A robust summarization system to explain document categorization

Alessandro Moschitti and **Fabio Massimo Zanzotto**
University of Rome Tor Vergata,
Department of Computer Science, Systems and Production,
00133 Roma (Italy),
{moschitti,zanzotto}@info.uniroma2.it

Abstract

The automatic delivery of textual information to interested users is often based on the notion of text categories. The approach generally adopted by news providers consists of categorizing news items in predefined classification schemes (e.g. *Sport*, *Politics*, *Economics*, etc.) and, then selectively delivering information to interested consumers. Robust automatic models successfully proposed for the text classification task use a bag-of-word abstraction of the documents to achieve good performances. This latter guarantees the applicability over very different corpora and the robust machine learning algorithms guarantee a high level of performances. The inherent limitation of this approach is given by the fact that the users perception of the adopted categories may differ from the one imposed. The robustness of the bag-of-word model may not be sufficient to coherently serve users' information needs. In this paper, we want to investigate a class-based model devoted to explain the categorization decision and the classification scheme to the final user. We propose the use of indicative and informative summaries as an explanation of the categorizer decision for the target document. The summaries are produced using the explicit class profile. This latter contains simple terms (i.e. words) as well as complex nominals and coarse event representations. The proposed information access paradigm is based on a robust machine learning approach to text classification that extracts explicit class descriptors (i.e. profiles) and a robust syntactic parser used to produce document descriptions. Specific experiments over a medical corpus have been settled to evaluate the impact of the document explanation model on the users' comprehension of the categorization process.

1 Introduction

The automatic delivery of textual information to interested users is often based on the notion of text categories: firstly, news providers tag news items according to a predefined classification scheme (e.g. general class scheme: *Sport*, *Politics*, *Economics*, etc.) and, secondly, deliver them to the interested information consumers. The more fine-grained is the classification structure the more specific information needs can be satisfied. However, a high granular classification has a first inherent limitation: the providers and consumers may have a different understanding of the classification scheme. A second limitation is a consequence of the first: it may be unclear why a given document, automatically tagged, should be relevant for the final user.

Some robust automatic approaches have been successfully proposed for the text classification (TC) task. A definition of *robustness* has been given in (Basili and Zanzotto, 2002): it is the capability of a system to show similar performance over different corpora. In this perspective, the GRC system proposed in (Basili and Moschitti, 2001) as well as other TC models (e.g SVM in (Joachims, 1998)) are robust since they can generalize well even if exposed to an high number of irrelevant features. The bag-of-word abstraction of the documents (obtained via a simple tokenization phase without selection) is sufficient to achieve good performances because features (i.e. words) are implicitly pruned out by the classification model if not relevant for its learning algorithm. The bag-of-word abstraction guarantees the applicability over very different corpora and the robust machine learning algorithms guarantee a high level of performances. However, the bag-of-word model may not be sufficient to coherently serve users' information needs.

In order to improve the user satisfaction in the interaction with the providers, two aspects have to be taken into account: (1) the categorization scheme and (2) the explanation of the class assignment. Once the categorization scheme adopted by the news provider is clear to the final users, these may better express their needs. Secondly, a concise document explanation may help the user to understand if the document is interesting for him. Moreover, the above explanation may help to recover authentic misclassifications of an automatic classification system. The user can better decide to thrust the system and read the news item, or, conversely, discard it. This may not be possible if he is exposed

Title: Combined laminoplasty and posterolateral fusion for spinal canal surgery in children and adolescents.
Spinal deformities, especially kyphosis and instability, after laminectomy for **tumors** and other diseases, are major clinical problems. Since 1981, combined laminoplasty and posterolateral fusion for the prevention of postlaminectomy spinal deformities was performed on eight male and two female patients aged two to 26 years (average, 13.9 years). The follow-up period was from six months to seven years and three months (average, three years and five months). Two patients died six and ten months postoperatively because of **brain metastases** (astrocytoma) and **lung metastases** (neuroblastoma), respectively. Good alignment with no instability of the cervical or thoracic spine was obtained for all patients, including the two who died. Laminoplasty combined with posterolateral fusion was found to be very effective in preventing the development of spinal deformities after spinal canal surgery for spinal **cord tumors** or other diseases in children and adolescents.

