Using Information Extraction to Classify Newspapers Advertisements

Ramón Aragüés Peleato, Jean-Cédric Chappelier and Martin Rajman

EPFL - DI - LIA - INR (Ecublens) - CH-1015 Lausanne - Switzerland
{aragues,chaps,rajman}@lia.di.epfl.ch

Abstract

This paper presents a text classification procedure that has been developed in the context of an information extraction project. In the prototype that has been developed for this project, newspaper advertisements are processed by three main modules: first of all, a classification module associates a category to the advertisement. Then, a tagging module identifies textual information units that are related to the associated category, and finally a predefined form for that category is filled with the tagged text.
The classification module, which is the main focus of this paper, consists in using a naive Bayes classifier and, at the same time, trying to fill all the predefined forms associated with all categories. Results of both methods (classification probabilities and filling scores) are then combined to provide a final classification decision. This mixed classification method is described and evaluated on the basis of concrete experiments carried out on real data. The purpose of the presented experiments is to precisely evaluate the impact of the information extraction step on classification accuracy. As one could reasonably expect, classification relying on information extraction alone doesn't perform very well but when used as a complement to the statistical approach it significantly improves the classification results.

Keywords: Text Classification, Information Extraction, Automated Newspaper Advertisements Processing, naive Bayes classification.

1. Introduction

The work reported in this paper has been carried out in the context of the development of a system able to automatically extract and structure information out of the textual content of newspaper advertisements. The system consists of three main modules, as illustrated in figure 1:

1. a classification module: the task of this module, which is the main focus of this paper, is to classify the processed advertisements into a set of 4 *a priori* known classes (*real estate*, *vehicles*, *employment* or *other*). As each of the classes is associated with a form that defines the fields in which the extracted information should be structured, the objective of the classification step is to identify which form has to be associated with the advertisement to guide the information extraction process.
2. a tagging module: the task of this module consists in labelling the textual content of the advertisement, in order to identify the information units that have to be extracted (segmentation) and the slots of the selected form they have to be associated with (tagging). The slots represent different features describing the category (e.g. make, colour, year, ... for class vehicles). Tagging is achieved by simultaneously using specialized lexica, regular expressions, word spotting techniques and relative position analysis (Aragüés et al., 2000).

