A Hybrid Approach to Optimize Feature Selection Process in Text Classification

Roberto Basili, Alessandro Moschitti, and Maria Teresa Pazienza

University of Rome Tor Vergata
Department of Computer Science, Systems and Production
00133 Roma (Italy)
{basili,moschitti,pazienza}@info.uniroma2.it

Abstract. Feature selection and weighting are the primary activity of every learning algorithm for text classification. Traditionally these tasks are carried out individually in two distinct phases: the first is the global feature selection during a corpus pre-processing and the second is the application of the feature weighting model. This means that two (or several) different techniques are used to optimize the performances even if a single algorithm may have more chances to operate the right choices. When the complete feature set is available, the classifier learning algorithm can better relate to the suitable representation level the different complex features like linguistic ones (e.g. syntactic categories associated to words in the training material or terminological expressions). In [3] it has been suggested that classifiers based on generalized Rocchio formula can be used to weight features in category profiles in order to exploit the selectivity of linguistic information techniques in text classification. In this paper, a systematic study aimed to understand the role of Rocchio formula in selection and weighting of linguistic features will be described.

1 Natural Language Processing and Text Classification

Linguistic content in Text Classification (TC) aims to define specific and selective *features* with respect to training and test documents. Previous works on NLP-driven text classification (e.g. [1]) suggest that word information (e.g. morphology and syntactic role) improve performances. In particular, lemmatization and POS tagging provide a linguistically principled way to compress the features set (usually obtained by traditional crude methods like stop lists or statistical thresholds, e.g. χ^2). Statistical unsupervised terminological extraction has been also applied to TC training [2]. It allows to detect more complex and relevant features, i.e. complex nominal groups typical of the different target classes. The results are improved TC performances, although the contribution given by such modules has not yet been accurately measured. When more complex features (e.g. words and their POS tag or terminological units) are captured it is more difficult to select the relevant ones among the set of all features. Data sparseness effects (e.g. the lower frequency of n-grams wrt simple words) interact with wrong recognitions (e.g. errors in POS assignment) and the overall information

F. Esposito (Ed.): AI*IA 2001, LNAI 2175, pp. 320–326, 2001.

may not be enough effective. The traditional solution is the *feature selection*, discussed for example in [7]. By applying statistical methods, (information gain, χ^2, mutual information ...), non relevant features are removed. The Major drawback is that features irrelevant for a class can be removed even if they are important for another one. The crucial issue here is how to give the right weight to a given feature in different classes. This is even more important when NLP (and, mainly, terminology recognition) is applied: some technical terms can be perfectly valid features for a class and, at the same time, totally irrelevant or misleading for others.

In this paper a systematic study aimed to understand the role of Rocchio formula in the selection and weighting applied to standard and linguistic features (e.g. terminological expressions), will be described.

2 A Hybrid Feature Selection Model

2.1 The Problem of Feature Selection

Automatic feature selection methods foresee the removal of noninformative terms according to corpus statistics (e.g. *information gain*, *mutual information* and χ^2), and the construction of new (i.e. reduced or re-mapped) feature space. A distinctive characteristic is the selection of features based on their relevance in the whole corpus instead of in a single category. Moreover, in [6], feature selection appears as a distinct phase in building text classifier. In order to account for differences in the distribution of relevance throughout classes, we should depart from the idea of a unique ranking of all corpus features. Features should be selected with respect to a single category. This can lead to retain features only when they are truly informative for some classes. In next section an extension of the Rocchio formula aiming to obtain feature weights that are also, at the same time, optimal selectors for a given class is presented.

2.2 Generalizing Rocchio Formula for Selecting Features

The Rocchio's formula has been successfully used for building profile of text classifier as follows. Given the set of training documents R_i classified under the topics C_i, the set $\bar{R}_i$ of the documents not belonging to C_i, and given a document h and a feature f, the weight Ω_f^h of f in the profile of C_i is:

$$\Omega_f^i = \max\left\{0, \frac{\beta}{|R_i|}\sum_{h\in R_i}\omega_f^h - \frac{\gamma}{|\bar{R}_i|}\sum_{h\in \bar{R}_i}\omega_f^h\right\} \quad (1)$$

where ω_f^h represent the weights of features in documents[1]. In Eq. 1 the parameters β and γ control the relative impact of positive and negative examples and determine the weight of f in the i-th profile. In [4], Equation 1 has been firstly

[1] Several methods are used to assign weights of a feature, as widely discussed in [5]

used with values $\beta = 16$ and $\gamma = 4$: the task was categorization of low quality images. In the next section a procedures for selecting an optimal γ, keeping fixed β to 1 value, will be presented.