Table 1: Ohsumed sample news item

only to the document category and to the title of the current actual news. For instance, given the title:

Combined laminoplasty and posterolateral fusion for spinal canal surgery in children and adolescents.

it is not clear why the document of the medical domain in Tab. 1 should be related to the *Neoplasms* class of Medical Subject Headings (MeSH[1]). If the user is provided also with an indicative summary represented by the complex nominals such as *tumors*, *metastases*, *lung metastases*, *brain metastases*, and *cord tumors*, he may better understand if this incoming document is related to the above class. Furthermore, this perception may be improved if an informative summary is presented. This latter is built using the sentences that contain the above concepts. Note that, in order to serve the purpose, this enriched information should be strictly related to the document class. The proposed summaries provide a sort of class-based explanation for the document content. It is worth noticing that such kind of information is already produced as side effect by the automatic categorization task although it is generally neglected.

In this paper, we want to investigate then two aspects: (1) a model for the explanation of the classification scheme; (2) a class-based document enriching that better explains the document category mapping. The proposed information access paradigm is based on a robust machine learning approach to text classification (Basili and Moschitti, 2001) that extracts explicit class descriptors (i.e. profiles). It is applied to a feature space richer than the ones generally used in IR systems. For this purpose a robust syntactic parser (Basili and Zanzotto, 2002) is used to produce the document descriptions. In order to introduce our class explanation model (Sec. 4), we will firstly describe the enriched feature space introduced for a better document representation (Sec. 2) and the document categorization model (Sec. 3). Finally we will describe preliminary experiments performed over a medical corpus that have been settled to evaluate the impact of the document explanation model on the users' comprehension of the classified documents.

[1]A complete description can be found in `http://www.nlm.nih.gov/mesh`

2 Representing documents for enriched categorization

The text classifiers based on a Vector Space Model (VSM) represent document as points in the space where generally words are adopted as dimensions. The knowledge bases that machine learning algorithms for text classification may produce strongly depend on the underlying VSM. These KBs affect the document explanation model and the class explicit representation we want to settle. Therefore, only if a rich text representation is settled, they may result expressive. Generally, available efficient machine learning algorithms for text categorization are applied over a document Vector Space Model (VSM) which is built upon the notion of words (or stems). This latter seems to be sufficient for achieving good results for the classification task itself. The use of more complex information does not improve dramatically the document categorization accuracy. However, such a VSM may not result sufficient for settling an explicative indicative document summary nor clear class descriptions.

More formally, the classification problem may be seen as a decision function f that maps documents ($d \in D$) into one or more classes, i.e. $f : D \rightarrow 2^C$, once a set C of classes is given. The function f is usually built according to an extensive collection of examples classified into classes C_i, often called *training set*. In the vector space model (Salton and Buckley, 1988), each document in D is seen as a set of couples <*feature,weight*> organized into vectors. In order to set this document representation, three parameters have to be settled: which are the document indexes, i.e. the content bearing *features* (generally the stemmed words, after stopword removal), that should be extracted from the documents; how these indexed features are *weighted* to enhance re-

trieval of the documents relevant to the user; and, finally, how the similarity function between documents and other informative structures (e.g. user profiles, categories, queries) is defined for modeling the conceptual relatedness.

With the purpose of modeling our document vector representation, we need, therefore, to define a few specific parameters. Given a training set, a feature $f \in \{f_1, ..., f_n\}$ to describe it, a generic document d of the corpus, let the following notations express:

- M, the number of documents in the training set,
- n_f, the number of documents in which the features f appears and
- o_f^d, the occurrences of the feature f in the document d (TF of feature f in document d).

We will use the ω_f^d weight[2] as follow:

$$\omega_f^d = \frac{l_f^d \cdot IDF(f)}{\sqrt{\sum_{r=1}^{n}(l_r^d \cdot IDF(r))^2}} \quad (1)$$

where the $IDF(x)$ is $log(\frac{M}{n_x})$ and

$$l_f^d = \begin{cases} 0 & \text{if } o_f^d = 0 \\ log(o_f^d) + 1 & \text{otherwise} \end{cases} \quad (2)$$

It is clear that stemmed words strongly limit the expressiveness of both the possible document explanations (indicative summaries) and the eventual class descriptions (i.e. the category profiles). In the next sections, an enriched set of document features is therefore introduced.