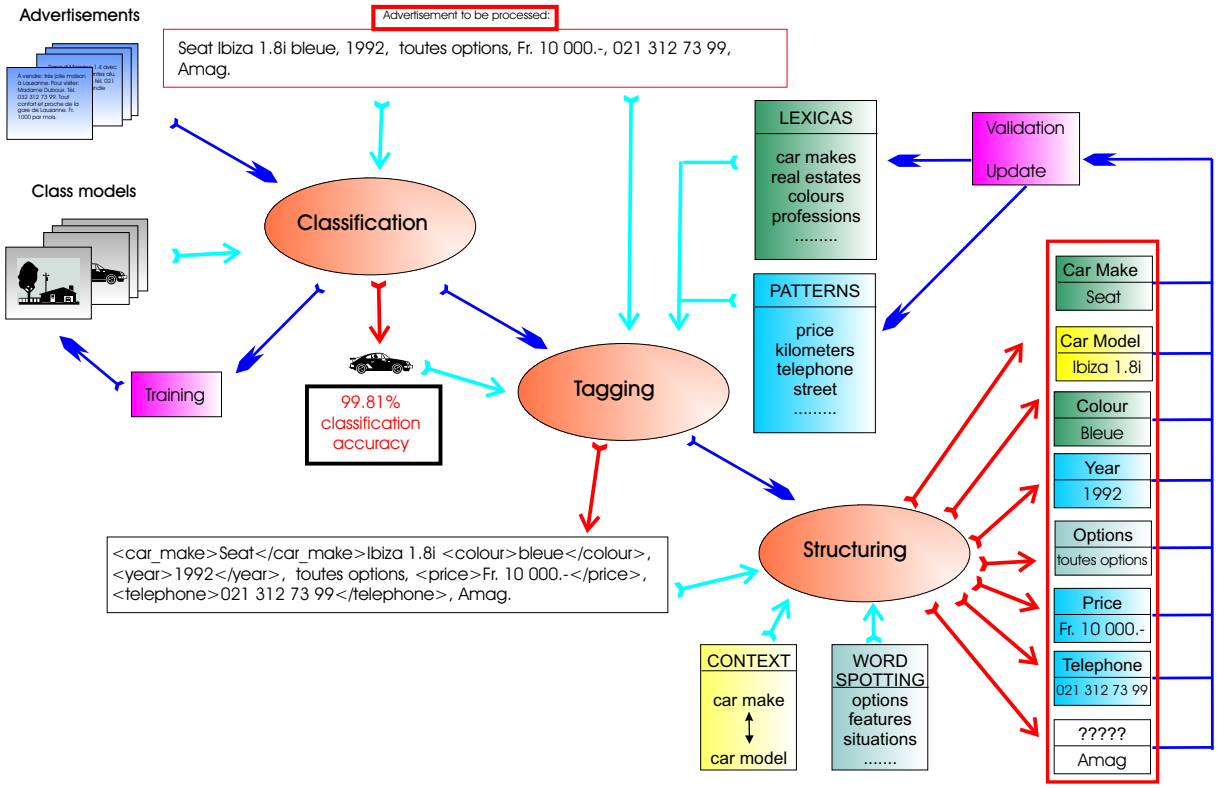


Figure 1: *Global architecture of the system for automatic processing of newspaper advertisements.*

3. a structuring module: this module is in charge of transforming the tagged text into an organized data structure (concretely a filled form). This involves extracting the tagged textual units, standarizing formulations[1], removing inappropriate punctuation, transforming abbreviations, etc. However, in the current prototype, this module remains quite simple as the tags used in step 2 closely correspond to the slots present in the associated forms.

As already mentioned earlier, the present paper focuses on the classification module, that simultaneously integrates statistical and information extraction classification approaches.

One approach to text classification is statistical methods, where naive Bayes or K-nearest neighbours (Hoyle, 1973; Yang, 1999) can for instance be considered. Such methods try to classify a text by comparing it to preexisting models of categories that have been learned from representative data. These models are represented in terms of word frequencies and co-frequences, a text being classified into a given category if it is "close enough" to the corresponding model. As it is described in section 2.1 the statistical part of the mixed classification approach uses a pure naive bayes techique.

Other approaches to text classification include techniques relying on information extraction. Work developed in the last years has shown that good classification scores can be achieved by assigning a category to a text depending on how much relevant information for that category can be extracted from the text (Riloff and Lehnert, 1994).

[1] for example, using the same format for all price indications.

When properly tuned, both approaches can lead to high classification scores, however, they both suffer from several limitations:

- when automatically trained on reference data, statistical methods mainly classify on the basis of word frequencies, and are therefore not very adequate to take into account the *a priori* knowledge about the importance of certain words for some particular category: some words may be important for a category in spite of the fact that they have a low frequency in the training data, for example a low frequency car make for the vehicles category.
- information extraction techniques usually obtain low recall scores because of the fact that documents not containing much relevant information are not properly classified, even if they do belong to some well defined category.

The classification procedure developed in our project aims at overcoming these limitations by mixing the two approaches. The idea is to test how much an *a priori* filling up of all predefined forms can bring to the classification procedure. The method considered consists in classifying with a standard statistical classifier (naive Bayes, section 2.1) and, in parallel, trying to fill all the forms in order to assign to each of the categories a score based on information extraction success (section 2.2). The results of the statistical and information extraction classifications are then combined to produce the final classification decision (section 2.3).

2. Classification Methods

2.1. Naive Bayes classification

For statistical classification, we used a standard naive Bayes classifier (Joachims, 1997; Mitchell, 1997). Let $w_1^n = w_1 ... w_n$ denote the n words representing the textual content of the advertisement to be classified and C a category. The classification score used is

$$P(C) \cdot \prod_{i=1}^{n} P(w_i|C)$$

where $P(C)$, the prior probability of category C, and $P(w_i|C)$, the likelihood of word w_i jnowing C, are estimated on a labelled training corpus.

A given advertisement is then classified in the category that maximizes the classification score, unless the scores for all the categories are below a given threshold[2], in whichx case, the advertisement is considered as unclassified.

The naive Bayes classifier used was RAINBOW from the BOW package (McCallum, 1996). We trained it on a corpus of 51,301 advertisements sampled over a period of five years. The average length of a advertisement in the training corpus was of 16.4 "words"[3]. The vocabulary contained 29,225 different words with an average occurrence frequency of 29.

2.2. Information Extraction based classification

Each advertisement category is associated with a form that describes the category in terms of specific slots (e.g. make, colour, year, ... for vehicles). The idea behind an information extraction classification is to find the category to be associated with an advertisement by computing a

[2] 0.95 for the experiments described below

[3] tokens strictly speaking

score related to the number and importance of the slots that can be filled for this advertisement by using various information extraction techniques. To compute such a score, each slot is given an *a priori* fixed weight (between 0 and 100) that measures the importance of the slot for the category. For instance, the slot "*make*" of the VEHICLE form had a weight of 100, whereas the slot "*colour*" only had a weight of 10. These weights where manually assigned by a human expert using his knowledge of the domain.

