What is my Style? Using Stylistic Features of Portuguese Web Texts to classify Web pages according to Users' Needs

Rachel Aires[1,2], Aline Manfrin[1], Sandra Aluísio[1], Diana Santos[2]

[1]NILC/ICMC-USP, [2]Linguateca, SINTEF
Instituto de Ciências Matemáticas e de Computação - ICMC-USP, SINTEF ICT
Av. Trabalhador São-carlense, 400. Centro. 13566-970. São Carlos – SP – Brazil, Pb 124 Blindern, 0314 Oslo, Norway
raires@icmc.usp.br, aline@nilc.icmc.usp.br, sandra@icmc.usp.br, diana.santos@sintef.no

Abstract

In this paper we investigate the use of stylistic features of Web texts in Portuguese to classify web pages according to users' needs, in order to improve Web Information Retrieval. We first describe a seven categories classification of users´ needs, which was the outcome of a qualitative analysis of two TodoBr logs (a major Brazilian search engine). We describe 46 shallow linguistic features, inspired by the works of Biber and Karlgren, and proceed describing the compilation of the corpus employed on the classifier training. Our aim is to obtain rules that can be applied on the classification of Web texts according to those seven users´ needs. Some experiments are reported, showing that it is possible, at least for some of the categories, to identify them reliably.

1. Introduction

The actual size of the Web and its variety of texts allow us to find almost any type of information. The size of the web in Portuguese was estimated in 5,090,230,228 words in early November 2002 (Aires & Santos, 2002). Current search engines do a good job in matching the documents' topic with the user's search topic. Although texts can be about the topic that the user is looking for, they may not fulfill his/her needs. The reason for this is that the user might be looking for a text about the same subject of the recovered texts, but belonging to a different genre, text type, register type, style or quality.

According to Karlgren (2000), style is the difference between two ways of saying the same thing. Systematic stylistic variation can be used to characterize the genre of documents. Genre depends upon context and can be defined as a group of documents that are stylistically consistent and intuitive to accomplished readers of the communication channel in question.

Biber (1988) has studied English texts variation using several variables, and found that texts vary along five dimensions. Registers would then differ systematically along each of these dimensions, relating to functional considerations such as interactiveness, involvement, purpose, and production circumstances, all of which have marked correlates in linguistic structure.

Other work that has explored relatively stable characteristics of texts to be used on text categorization consists of the studies presented in Karlgren (2000), two of which are particularly interesting for our work. The first one (Karlgren, 2000: Chapter 7) was carried out with features similar to Biber's, but concentrating on those easy to compute with a POS tagger. Using texts from the Brown Corpus, three experiments were performed, with two, four or fifteen categories, respectively, correctly classifying 478, 366 and 258 texts out of 500. The second study (Karlgren, 2000: Chapter 16) explored how an interactive system could be designed to incorporate stylistic information in its interface, categorizing retrieval results by genre, and displaying the results using this categorization. In this experiment eleven categories were employed and an user-centered evaluation was performed. The users were asked to execute two tasks each, using the prototype of the interface which uses stylistic features and the web search engine Altavista. Karlgren concluded that most users used the interface as intended and many searched for documents in the genres the results could be expected to show up in.

We believe that simple stylistic items like word-based statistics, text-based statistics and statistics on specific items, used by Biber, Karlgren and others, can be used as well to automatically classify texts according to basic users´ needs, decreasing the user effort to find the information he is looking for.

The goal of our study was to find regularities in a corpus composed by web pages in Portuguese, which could be used in rules to classify texts according to users´ needs. This work is part of a larger project that consists on the development of a linguistically motivated approach for information retrieval for Portuguese, named *Linguarudo*[1]. *Linguarudo* explores features of the language (Portuguese) during the interpretation of queries, matching and ranking. The results of the work presented here will be used on the dialog interface with the users. Our approach, by default, tries to detect automatically the user's need from his enquiries in natural language, based on pre-defined typical ways of posing questions, but also allows the user to choose the type of need his query is related to.

In the following sections we present the setup of our study and the results of the experiments carried out. The paper ends with a discussion of the results, limitations of the work and further work to be conducted.

2. Experimental Setup

2.1 Seven users´ needs

The classification in seven categories of users´ needs was the outcome of a qualitative analysis of two TodoBr[2] logs (a major Brazilian search engine). We selected these seven items as the most common users' needs by analyzing the logs of November 1999 and July 2002. This classification is based on what the user wants:

1) A definition of something or to learn how or why something happens. For example, what are the northern

[1] http://www.nilc.icmc.usp.br/nilc/projects/linguarudo.html
[2] www.todobr.com.br

lights? For this need, the best results would be presented by dictionaries and encyclopedias, or even textbooks, technical articles and reports and texts of the informative genre.
2) To learn how to do something or how something is usually done. For example, find a recipe of his favorite cake, learn how to make gift boxes, or how to install Linux on his computer. Typical results are texts in the instructional genre, such as manuals, textbooks, readers, recipes and even some technical articles or reports.
3) A comprehensive presentation or survey about a given topic, such as a panorama of 20th century American literature. In this case, the best results should be texts of the instructional, informative and scientific genres, e.g. textbooks, reportages and long articles.
4) To read news about a specific subject. For example, what is the current news about the situation in Israel, what were the results of the soccer game on the day before or find about a terrible crime that has just happened in the neighborhood. The best answers in this case would be texts of the informative genre, e.g. news in newspapers and magazines.
5) To find information about someone or some company or organization. For example, the user wants to know more about his blind date or to find the contact information of someone he met in a conference. Typical answers here are personal, corporation and institutional web pages.
6) To find a specific web page that he wants to visit, but does not remember its URL. For this type of need the results could be from any type of text or genre. The only way to identify this need would be the interface asking the user what type of page he is looking for.
7) To find URLs where he can have access to a given online service. For example, he wants to buy new clothes or to download a new version of software. The best answer to this kind of request is commercial text types (companies or individuals offering products or services).
These seven types are not claimed, however, to cover all kinds of user needs. Users may do all kinds of unpredictable searches, and we are not presuming to be able to recover their intentions by looking only at the logs[3].