2.1 Extending the word-based document representation

The VSM based on simple words lacks in expressiveness. In fact, words, considered independent, provide only singleton surface forms. These latter are only a small part of the key concepts expressed in the documents and, moreover, are generally polysemic, i.e. denote more than one concept. The consequence is a very poor representation from the user point of view. A large part of relevant concepts in the domain is expressed by collocations of more than one word (i.e. complex terms like *interim dividend* in the financial domain). These collocations have also the beneficial property of denoting generally only one concept. This is also considered valid in the terminology extraction approaches (Jacquemin, 2001): n-grams following specific syntactic prototypes as, for example, the *risk factor* meeting the `Noun Noun` syntactical constraint or *interim dividend* following the `Adjective Noun` constraint are certainly less polysemic than the isolated compounding words, i.e. *risk*, *interim*, *factor*, and *dividend*. Therefore, a document and a class descriptor based on such kind of complex concept denotations can be more informative since less ambiguous than words (or stems).

The above analysis assumes that important concepts are denoted by nouns or nominal compounds such as *equation of fluidomechanics* whilst the verbal information is neglected. However, specific relations between (simple) concepts are often expressed by verb-governed surface forms such as *companies buy shares*. However, even if this information seems to be crucial "as it is" for the description of the class, due to the fact that the verb arguments may be very distant and in relatively free order, an approximated version can be used in the vector space model. The document class explanation and the concise class description we want to produce is thus based on:

- concepts expressed with simple surface forms, i.e. words;
- concepts expressed with complex surface forms, i.e. complex terms;
- simple relations between concepts based on verbal contexts;

They can be sufficient to understand the bunch of documents forming a class and to judge the document-class mapping. However, to support the discovery of such an explicit description of the class knowledge, both a suitable vector space model has to be defined and tools for extracting such an information have to be defined. Simple techniques based on barrier words are not sufficient. These approaches show their limits if applied to long distance dependencies such as the verb argumental relation.

2.2 Mapping documents in the systematic vector space

The extraction/selection of these surface forms from the texts needs a suitable language driven model. In the terminology extraction techniques (Jacquemin, 2001), a syntactic model of the textual phenomena is generally used. We will here rely on the extended dependency-based representation formalism (XDG, (Basili et al., 2000)). An XDG is a graph whose nodes are constituents and whose arcs are the syntactic relations among constituents. The constituents taken in consideration are generally *chunks* (Abney, 1996), i.e. non-recursive kernels of noun phrases (`NPK`), prepositional phrases (`PPK`) and verb phrases (`VPK`) such as *five patients*, *by non invasive methods*, *were evaluated* respectively. On the other hand, arcs make explicit the syntactic relations between chunks, i.e. the inter-chunks relations such as verb-subject, verb-object, verb-modifier, and noun-prepositional modifiers. In Fig. 1, a sample XDG

[2] It is the *ltc* in (Salton, 1991)

is depicted: chunks are the words between square brackets (i.e. VPK,NPK,PPK) while inter-chunk dependencies are depicted as arrows (i.e. subj for the subject relation, V_PP for the verb prepositional modifier relation and NP_PP for the noun-prepositional modifier relation). It is worth noticing that the chunk layer is build on a part-of-speech tagged text. Under this representation, the surface

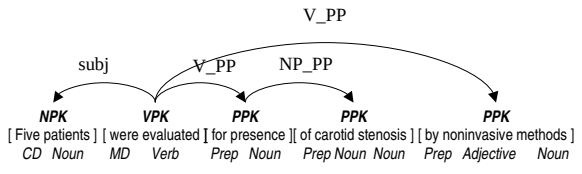

Figure 1: Example XDG

form prototypes for the complex concepts can be detected via `NPK PPK*` regular expressions on the XDG node sequence. The selected sequence is accepted if the portion of the graph is totally connected. This limits the generative power of the previous regular expression. It is worth noticing that not all the PPK are accepted (i.e. `PPK`s containing pronouns are refused) All the accepted subsequences `(Noun|Adjective)* Noun` of an `NPK` are selected as representative of concepts represented in the document and then added as features. On the other hand, relations among concepts are extracted by verb-dependencies. Verb argument couples are more informative for describing the target class. For example, *(buy,(dirobj,'share'))* or *(complete,(dirobj,'acquisition'))* in an economic corpus suggest that the text collection refers to (the changes of) company assets. The same information is used in some natural language information retrieval engine like in (Strzalkowski et al., 1998). The adoption of robust syntactic parsing techniques based on processing module cascades (Basili et al., 2000) makes possible the selection of the above surface forms on a large scale. The parser includes a tokeniser, a part-of-speech tagger, a chunker, and a shallow syntactic analyser.

