :

$$S = \{C_j \mid \max(f_j(T))\}. \qquad (6)$$

This set gathers all category codes having the maximum likelihood to observe T. This set may have several items. We decide to keep only one item, Cat, solution of the tagging. A heuristics aims to keep the category having the smallest code. But if the cardinal of S is null another heuristics takes into account a seed term performing the choice of Cat. Finally, Cat is assigned to T:

$$\text{Cat} = \begin{cases} C_j & \text{if } \operatorname{card}(S) = 1 \\ \min(S) & \text{if } \operatorname{card}(S) > 1 \\ C_p & \text{if } \operatorname{card}(S) = 0, T_p \in C_p \end{cases} \qquad (7)$$

3 Semantic Tagging

3.1 Thesaurus structure

Our aim is to test the lexical cohesion of a group of terms considered as a class [Loukashevich and Dobrov, 2000; Elman, 2000]. To do this we try to tag the cluster of terms with a category of a reference thesaurus. The thesaurus contains terms dispatched (non exclusively) under sub-categories themselves dispatched (exclusively) under categories:

- level 0 : terms, around 120,000 ,
- level 1 : sub-categories, 873 in all (semantic fields),
- level 2 : categories (themes), 26 in all.

We have lemmatized with our own tool to process the terms present in the thesaurus as terms of classes obtained with the classifier. This lemmatization reduces the number of forms to around 100,000.

example :

sub-category range	code
773-791	10 (art)

sub-category	code
photography	775

775 (sub-category code)

device ; camera ; photo booth ; photorama ; photographic rifle

Before using the consensus model we extract terms classes with our term classifier Galex (Graph Analyzer for Lexicometry)(Turenne 2000). The corpus is a collection of full texts referring to one subject. A term extractor extracts noun groups considered as terms of the domain included in the corpus. Hence the classifier makes classes of terms that we want to evaluate as a semantic field of the corpus subject.

Our experiments more often take into account a medical corpus talking about coronary diseases and mainly constituted by medical reports.

Terms of the class	Collected codes	Decision
Anomaly, coronary spasm, thallium, prognosis, methergin, woman, question, left function, segment, right coronary	210,248,328,345,326, 331,391,331	Heart(331)

Table 1. Automatic tagging of a group of terms.

We gather the total score obtained for each code of the thesaurus (Table 1.). The code having the maximal frequency is assigned as the semantic tag to qualify the lexical cohesion of the given class of terms [Turenne 2000] (Table 2).

3.2 Algorithm

Algorithm 1 first level tagging

```
1- Collect all codes of the terms included in a class.
2- **If** a term is composed
            search the code of the whole term if it exists
            **else** collect the codes of the first and
            the last word of the term. Except if the
            last word is an adjective.
3- **If** it exists, choose the most frequent code as tag
   of the class
        **if** several codes have the same maximum
                frequency
                        choose the smallest one
                **else** search the codes of the pole term
                        choose the smallest one.
```

(the heuristics consisting in choosing the smallest code for convenience can be explained by the fact that the least weighted code is more general)

Algorithm 2 second level tagging and global tagging

```
1- **Assign** each code of the interest center from the 1st-level
to a code of the 2nd-level included in a range of code of
the 2nd-level. For instance: code 248 is the range 230-267
related to "matter"; so the node 2 for the class is "matter".
2- **Select** the themes of the corpus which are prevailing:
- Collect all codes from all classes ,
- Calculate the frequency of each code,
- Sort codes by increasing order of frequency.
3- **Choose** the 3 at the top of the list (frequency > 3)
```

3.3 Results

Class	**Category (lev. 1)**	**Category (lev. 2)**
Star , planet , solar system , satellite ,earth, moon, comet	World	Matter in general
human rights, council of Europe, minorities, convention , parliamentary assembly, committee of ministers, countries	Council	Voluntary action

Table 2. Automatic tagging of manually created classes.