For classification, the systems applies to the advertisement to be classified various automated information extraction procedures (Aragüés et al., 2000), in order to fill as many slots as possible for each category. The chosen category is then the one that maximizes the sum of the weights of the filled slots, unless the sums for all categories are below a given threshold[4], in which case the advertisement is considered as unclassified.

2.3. Mixed classification

In the mixed approach, each advertisement is first classified with both the statistical and the information extraction based classifiers. Classification results are then combined according to the following rules:

- If an advertisement is unclassified for both methods, the final result is "unclassified";
- If an advertisement is classified for only one of the two methods, it is classified according to this method;
- If both methods associate a class with the advertisement:
  - If both methods agree on the class, the final result is the corresponding class;
  - If the two methods disagree then two methods were tested: either the final result is "unclassified" (balanced mixture) or the advertisement is classified according to one of the methods on the basis of an *a priori* priority decision (mixture with priority).

The rationale behind the proposed approach is to improve classification decisions is the following:

- Statistically unclassified advertisements will have additional chances to be classified by the information extraction based classifier, and therefore recall will be improved.
- Many statistically misclassified advertisements can be detected in the mixed approach on the basis of their low filling score[5]. In the balanced mixture approach, such advertisements will be considered unclassified leading to higher precision. In the mixture with priority a approach, the precision will not be affected [6].

3. Experiments and Results

To evaluate the impact of the mixed classification techniques, results were computed on various test sets built from real data.

3.1. The test base

Tests were made on a set of 2,856 advertisements distributed over three classes and sampled over a period of 6 months. The size of the vocabulary for that test set was of 7,564 defined

[4] set to 50 in the reported experiments

[5] indicating that in spite of its statistical relevance, the advertisement actually does not contain relevant information

[6] Except for the case where the statistical classifier considered the advertisement as unclassified and the information extraction based classifier classified it in a wrong category.

word forms. The average length of a advertisement was about 25.5 words and the average word occurrence frequency was 9.6. As for the training set, the reference categories assigned to the advertisements were defined manually by a human expert. There was no intersection between the test set and the training set.

Performances on the test set of the three methods (naive Bayes, information extraction based and mixed) are presented hereafter with their confusion matrix and average[7] precision and recall. For a given class, precision is the number of well classified advertisements over the number of advertisements classified in this class and recall is the number of well classified advertisements over the number of advertisements in the class.

3.2. Results

The following confusion matrices summarize the results for the three considered classification methods. A well classified advertisement will contribute to the scores that appear on the diagonal of the matrix. Advertisements outside the diagonal have been either mis- or unclassified in the class corresponding to the column they appear in. Class identifiers are the following: 1:Real Estate, 2:Employment and 3:Vehicles.

Naive Bayes

class	1	2	3	unclassified	Total	Recall
1	1686	16	1	46	1749	96.39%
2	6	781	0	40	827	94.44%
3	0	0	234	2	236	99.15%
Total	1692	797	235	88 (3.13%)	2812	avg=**96.66%**
Precision	99.65%	97.99%	99.57%	avg=**99.07%**		

Information Extraction based

class	1	2	3	unclassified	Total	Recall
1	1620	25	1	103	1749	92.62%
2	24	658	2	143	827	79.56%
3	3	0	229	4	236	97.03%
Total	1647	683	232	250 (8.89%)	2812	avg=**89.74%**
Precision	98.36%	96.34%	98.71%	avg=**97.80%**		

Balanced mixed classification

class	1	2	3	unclassified	Total	Recall
1	1684	7	0	58	1749	96.28
2	5	777	0	45	827	93.95
3	0	0	231	5	236	97.88
Total	1689	784	231	108 (3.8%)	2812	avg=**96.04%**
Precision	99.7%	99.11 %	100 %	avg=**99.60%**		

[7] over the three classes

In the experiment with balace mixed classification, the distribution of advertisements unclassified due to disagreements between two methods is the following ("IE" stands for "Information Extraction"):

class	disagreements	statistical classifier correct	IE-based classifier correct
1	35	22 (63 %)	13 (37 %)
2	23	22 (96 %)	1 (4 %)
3	3	3 (100 %)	0 (0 %)
total	61	47 (77 %)	14 (23 %)

On the basis of these results, we tested mixed classification giving priority to the statistical method, i.e. the answer in case of disagreement between the two methods is the one of the statistical classifier. The corresponding results are given in the following confusion matrix.