3 Effective domain modeling in Rocchio text classifier

Profile-based are traditionally adopted for computationally efficient text classification. They are characterized by a function f based on a similarity measure between the synthetic representation of each class C_i and the incoming document d. Both representations are vectors and similarity is traditionally estimated as the cosine angle between the two. The description $\vec{C}_i$ of each target class (C_i) is usually called *profile*, that is a vector summarizing all training documents d such as $d \in C_i$. Vector components are called *features* and they refer to independent dimensions in the space in which similarity is estimated. The i-th component of a vector representing a given document d is a numerical weight associated to the i-the feature w that occurs in d. Similarly, profiles are derived from the grouping of positive instances d in class C_i, i.e. $d \in C_i$.

A newly incoming document is thus considered member for a given class *iff* the similarity estimation overcomes established thresholds. The latter are parameters that adjust the trade-off between precision and recall.

3.1 Selecting relevant features from categories

Feature Selection techniques (Yang and Pedersen, 1997) have been introduced in order to select the most informative features according to training corpus statistics, and reduce (via pruning or remapping) the size of the feature space.

They are based on probability distributions of features in training data and modeled according to statistical selectors (e.g. χ^2, mutual information, information gain,..). Ranking of features is carried out and selection is applied by removing the lower features in ranking (i.e. thresholding). Notice how ranking is uniform throughout categories: it may emphasize only features f which are globally informative by pruning those f relevant only for a given (or few) class(es). The selection does not depend on differences in relevance among classes. More generic classes (e.g. with low χ^2 values) may result in a very poor profile, i.e. fewer features. A side effect of this choice is to equally distribute relevance of features over classes.

In order to account for differences in the distribution of relevance throughout classes, we should depart from the idea of a uniform ranking. Features should be selected only via comparative analysis of evidences related to a single category. This can lead to discard features *only* when they are truly not informative for all classes. The result is a suitable class-based ranking, so that the obtained feature scores can be straightforwardly assumed as features weights in the class profile (Basili et al., 2001).

In next section an extension of the Rocchio formula, the Generalized Rocchio Classifier (GRC) aiming to obtain feature weights that are, at the same time, optimal selectors for a given class is presented.

3.2 Generalized Rocchio formula

The generalized Rocchio formula has been employed for weighting linguistic features of category profiles (Basili et al., 2001) as follows. Given the set of training documents R_i classified under the topics C_i, the set $\bar{R}_i$ of training documents classified in $C_j \neq C_i$, a

document d and a feature f, the weight Ω_f^i assumed by f in the profile of C_i is:

$$\Omega_f^i = \max\left\{0, \frac{1}{|R_i|}\sum_{d\in R_i}\omega_f^d - \frac{\gamma_i}{|\bar{R}_i|}\sum_{d\in \bar{R}_i}\omega_f^d\right\} \quad (3)$$

where ω_f^d represents the weights of features f in documents d. In Eq. 3 the parameters γ_i control the relative impact of negative examples and determine the weight of f in the i-th profile. Features with negative difference between positive and negative relevance are set to 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).In this way, the minimal set of truly irrelevant features (giving 0 values for all the classes) can be better captured and removed. Therefore, the γ_i setting, fitted with respect to the classification performance, has three main objectives:

- First, noise is drastically reduced without direct feature selection (i.e. without removing any feature).
- Second, the ranking based on Ω_f^i scores depends on i and can be directly used as weights for f in the corresponding feature space.
- Linguistic features as well as the other ones receive a weight proportional to their contribution in the classification accuracy. Note that as a parameter γ_i for each category is provided, a linguistic feature may assume different weights in different categories. This defines the best suitable set of concepts (i.e. the features with higher weights) for the target category.

In (Basili and Moschitti, 2001) an approach for selecting the optimal γ_i parameters has been proposed. This methodology uses a set of training documents for profile building and a second different subset (the *estimation* set) for finding parameters which optimize the TC accuracy.