Selecting Features via Rocchio's Formula Parameters The relevance of a feature deeply depends on the corpus characteristic and, in particular, on the differences among the training material for the different classes, i.e. size, the structure of topics, the style of documents, ... This varies very much across text collections and across the different classes within the same collection. Notice that, in Equation 1, features with negative difference between positive and negative relevance are set to 0. This implies a discontinuous behavior of the Ω_f^i values around the 0. This aspect is crucial since the 0-valued features are irrelevant in the similarity estimation (i.e. they give a null contribution to the scalar product). This form of selection is rather smooth and allows to retain features that are selective only for some of the target classes. As a result, features are optimally used as they influence the similarity estimation for all and only the classes for which they are selective.

As feature weights relies on the γ and β setting, fitting them with respect to the classification performance has two main objectives:

- First, noise is drastically reduced without direct feature selection (i.e. without removing any feature).
- Second, the obtained ranking provides scores that can be directly used as weights in the associated feature space.

Notice that each category has its own set of relevant and irrelevant features and Eq. 1 depends for each class i on γ and β. Now we assume the optimal values of these two parameters can be obtained by estimating their impact on the classification performance, independently for each class i. This will result in a vector of (γ_i,β_i) couples each one optimizing the performance of the classifier over the i-th class. Hereafter we will refer to this model as the $Rocchio_{\gamma_i}$ classifier. Finally, it has to be noticed that combined estimation of the two parameters is not required. For each class, we fixed one parameter (β_i indeed) and let γ_i vary until the optimal performance is reached. The weighting, ranking and selection scheme used for $Rocchio_{\gamma_i}$ classifier is thus the following:

$$\Omega_f^i = \max\left\{0, \frac{1}{|R_i|}\sum_{h\in R_i}\omega_f^h - \frac{\gamma_i}{|\bar{R}_i|}\sum_{h\in \bar{R}_i}\omega_f^h\right\} \tag{2}$$

In our experiments, β has been set to 1, Equation 2 has been applied given the parameters γ_i that for each class C_i lead to the maximum breakeven point[2] over a test set. By using this formula, when γ_i is increased only features very representative for the target class i assume a relevant weights. Alternatively features

[2] It is the threshold values for which precision and recall coincide (see [6] for more details).

that are moderately informative for a category i but that are at same time poor relevant for all other categories may be heavily weighted. In this perspective γ_i parameter acts as a domain concept selector. In next sections the above characteristic of γ_i is verified over (possibly complex) features extracted by Natural Language Processing techniques. As such features are more representative than simple words the effectiveness of $Rocchio_{\gamma_i}$ selection can be emphasized and measured.

2.3 The Role of NLP in Feature Extraction

Main objective of this section is to describe the role of linguistic information in the representation of different classes in a TC task. We underline that these latter are often characterized by sets of *typical* concepts usually expressed by multiwords expressions, i.e. linguistic structures synthesizing widely accepted definitions (e.g. "*bond issues*" in topics like "*Finance* or *Stock Exchange*"). These sets provide useful information to capture semantic aspects of a *topics*. The multiword expressions are at least in two general classes useful for TC: Proper Nouns (PN) (e.g. like locations, persons or artifacts) and Terminological expressions, which are more relevant triggers than PN for the classification decisions. Their detection results in a more precise set of features to be included in the target vector space. The identification of linguistically motivated terminological structures usually requires external resources (thesaura or glossaries): as extensive repositories are costly to be developed and simply missing in most domains, an enumerative approach cannot be fully applied. Automatic methods for the derivation of terminological information from texts can thus play a key role in content sensitive text classification.

Previous works in the NLP research area suggest that the semantic classes related to terms depend strongly on the underlying domain. As terms embody domain specific knowledge we expect that their derivation from a specialized corpus can support the matching of features useful for text classification. Once terms specific to a given topics C_i are available (and they can be estimated from the training material for C_i), their matching in future texts d should strongly suggest classification of d in C_i. In this work, the terminology extractor described in [2] has been adopted in the training phase. Each class (considered as a separate corpus) gives rise to a set of terms, T_i. When available, elements in T_i can be matched in future test documents. They are thus included in the final set of features of the target classifier. Other features provided by linguistic processing capabilities are lemmas and their associated POS information able to capture word syntactic roles (e.g. *adjective*, *verb*, *noun*). Those irrelevant features, that are not necessarily produced via complex linguistic processing (e.g. single words), are correctly smoothed by Eq. 2 and this also helps in a more precise weight of the NLP contribution. This results in a hybrid feature selection model where grammatical and statistical information are nicely investigated.

3 Experimenting Hybrid Feature Selection in Text Classification

In these experiments the Equation 2 is applied to several sets of linguistically derived features. Reuters, version 3, corpus prepared by Apté [6] has been used as reference corpus. It includes 11,099 documents for 93 classes, with a fixed splitting between test (TS) and learning data (3,309 vs. 7,789). The linguistic features described in Section 2.3 have been added to the standard set. They consist of:

- Proper Nouns: +PN indicates that the recognized proper nouns are used as features for the classifiers; -PN is used instead to indicate that proper nouns recognized in texts are removed from the set of valid features during training
- Terminological Expressions (+TE)
- Lemmas (-POS) and Lemmas augmented with their POS tags (+POS)

In Table 1 is reported the BEP of the three feature sets: the comparison is against the baseline, i.e. the best non linguistic result. Note that for each feature set, indicated in the table, re-estimation of the γ_i parameters has been carried out. The above table shows the overall positive impact (by microaveraging the

Table 1. Breakeven points of $Rocchio_{\gamma_i}$ on three feature set provides with NLP applied to Reuters version 3.

