AUTOMATIC SPECIALIZED VS. NON-SPECIALIZED TEXTS DIFFERENTIATION: A FIRST APPROACH

M. Teresa Cabré[1]
Iria Da Cunha[1,2]
Eric Sanjuan[3]
Juan-Manuel Torres-Moreno[2,3]
Jorge Vivaldi[1]
[1] *Institut Universitari de Lingüística Aplicada (Spain)*
[2] *Laboratoire Informatique d'Avignon (France)*
[3] *École Polytechnique de Montréal (Canada)*

ABSTRACT: In this paper we would like to show that certain grammatical features, besides lexicon, have a strong potential to differentiate specialized texts from non-specialized texts. We have developed a tool including these features and it has been trained using machine learning techniques based on association rules using two sub-corpora (specialized vs. non-specialized), each one divided into training and test corpora. We have evaluated this tool and the results show that the strategy we have used is suitable to differentiate specialized texts from plain texts. These results could be considered as an innovative perspective to research on domains related with terminology, specialized discourse and computational linguistics, with applications to automatic compilation of Languages for Specific Purposes (LSP) corpora and optimization of search engines among others.

Keywords: Specialized Text, General Text, Corpus, Automatic Tool, Languages for Specific Purposes, Search Engines.

INTRODUCTION

There are several works about the differences between general and specialized texts. Most of them consider that lexicon is the most distinguishing factor (besides being the most visible) to carry out this differentiation. It is well-known that terms (units of the lexicon with a precise meaning in a particular domain [Cabré, 1999]) show the specialized content of a subject; therefore, they appear inevitably in texts of their domain. Kocourek (1982: 42) states that:

> La langue de spécialité est une variété de langue à dominante cognitive dont les ressources, qui sous-tendent les textes sur tous le plans linguistiques, sont marqués par des caractères graphiques, par des tendances syntaxiques et, surtout, par un ensemble des unités lexicales qui reçoivent dans les textes une précision sémantique métalinguistique.

Thus, other characteristic features of specialized texts (as grammatical features, both morphological and syntactic) can be considered as specific of these texts. Features as verbal flexion related to grammatical person, verbal tense or verbal mode have been underlined in some works (Kocourek, 1982, 1991).

Some authors, using small corpora, have established some grammatical phenomena that may differentiate specialized texts. In some cases, they have considered only a very limited number of features of a single category; in other cases, a scarce number of texts has been analyzed manually. Hoffmann (1976) analyzes the frequency of names and verbs into a general corpus and a specialized corpus. Some authors have studied verbs into specialized French corpora (Coulon, 1972; Cajolet-Laganière and Maillet, 1995; L'Homme, 1993, 1995). More works where differences between general and specialized texts are shown can be found in Cabré (2007).

The question we want to answer is: Is it possible to find specific characteristics into specialized texts moreover of their discourse conditions, that are external to the text, or the terminology they have?

In this paper we would like to show, using a specific software tool that we have developed, that certain grammatical features, besides lexicon, have a strong potential to differentiate specialized texts from non-specialized texts. Although this subject has not been studied in depth in the literature, we have carried out some preliminary works about it (Cabré et al., 2010; Cabré, 2007).

Moreover, the automatic tool we have developed is going to be very useful for two tasks: the automatic constitution of corpora of specialized texts and the optimization of search engines (for users searching specialized texts).

In Section 2 we explain the methodology of our work. In Section 3 we show the experiments we have carried out and the obtained results. In Section 4 we present some conclusions.

METHODOLOGY

The methodology to carry out this work has several stages. In the first place, we have selected some linguistic features that may be characteristic of specialized texts and general texts. The experiments where these features were detected are Cabré et al. (2010) and Cabré (2007). Table 1 shows them.

Table 1. Linguistic features used in our work.