Mixed classification with priority to naive Bayes

class	1	2	3	unclassified	Total	Recall
1	1706	19	1	23	1749	97.54%
2	6	799	0	22	827	96.61%
3	0	0	234	2	236	99.15%
Total	1712	818	235	47 (1.7%)	2812	avg=**97.77%**
Precision	99.65%	97.68%	99.57%	avg=**98.97%**		

3.3. Significance of the results

In order to evaluate the statistical significance of the differences between the results obtained for the three methods, we ran several separate evaluations on random subsamples[8] of the test set. This second set of measures was used to estimate the variances[9] of the performance measures presented above.

For 10 runs, we obtained the standard deviations that are given in the table in section 4 and the statistical significance of the differences between the different methods with relation to the naive bayes approach is indicated in the following table:

	Balanced mixture		**Priority to bayes**	
	Precision	Recall	Precision	Recall
Significant at 99%	better	no discernable difference	no discernable difference	better

These results show that the improvement of precision with the balanced mixed classification is significantly better (at 99%) than the pure statistical method, and the decrease in recall is not significant at 99%[10].

Concerning mixed classification with priority to the statistical method, the decrease in precision appeared not significant (even at 95%) but the improvement of recall is significant at 99%.

[8] generated with bootstrap methods

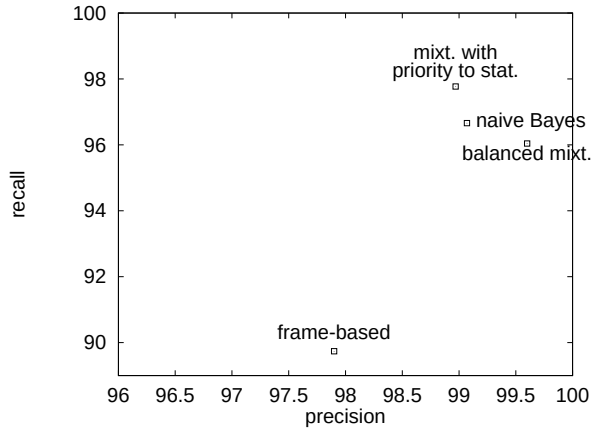
[9] bootstrap estimate of the variance

[10] at 95% it appears significant

4. Discussion

The average results are summarized in the following table and figure:

	precision (std deviation)	recall (std deviation)
statistical	99.07% (0.21)	96.66% (0.45)
balanced mixture	**99.60%** (0.18)	96.04% (0.69)
mixture w.p. stat.	98.97% (0.22)	**97.77%** (0.32)



Although the texts considered are very short (37 words in average), the results obtained with the naive Bayes classifier for newspaper advertisements are very high for both precision and recall when compared to other aplications referenced in the literature (Yang, 1999).

Concerning the information extraction based classification, results are significantly worse than those obtained with the naive Bayes classifier. As expected, Information Extraction based classification does not performe very well alone. It should only be used as a complement to other classification techniques in order to provide additional improvement.

Using the balanced mixed classification significantly improves the precision score while preserving the recall (not significant difference at 99%). This is due to the impact of advertisements that were well classified with the naive Bayes classifier but become unclassified with the addition of the information extraction based classifier. It is important to notice that the improvement of precision is good for systems where the accuracy of classification is more important than exhaustivity, as it is the case when questioning very redundant text collection such as the Internet.

In the case of mixed classification with priority to the statistical method, the recall score is (at 99%) significantly better than the one obtained with the naive Bayes classifier, while no significant difference (at 99%) is observed for the precision score. This classification method should then be used when exhaustivity is important, i.e. when it is important to retrieve all (or almost all) the relevant documents even if more unrelevant documents have then to be processed.

Finally, it should be mentioned that in the set of 12 advertisements that remained misclassified with the balanced classifier, 4 appeared to be really ambiguous (i.e. ambiguous even for human experts) and 1 was out of scope (i.e. corresponding to none of the categories). This gives an even better perspective to the results obtained by such a fully automated system.

References

Aragüés R., Chappelier J.-C., and Rajman M. (2000). Automated information extraction out of newspaper personal advertisements. In *submitted to Applied Natural Language Processing (ANLP'2000)*.

Hoyle W. (1973). Automatic indexing and generation of classification systems by algorithm. *Information Storage and retrieval*, 9(4):233–242.

Joachims T. (1997). A probabilistic analysis of the Rocchio algorithm with TFIDF for text categorization. In *Proceedings of International Conference on Machine Learning (ICML)*.

McCallum A. K. (1996). Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering. `http://www.cs.cmu.edu/~mccallum/bow`.

Mitchell T. M. (1997). *Machine learning*. McGraw-Hill.

Riloff E. and Lehnert W. (1994). Information extraction as a basis for high-precision text classification. *ACM Transactions on Information Systems*, 12(3):296–333.

Yang Y. (1999). An evaluation of statistical approaches to text categorization. *Information Retrieval Journal*.