Class	**Category (lev. 1)**	**Category (lev. 2)**
Contrast medium, development, anterior askew, technique, test, treatment, ventricle, ventriculography	Medicine	Medicine
Single incidence, obstruction diameter, possible to expand, angioplasty	Dimension	Dimensions
Aortic valve, coronary, aorta, circulation	Heart and vessel	Body
Cardiac catheter technique, lesion, examen, angor, catheter	Method	Order

Table 3. Automatic tagging of automatically created classes.

Number of occurrences	**Code of category**	**Name of the category**
31	*383*	*Disease*
23	*331*	*Heart*
21	*391*	*Medicine*
11	792	Job
10	185	Period

Table 4. Most frequent categories assigned to the clusters.

The second generalization processing performs the selecting of codes related to each class (Table 3.) and to class them by increasing order to keep the most frequent ones (Table 4.). The 3 codes being the most frequent are considered to qualify the whole set of classes and then the corpus.

Concerning the medical corpus, an overview of the assignation process shows the following results: 21 classes had been tagged by a medical category and 43 had been tagged by a state category; finally 85% had been "logically" classified in a theme related to their semantic content of the corpus. This last result remains a human appreciation and underlines the difficult problem of evaluation balancing between personal point of view and automatic robust heuristics.

Finally, we tested our tagging algorithm on a mixed thematic corpus. To do this we built a corpus from an encyclopedia (Britannica™) concerning aeronautics, and the history of Russia. The size of the corpus was enough small, around 70,000 words. The corpus processing leads to a discrimination of the subjects through the extraction of term classes obtained with our Galex classifier [Turenne, 2000]. From 61 extracted classes, 27 were related to the history of Russia, 19 were related to the field of aeronautics and 15 were ambiguous. From the whole set of classes around 75 % could be successfully assigned to their respective themes.

4 Related Work

[Grefenstette, 1996] exploits thesauruses to test lexical cohesion of word pairs (target word / most significant contextual word). Results aim at comparing the cohesion of the most significant words either by a syntactic method (relation adjective-noun, noun-verb…), or by a method of co-occurrence based on a window of 10 words (left and right), the most significant words being determined with a Jaccard coefficient. As the probability that 2 terms belong to the same category is less than 1%, the use of a thesaurus looks helpful but not really sufficient on its own to determine the semantics of a pair. [Tishby et al., 1999; Feldman, 1997] have set up a hierarchy of concepts manually related to a theme according to a corpus of documents for a given theme. In their study term distributions are compared for a given node of the hierarchy to a calculated relative entropy. The *Syndikate* system [Hahn, 1997] proposes to choose the most significant concept from an ontology (i.e. subsumer) according to a description logics reasoning. This system exploits : firstly, qualitative knowledge about linguistic properties present in free texts, and secondly structural configuration in knowledge bases of a domain (345 concepts and 347 relations). [Valtchev et al., 2001] has studied fusion of Concept lattices with Formal Concept Analysis (FCA). FCA is a discipline that studies the hierarchical structures induced by a binary relation between a pair of sets. The structure, made up of the closed subsets ordered by a set-theoretical inclusion, satisfies the properties of a complete lattice. [Valtchev et al., 2001] provide the foundation of an efficient lattice assembly procedure carrying out a filtering of the direct product of the partial lattices, which retrieves the concepts of the global lattice and their precedence links. They base their algorithm on the intersection of attributes. We cannot performs such fusion since we do not have at our disposal only sets of terms without their attributes.