POS	Tag meaning	% in generalist text	% in specialized text
A	Determiner	10.00	9.90
C	Conjunction	6.79	7.62
D	Adverb	10.30	10.54
E	Especifier	4.39	5.49
JQ	Qualifier adjective	8.43	9.00
J	Adjective	4.56	4.48
N4	Proper noun	8.05	6.34
N5	Common noun	10.53	10.59
P	Preposition	10.35	10.34
R	Pronoun	6.34	7.03
T	Date	0.42	0.07
VC	Verb (participle)	4.51	4.47
V1P	Verb (first person, plural)	0.25	1.16
V1S	Verb (first person, singular)	0.13	0.24
V2	Verb (second person)	0.03	0.05
V	Verb	10.38	10.12
X	Number	4.54	2.56
Total	—	100.00	100.00

The full meaning of these POS (Parts of Speech) tags can be seen on the following URL: http://www.iula.upf.edu/corpus/etqfrmes.htm. Some POS tags are produced by simplification of the full tag (ex. 'A' is a simplification of 'AMS', 'AMP',…).

In the second place, we have compiled a corpus, divided into two sub-corpora:

1. A sub-corpus including texts from the specialized domain of economics, mainly scientific papers, books, theses, etc. (with 292,804 tokens corresponding to 9.243 sentences).

2. A sub-corpus with plain language form newspapers (with 1.232,512 tokens corresponding to 36.236 sentences).

Texts of both sub-corpora have been extracted from the Technical Corpus of the Institute for Applied Linguistics (IULACT) of the Universitat Pompeu Fabra of Barcelona. It consists of documents in Catalan, Spanish, English, German and French; although the search through *bwanaNet* is at the moment restricted to the first three of these languages. It contains texts of several specialized domains (economics, law, computing, medicine, genome and environment) and plain texts from newspapers. All the texts are tagged with POS tags. This corpus is accessible on-line via http://bwananet.iula.upf.edu/. Further details on these resources are shown at Vivaldi (2009). In this experiment we only use texts from economics. This is a field where there is a large overlap between topics and vocabulary in specialized and non specialized publications, making the task even harder.

In the third place, we have developed a tool including the mentioned linguistic features and we have trained it using these two sub-corpora. The machine learning approach that we used is based on association rules, one of the most-known methods to detect relations among variables into large symbolic (i.e. non numerical) data (Amir et al., 2004).

We choose to work on sentences instead of entire documents. Indeed, documents can be classified using contextual information about their structure or statistical information about their specific vocabulary. At sentence level, none of these informations can be used. Therefore, the application that we propose not only allows to classify texts, it also allows to look for technical statements inside non specialized documents.

In the fourth place, we have evaluated the results of the tool. This evaluation is based on the capacity of the tool to differentiate sentences coming from specialized texts from others over the mentioned test corpora (specialized and non-specialized).

EXPERIMENTS AND RESULTS

In our machine learning experiments with association rules, we have randomly selected 9,000 sentences from each corpus. Therefore the experiment has been carried out on a set of 18,000 sentences with a total of 112,870 tokens. We used 90% of both corpora (specialized and non-specialized) for training and the remaining 10% of them for test corpora, repeating this split 30 times at random. For the training, we have used sentences level (although we have tested that only sentences with more than six words can be classified). We have a machine learning strategy based on the combination of lexical features (lemmas) and grammatical features (POS tags).

Table 2 shows an example of plain text and its corresponding generated test corpus text. In bold we have marked the category GEN, which is indicating that this sentence is classified as part of a non-specialized text. Observe that "plain text" section includes the sentence as found in the general corpus while the "generated corpus text" section includes just a list of the lemma/tags found in such sentence.

Table 2. Example of plain text and generated test corpus text.