4 The explanation system

Once a rich document and domain representation have been defined the classification/explanation system can be easily modeled. The vector representing an incoming document includes the informative structures defined in Section 2.1 together their weights (Sections 2 and 3). Several strategies can be applied to select the more important document concepts related to a target domain. Our aim is to provide two summaries as explanation of the current document categorization: one *indicative* and one *informative*. These summaries will show important concepts inside the document with respect to the target category. For this purpose, the set of document features that are contained in the profile of target category are considered. Then, the feature f are ranked according to $\omega_f^d \cdot \Omega_f^i$, i.e. the product between the weights of f in document and in profile vectors.

The indicative summary of the document d is defined as the set $R_k(d)$ of k top ranked features (*k-best features*). As the document features even contain complex terms the resulting set of summary concepts may be more descriptive than those based on words only. We will call such an explanation as the *summary based on best features* (S_{bf}).

The informative summary should contain the more meaningful paragraphs (*m-best paragraphs*). The paragraphs that contain some of the best k features are ranked according to ω_p^i weight defined in the following.

Given a paragraph p in a document d, the set of the best k paragraph features can be defined as $S_k(p,d) = \{f : f \in p, f \in R_k(d)\}$, where f is a feature in p. The paragraph weight is then defined as follows:

$$\omega_p^i = \sum_{f\in S_k(p,d)} \omega_f^d \cdot \Omega_f^i, \quad (4)$$

where p is a paragraph of d.

The *informative summary based on the best paragraphs* (S_{bp}) is obtained by picking-up the top m paragraphs ranked according to Eq. 4. The parameter m establishes the rate of the document paragraphs shown as explanation.

A base-line version of the proposed explanation model can be obtained by replacing the $\omega_f^d \cdot \Omega_f^i$ weights with the simpler *document frequency* (i.e. the number of category documents in which the features f appears). Hereafter we will refer to these simpler explanation models as the *frequency summary based on features* (S_{ff}) and *frequency summary based on paragraphs* (S_{fp}) . In next section the above explanation models will be evaluated contrastively.

5 Experiments

For our experiments we adopted the Ohsumed corpus[3]. From the 50,216 medical abstracts of the year 91, the first 20,000 belonging to one of the 23 *MeSH diseases* categories have been used.

The classifier accuracy is provided by means of *f-measure* with equal importance assigned to recall and precision. Cross validation has been applied for measuring the system performances. 30 splits between training and test-set (about 70% and 30%) have been carried out. The global performance of a systems is given by averaging the global classifier

[3]It has been compiled by William Hersh and it is currently available at `ftp://medir.ohsu.edu/pub/ohsumed`

Table 2: Linguistic contribution to the Generalized Rocchio classifier performances on Ohsumed corpus

	Tokens	Ling. Feat.
Category	f_1	f_1
Pathological Conditions	48.78	49.36
Cardiovascular Diseases	77.61	77.48
Neoplasms	71.34	79.38
Hemic & Lymphatic Dis.	65.80	65.93
Neonatal Dis.& Abnormal.	50.05	52.83
Skin & Connect. Tissue Dis.	60.38	60.53
Microaverage	65.81	65.90

performance[4] over all 30 splits. As Table 2 shows, some of the classes are positively affected by the more structured VSM. However, the overall classifier performances are not significantly improved by these complex structures as also previously pointed out for IR models (Strzalkowski et al., 1998). Category profiles seem well subsumed by the *"bag of words"* models (i.e. Tokens).

Even if the overall classification performances are not dramatically affected, the proposed VSM allows the extraction of more informative class description since complex terms receive a high ranking weight in the class profile. To obtain a class description, it is sufficient to rank profile features according to the their weights (i.e. Ω_f^i of Equation 3). The column 1 of the Table 3 shows the top 31 complex terms of *Cardiovascular Disease* category. The features seem to be conceptually closer to the target domain. In column 2 the complex terms ordered by frequency inside the category are shown. We observe that some non-relevant features as well as non specific terms, i.e *normal subject*, *control subject*, *risk factor*, *side effect*, *appo patients* and so on have reached the top of ranking positions. As suggested in (Daille, 1994), frequency seems to be a good indicator of domain relevance, however cross-class techniques, as the one proposed, eliminates the unspecific and useless terms.