5 Conclusion and perspectives

A number of approaches focuses on class extraction by similarity. This is a difficult task since no a priori knowledge is assumed. In our paper, we have presented for this purpose a variant of a Naive Bayesian approach. It is an evaluation of our clustering method by categorizing clusters. Before performing the tagging task, a classifier extracts a set of term classes from a corpus. Then, a tagging strategy implements a Bayesian decision rule so as to choose which semantic tag is more related to the whole set of terms of a given class. The tag belongs to a set of categories defining a hierarchy of a general thesaurus. The tagging tries to implement a kind of fusion of a simple level hierarchy of terms with a set of term clusters. The experiments revealed some difficulties to perform efficiently with corpora constituted of email data. Noun phrases are not referenced in the thesaurus or do not have common usages if they belong to the same cluster. With encyclopedic texts we get more interesting results with a rate of 70% of good tags. The assignation of tag to the whole set of clusters works very well with a rate of 100% whatever the theme of the corpus processed. The interest of this approach lies in the simplicity of its implementation and in its efficiency to give a rough information about generality of a class according to a reference. We expect to use such semantic tagging in a filtering process to analyze the capability of categories, provided by clusters, to match with a probable category assigned to a new document.

References

[Bernardo and Smith, 2000] Bernardo J. and Smith A., *Bayesian Theory*, John Wiley & Sons Ltd, 2000.

[Elman, 2000] Elman J., On the Generality of Thesaurally derived Lexical Links, in the Proceedings of *International Conference of Textual Statistical Data Analysis (JADT)*, Lausanne, 2000.

[Feldman and Dagan, 1997] Feldman R., Dagan I., Knowledge Discovery in Textual Databases (KDT), in the Proceedings of the *1st International Conference on Knowledge Discovery (KDD),* Montreal, 1997.

[Frank and Paynter, 1999] Frank E, Paynter G.W. , Domain-specific Learning Algorithms for Keyphrase Extraction, in the Proceedings of the *16th International Joint Conference on Artificial Intelligence*, Stockholm, 1999.

[Grefenstette, 1996] Grefenstette G., Evaluation Techniques for Automatic Semantic Extraction: Comparing Syntactic and Window Based Approaches, in *Corpus processing for Lexical Acquisition* ed. J.Pustejovsky MIT 1996.

[Hahn and Schnattinger, 1997] Hahn U., Schnattinger K., Knowledge Mining from Textual Sources, in the Proceedings of the *International Conference on Information and Knowledge Management (CIKM)*, Las Vegas, 1997.

[Hearst, 1999] Hearst M., Untangling Text Data Mining, in the Proceedings of the *Association for Computational Linguistics Conference*, University of Maryland, 1999.

[Loukashevich and Dobrov, 2000] Loukashevich N., Dobrov B., Thesaurus as a Tool for Automatic Detection of Lexical Cohesion in Texts, in the Proceedings of the *International Conference of Textual Statistical Data Analysis,* Lausanne, 2000.

[Mladenic and Grobelnik, 1999] Mladenic D. and Grobelnik M., Feature selection for unbalanced class distribution and Naive Bayes, in the Proceedings of *16th International Conference on Machine Learning*, Morgan Kaufmann Pub., San Francisco, 1999.

[Tishby et al., 1999] Tishby N., Pereira F. and Bialek W., The Information Bottleneck Method, in the Proceedings of the *37th Annual Allerton Conference on Communication Control and Computing*, University of Illinois, 1999.

[Turenne, 2000] Turenne N., *Apprentissage statistique pour l'extraction de concepts à partir de textes. Application au filtrage d'informations textuelles*, PhD thesis of the Université Louis-Pasteur, Strasbourg, 2000.

[Valtchev et al., 2001] Valtchev P, Missaoui R, Lebrun P A partition-based approach towards constructing Galois (concept) lattices, Forthcoming in *Discrete Mathematics*, 2001.

[Yarowsky, 1992] Yarowsky D., Word-Sense Disambiguation using Statistical Models of Roget's Categories trained on Large Corpora, in the Proceedings of the *Computational Linguistics Conference (COLING)*, Nantes, 1992.

[Zernik, 1991] Zernik U. , Train 1 vs Train 2: Tagging Word Sense in a Corpus , in Zernik,U (ed.) *Lexical Acquisition: Exploiting on-Line Resources to Build a Lexicon*, Hillsdale, NJ: Lawrence Erlbaum Associates, 1991.