Plain text	Tras el acuerdo con los pilotos, la dirección de Alitalia concluyó ayer de madrugada la negociación con los sindicatos del personal de tierra, que aceptaron 2.500 despidos (la propuesta inicial era de 3.500), la congelación de los salarios durante dos años y el bloqueo del fondo de previsión social durante el mismo periodo, para evitar la quiebra de la compañía.
Generated corpus text	**GEN** ser congelación despido previsión tierra dos dirección el tras para quiebra periodo negociación mismo piloto bloqueo = salario A Alitalia C D de N4 N5 personal compañía fondo P R que JQ V propuesta num X social con ayer aceptar madrugada sindicato concluir año inicial durante acuerdo y evitar

We consider association rules of the form X=>D where X is a set of at most 5 lemmas and/or tags, D is the decision: SPE for specialized and GEN for general. For a rule to be valid, X has to be included in more than 0.5% of the sentences (this is called the support of the rule) and more than 90% of these

sentences that include X have to be in category D (this is called the confidence of the rule). Since the right part of the rule is restricted to a few numbers of categories, we shall refer to these rules as decision rules. This kind of rules can be computed using standard GPL packages like "Apriori" by Christian Borgelt (http://www.borgelt.net/apriori.html).

Our experiments show that this strategy allows us to obtain 46,148 decision rules. It appears that:

1) 60% of the rules induce category SPE, which means that there are more implicit decision rules among specialized texts than non specialized ones.

2) 78% of the rules include at least one grammatical tag which shows that this information is significant to distinguish between these two categories.

Table 3 gives the list of POS tags that are effectively used in the resulting decision rules.

Table 3. Tags included in rules with the percentage of rules using them.

POS	% of rules using them
A	17.36
C	12.26
D	17.72
E	6.81
J	6.69
JQ	14.91
N4	12.11
N5	17.88
P	17.77
R	11.26
T	0.17
V	17.20
VC	6.70
X	4.48

Here is a sample set of 10 rules randomly extracted from the total list of decision rules. Rules are given in Prolog format: the decision is on the left and the two figures give respectively the support and the confidence of the rule.

SPE ← europea N4 JQ N5 (50, 100.0)

SPE ← millones X JQ P (70, 100.0)

GEN ← anunciar N4 P = (80, 98.3)

GEN ← ayer uno R N4 (10, 100.0)

SPE ← función C JQ D (12, 93.1)

GEN ← Gobierno haber VC V (60, 100.0)

GEN ← España que P = (100, 100.0)

SPE ← embargo sin de N5 (70, 100.0)

SPE ← internacional a R N5 (12, 90.8)

GEN ← presidente en R JQ (80, 93.0)

Therefore each rule indicates that if a given set of lemmas and tags is included in one sentence, there is a specific probability to classify the sentences as general (GEN) or specialized (SPE). As an example, the first rule may be read as follows: if the sentence under analysis includes the lemma "europea" and words with the POS tags: "N4", "JQ" and "NQ" then such sentence may be classified as specialised (SPE). The coverage of this rule is 50% with a 100% of precision.

Once this set of rules is available, it is possible to build a classifier that, given a sentence, looks for the set of rules that match the sentence and chooses the rule that has the highest confidence. One important feature of this type of classifier is that it indicates when it cannot take a decision.

Finally, for a given text under analysis if more than half of the sentences it contains belong to a given category the text is considered to belong to such category.

To evaluate the results of the classifier based on the total set of decision rules (Classifier_1) we have used precision, recall and F-Score measures. These results are shown in Table 4.

Table 4. Results of Classifier_1.

	Precision	**Recall**	**F-Score**
GEN	0.7602	0.8875	0.8190
SPE	0.8671	0.7239	0.7890
Average	0.8137	0.8057	0.8040

We have carried out another experiment using for the classifier (Classifier_2) only the association rules including at least one grammatical feature (POS tags). This is a subset of 36.217 rules (78%).

Results obtained by Clasifier_2 are shown in Table 5.

Table 5. Results of Clasifier_2.