5.1 Evaluation of the different summaries

The aim of these experiments is to measure the effectiveness of our explanation methods. This objective can be achieved in several ways. As our purpose is to design a document filtering system based on users' information needs we have implemented a specific experimental procedure to test the user satisfaction. A randomly generated set of about 200 documents (UTS) has been selected from the classified test-set.

[4]We adopt the *microaveraging* (Yang, 1999). It is the mean over all decision taken by the systems. This includes the decision over all categories.

GRF	*Frequency*
myocardial infarction	myocardial infarction
coronary angioplasty	coronary artery
coronary artery	risk factor
essential hypertension	coronary angioplasty
acute myocardial infarction	congestive heart failure
congestive heart failure	acute myocardial infarction
myocardial ischemia	pulmonary hypertension
hypertensive patients	essential hypertension
ventricular function	myocardial ischemia
arterial pressure	ventricular tachycardia
ventricular tachycardia	arterial pressure
pulmonary hypertension	hypertensive rat
hypertensive rat	ventricular function
cardiovascular disease	hypertensive patients
coronary angiography	vascular resistance
cardiac catheterization	cardiac arrest
atrial fibrillation	atrial fibrillation
cardiac arrest	appo patients
cardiac output	cardiac output
thrombolytic therapy	control subject
mitral regurgitation	significant difference
hypertrophic cardiomyopathy	consecutive patients
vascular resistance	chest pain
angina pectoris	cardiac catheterization
antihypertensive agent	hypertrophic cardiomyopathy
doppler echocardiography	side effect
unstable angina	pulmonary artery
enzyme inhibitors	cardiovascular disease
atrial pressure	cardiac death
coronary disease	thrombolytic therapy
mitral stenosis	normal subject

Table 3: Complex term Ohsumed Cardiovascular disease class descriptor: GRC vs. simple frequency

The user has to evaluate if an incoming document d is correctly labeled in the category C, i.e. if d belongs to C according to his own perception of the classification scheme. Documents are presented to the users together to a category C that may or may not be the true category of the document according to the classification scheme (50% are correct). The user is asked to state its *acceptance*, or its *rejection* with respect to the shown class C. For each document $d \in UTS$, the user goes through 3 steps that make available different kinds of information:

- *Indicative summary*, made of the document title and the S_{bf} (set of best features) or S_{ff} (set of frequent feature) defined in Section 4.
- *Informative summary*, including the S_{bp} (set of the *best* paragraphs) or S_{fp} (set of the *frequent* paragraphs).
- *Full document* where the entire document is shown for final decision.

The S_{bp} is displayed after the user has been exposed to the S_{bf} while S_{fp} is shown after S_{ff}. This means that it was not possible to measure the S_{bp} and S_{fp} independently from the related indicative summaries. In case of a wrong category is proposed (with respect to the test information in Oshumed), the system always provides its best explanation. The user has thus no information about the correctness of the proposed class, so that he relies only on explanations.

The first (users 1,2,3, and 4) tested the machine learning explanation models (i.e. S_{bf} and S_{bp}); the other users tested the k-frequent explanation models (i.e. S_{ff} and S_{fp}). We define the explanation *score* as the user coherence with its own final decision. This is the number of matches between

Table 4: Evaluation of the class explanation model

User	Classifier	Ind. Summary	Inf. Summary
1	0.7450	0.8431	0.9019
2	0.5890	0.7945	0.8493
3	0.6557	0.7950	0.8852
4	0.6534	0.7326	0.8712
avg.	0.6608	0.7913	0.8769
5	0.7647	0.8921	0.9705
6	0.5791	0.6582	0.7861
7	0.7523	0.8012	0.8802
avg.	0.6987	0.7838	0.8789