Base-Line	+POS-PN	+PN+TE	+PN+TE+POS
83.82%	83.86%	84.48%	85.13%

BEP of all 93 categories) of using diverse NLP capabilities. However, individual analysis for each category is required for a better understanding of the selectivity of γ_i. Figure 1 shows the performance (BEP) of some classes wrt the adopted γ for profile learning. Results show that $Rocchio_{\gamma_i}$ weighting scheme proposes as a robust filtering technique for sparse data in the training corpus. It is to be noticed that the behavior of the γ_i parameters is tightly related to the categories (i.e. to the training material available for them). Different categories show quite different values of γ_i able to optimize performances. This seems related to the inner conceptual nature of the categories themselves. A suitable setting represents thus a promising model to select relevant features, that well reflect the semantic role of each of them. For this reason, we applied the best obtained settings to weight linguistic features in profiles and carried on a contrastive analysis against the baseline figures derived by only using standard features.

A second major outcome, is that the comparative evaluation of simpler against linguistically motivated feature sets confirm the superiority of the latter. The $Rocchio_{\gamma_i}$ applied to linguistic material supports thus a computationally efficient classification. This is mainly due to the adoption of the optimal selection and weighting method proposed in Equation 2, which optimizes features

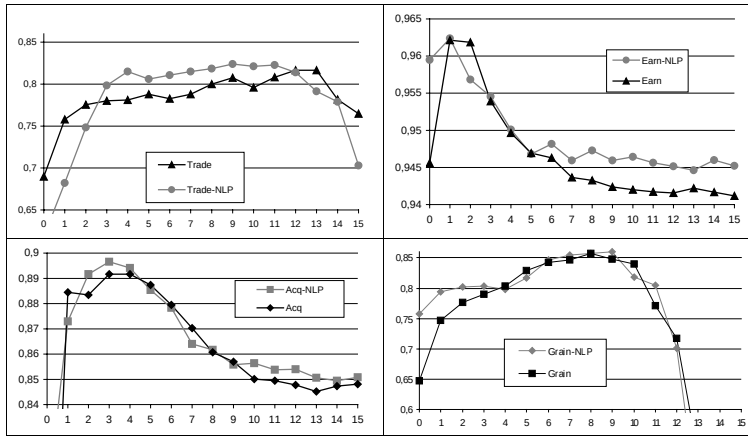

Fig. 1. Break-even point performances of the Rocchio classifier according to different γ values for some classes of Reuters 3 Corpus

vs. performances (see Figure 1). It seems that a suitable parameter setting for the γ_i provides a systematic way to filter (and emphasize) the source linguistic information. It has to be noticed that in the experiments we obtained a source set of 9,650 features for the Reuters 3 *acq* category. After γ_{acq} setting, only 4,957 features are assigned with a weight greater than 0.

Figure 1 also shows that the NLP plots (for the selected classes i) have values systematically higher than plots of tests over standard features. This is true for each γ_i value. Notice that if a non optimal γ is chosen for a given class, it is possible that classifiers trained with standard feature outperform those trained with NLP, as evidently data sparseness in linguistic data creates dangerous noise. This is the mainly reason for previous failures in the adoption of Rocchio weighting. The results show how setting of suitable γ_i values are critical for the optimal use of linguistic information. The suggested model (Eq. 2) thus represents an efficient approach to hybrid feature selection in operational (large scale) TC systems.

References

[1] R. Basili, A. Moschitti, and M.T. Pazienza. Language sensitive text classification. In *Proceeding of 6th RIAO Conference, Collège de France, Paris, France*, 2000.

[2] R. Basili, A. Moschitti, and M.T. Pazienza. Modeling terminological information in text classification. In *Proceeding of 7th TALN Conference*, 2000.

[3] R. Basili, A. Moschitti, and M.T. Pazienza. NLP-driven IR: Evaluating performances over text classification task. In *Proceeding of the 10th IJCAI Conference, Seattle, Washington, USA*, 2001.

[4] David J. Ittner, David D. Lewis, and David D. Ahn. Text categorization of low quality images. In *Proceedings of SDAIR-95*, pages 301–315, Las Vegas, US, 1995.

[5] G: Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. *Information Processing and Management*, 24(5):513–523, 1988.
[6] Y. Yang. An evaluation of statistical approaches to text categorization. *Information Retrieval Journal*, May, 1999.
[7] Y. Yang and Jan O. Pedersen. A comparative study on feature selection in text categorization. In *Proceedings of ICML-97*, pages 412–420, Nashville, US, 1997.