	Precision	**Recall**	**F-Score**
GEN	0.7582	0.8959	0.8213
SPE	0.8749	0.7182	0.7889
Average	0.8166	0.8071	0.8051

This evaluation indicates that elimination of rules exclusively based on lemmas does not significantly degrade classifier performance. In fact, is seems that it lightly improves the average F-score. This shows that classifier performance mostly relies on rules with tags. Table 6 gives, for each tag, the percentage of decisions that used them.

Table 6. Tags used in decisions and percentage of decisions using them.

POS	**% of decisions using each tag**
C	4.62
D	3.30
E	1.67
J	1.17
JQ	3.21
N4	2.49
N5	8.71
P	9.63
R	2.9
T	0.11
V	2.05
VC	2.62
X	1.89

CONCLUSIONS

The results we have obtained until now show that the strategy we used in this work (machine learning techniques using association rules based on lexical and grammatical features) is suitable to differentiate specialized and plain texts. Moreover, we have shown that grammatical features are discriminant enough for this task.

In this application we choose GEN as default decision, but other strategies could be used. In particular we could use Hidden Markov Models (HMM), which would be a complementary approach. HMM are based on short sequences of tokens, meanwhile decision rules are based on small bags of tokens. We shall consider this enhancement in the future.

We think that these results constitute an innovative perspective to research on domains related with terminology, specialized discourse and computational linguistics, like for example automatic compilation of LSP corpora or optimization of search engines. Further experiments will be conducted using other corpora and in areas other than economics.

REFERENCES

AMIR, A.; AUMANN, Y.; FELDMAN, R. & FRESKO, M. (2005). Maximal Association Rules: A Tool for Mining Associations in Text. *Journal of Intelligent Information Systems, 5*(3), 333-345.

CABRÉ, M.T.; BACH, C.; DA CUNHA, I.; MORALES, A. & VIVALDI, J. (2010). Comparación de algunas características lingüísticas del discurso especializado frente al discurso general: el caso del discurso económico. In Proceedings of the *XXVII Congreso Internacional de AESLA: Modos y formas de la comunicación humana* (AESLA 2009). Ciudad Real: Universidad de Castilla-La Mancha.

CABRÉ, M.T. (2007). Constituir un corpus de textos de especialidad: condiciones y posibilidades. In M. BALLARD, & C. PINEIRA-TRESMONTANT. *Les corpus en linguistique et en traductologie* (pp. 89-106). Arras: Artois Presses Université.

— (1999). *La terminología. Representación y comunicación*. Barcelona: IULA-UPF.

CAJOLET-LAGANIÈRE, H. & N. MAILLET (1995). Caractérisation des textes techniques québécois. *Présence francophone, 47*, 113-147.

COULON, R. (1972). French as it is written by French sociologists. *Bulletin pédagogique des, IUT18*, 11-25.

HOFFMANN, L. (1976). *Kommunikationsmittel Fachsprache – Eine Einführung*. Berlin: Sammlung Akademie Verlag.

KOCOUREK, R. (1991). *La langue française de la technique et de la science. Vers une linguistique de la langue savante*. Wiesbaden: Oscar Branstetter.

KOCOUREK, R. (1982). *La langue française de la technique et de la science*. Wiesbaden: Brandstetter (2nd ed., 1991).

L'HOMME, M.C. (1993). *Contribution à l'analyse grammaticale de la langue despécialité : le mode, le temps et la personne du verbe dans quelques textes, scientifiques écrits à vocation pédagogique*. Québec: Université Laval.

— (1995). Formes verbales de temps et texte scientifique. *Le langage et l'homme 31*(2-3), 107-123.

VIVALDI, J. (2009). Corpus and exploitation tool: IULACT and bwanaNet. In P. CANTOS GÓMEZ & A. SÁNCHEZ PÉREZ. *A survey on corpus-based research. Proceedings of I International Conference on Corpus Linguistics* (CICL-09). (pp. 224-239). Universidad de Murcia.