the explanations-driven decisions and the final one (based on the entire document). In Table 4 the performances of 7 users are reported. Users have been divided in two groups. In columns *Ind. Summary* and *Inf. Summary*, the scores of the explanation models based on indicative and informative summaries are respectively reported. In *Classifier* column is reported the user satisfaction with respect to the category assigned by the classifier. In the *avg.* rows, the average of the corresponding user group is shown. Two main trends can be observed. First, the category assigned by the classifier seems to be the least satisfying, i.e. its agreement score (with the final user opinion) is the lowest. If the S_{bf} are added for explaining the category label provided the mean score increase of about 13% (79.13% vs. 66.08%). As the explanation model becomes richer, i.e. also the S_{bp} are provided, the user better appreciates the final document content. This reflects in a further increase of about 8% with respect to the feature based model. The overall improvement of the user satisfaction of the combined explanation model is around 21% (87.69% vs. 66.08%). The second aspect is that even the explanation models based on the simple frequency are helpful. In this case, the S_{ff} and S_{fp} improve the baseline of about, respectively, 9% and 18%. Therefore, adding explanatory information about the document category is always effective. However, the machine learning approach to feature selection proposed seems more promising as it better improve (+21%) the baseline explanation (i.e. category and title only) than the document frequency heuristic (+18%). It is worth noting that a direct comparison between the two explanation systems requires a larger set of users. Users show high variability in their decision revision process. A feasible solution to limit the variability could be testing the target users with both two explanation systems. However, this requires the doubling of the effort in testing process and it goes beyond the purpose of this preliminary experiments. A further advantage of the S_{bf} wrt S_{ff} is that the first actually selects and presents to the user only 4 features, on average, with respect to 8 shown by the second. In the machine learning summary approach the reader is exposed to less than half number of features when he has to take his decision. The compression of relevant information is mainly due to the selection technique of Section 3.

6 Conclusion

In this paper, a robust explanation system for text categorisation has been presented. This is based on two types of summaries that aim to improve the user satisfaction with respect to the delivered documents. The user simply reading the proposed summaries can decide (without reading the entire document) if the document meets his own interests. Both indicative and informative summaries are obtained by using a robust machine learning approach (GRC) together with a robust parser ($CHAOS$) to select the concepts and paragraphs related to the target category. A preliminary evaluation of our explanation model has been carried out by testing the users' satisfaction. The summary-based explanation seems to be a promising solution for the motivation of the automatic categorisation.

References

Steven Abney. 1996. Part-of-speech tagging and partial parsing. In G.Bloothooft K.Church, S.Young, editor, *Corpus-based methods in language and speech*. Kluwer academic publishers, Dordrecht.

R. Basili and A. Moschitti. 2001. A robust model for intelligent text classification. In *Proceedings of the thirteenth IEEE International Conference on Tools with Artificial Intelligence, November 7-9, 2001 Dallas, Texas.*

Roberto Basili and Fabio Massimo Zanzotto. 2002. Parsing engineering and empirical robustness. *Natural Language Engineering*, to appear.

Roberto Basili, Maria Teresa Pazienza, and Fabio Massimo Zanzotto. 2000. Customizable modular lexicalized parsing. In *Proc. of the 6th International Workshop on Parsing Technology, IWPT2000*, Trento, Italy.

R. Basili, A. Moschitti, and M.T. Pazienza. 2001. NLP-driven IR: Evaluating performances over text classification task. In *Proceedings of IJCAI 2001 Conference, Seattle, USA.*

B. Daille. 1994. Study and implementation of combined techniques for automatic extraction of terminology. In *The Balancing Act: Combining Symbolic and Statistical Approaches to Language, WorkShop of the ACL.*

Christian Jacquemin, editor. 2001. *Spotting and Discovering Terms through Natural Language Pro-*

cessing. The MIT Press, Cambridge, Massachussets, USA.

T. Joachims. 1998. Text categorization with support vector machines: Learning with many relevant features. In *In Proceedings of ECML-98*, pages 137–142.

G: Salton and C. Buckley. 1988. Term-weighting approaches in automatic text retrieval. *Information Processing and Management*, 24(5):513–523.

G. Salton. 1991. Development in automatic text retrieval. *Science*, 253:974–980.

Tomek Strzalkowski, Gees C. Stein, G. Bowden Wise, Jose Perez Carballo, Pasi Tapanainen, Timo Jarvinen, Atro Voutilainen, and Jussi Karlgren. 1998. Natural language information retrieval: TREC-7 report. In *Text REtrieval Conference*, pages 164–173.

Yiming Yang and Jan O. Pedersen. 1997. A comparative study on feature selection in text categorization. In *Proceedings of ICML-97*, pages 412–420, Nashville, US.

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