2.2 The Corpus of Web texts

According to Gorsuch (1983: 332, apud Biber 1988: 65), the data in a factor analysis should include five times as many texts as linguistic features to be analyzed. Although we are carrying out a different kind of analysis, we followed this recommendation.
In our experiment we created a corpus with 511 texts extracted from the Web, 73 for each type of need[4] plus additional 73 texts that would not answer any of the six types used (we call it "others"), in order to have a balanced corpus. Picking up the same number of texts for each type we ended up with considerable differences in the size of the parts of the corpus concerning the number of words, as can be seen in Table 1. We did not consider this difference in the size in words a problem for our study as the training instances are the texts, not their words.
The selection of the texts was carried out by five different persons who were instructed to maximize the variety of genres and subjects that could be relevant for the types of needs 1 to 5 and 7. We have used websites which were already known to contain the sort of things we look for. All the text in the page was used (the web pages were automatically converted into plain text, resulting in losing any text that was part of a picture), and links were not followed. As the variants of Portuguese differ on the lexical, morphological and syntactic levels we decided to use only one variant – the Brazilian Portuguese – in order to prevent interference in the classifier training. The resulting corpus has 640,630 words.

1	2	3	4	5	7	others
76,841	51,959	19,6450	39,533	67,601	39,951	168,295

Table 1: Corpus size per type of user need

It should be noted that while Biber's 481 texts amounted to a corpus with approximately 960,000 words, due to the fact that Web pages/texts are often smaller than texts in other media we only achieved 640,630 words. Another alternative to create the corpus would be to randomly select from a Brazilian Web collection like WBR-99 (Calado, 1999). We avoided this alternative because we would have to classify those pages according to the user's needs we were interested in.

2.3 Stylistic Features

The 46 features[5] used in our study were based on the ones in Biber (1988) and Karlgren (2000). We did not rely on POS taggers, parsers or analysis in other levels, in order not to have to revise manually their output, otherwise errors could interfere with our results. We used mainly closed lists and employed 5 word-based statistics: type/token ratio (3), capital type token ratio (4), digit content (5), average word length in characters (6), long words (>6 chars) count (7); and 5 text-based statistics: character count (1), average sentence length in characters (2), sentence count (8), average sentence length in words (9), text length in words (10). The remaining 36 statistics were based on specific items:

- the subjective markers "acho", "acredito que", "parece que" and "tenho impressão que" ("I think so", "I believe that", "it seems that", "have the impression that") (11);
- the present forms of verb to be "é/são" ("is/are") (12);
- the word "que" (can be: noun, pronoun, adverb, preposition, conjunction, interjection, emphatic particle) (13);
- the word "se" ("if/whether" and reflexive pronoun) (14);
- the discourse markers "agora", "da mesma forma", "de qualquer forma", "de qualquer maneira" and "desse modo" ("now", "on the same way", "anyway", "somehow" and "this way") (15);
- the words "aonde", "como", "onde", "por que", "qual", "quando", "que" and "quem" on the beggining of questions (wh-questions) (16);

[3] See Aires & Aluísio (2002) for a preliminary investigation on making intentions explicit.
[4] Except for type 6, which, as explained above, can correspond to any kind of text.
[5] Numbers after the description of the category indicate the feature number used in the classifier.

- "e", "ou" and "mas" as sentence-initial conjunctions ("and", "or", "but") (17);
- amplifiers (18), conjuncts (19), downtoners (20), emphatics (21);
- persuasive verbs (22), private verbs (23), public verbs (24);
- number of definite articles (25); number of indefinite articles (26);
- first (27), second (28) and third person pronouns (29);
- number of demonstrative pronouns (30);
- indefinite pronouns and pronominal expressions (31);
- number of prepositions (32);
- place adverbials (33); time adverbials (34);
- number of adverbs (35);
- number of interjections (36);
- contractions (37);
- causative (38), final (39), proportional (40), temporal (41), concessive (42), conditional (43), "conformative" (44), comparative (45) and consecutive conjunctions (46).

2.4 The Classification Algorithm

We calculated the 46 features over the texts using a Perl script and used them to train a classifier using the J48 algorithm available on the Weka collection of machine learning algorithms (Witten & Frank, 2000). J48 is the Weka implementation of the decision tree learner C4.5. C4.5 is a well known classification algorithm, it was the best one from the other seven algorithms from Weka we have tried, it has already been used in similar studies (Karlgren, 2000) and it generates easily understanding clear rules. The others algorithms we have used were: ZeroR (13.7%), Conjunctive Rule (25.73%), OneR (29.31%), FLR (30.81%), HyperPipes (30,81%), Decision Table (44.36%) and Part (44.95%). To test the generated classifiers we did a 10-fold cross-validation test.

3. Results

We have trained classifiers using 2, 3 (2 categories plus "others"), 4, 5 (4 categories plus "others"), 6 and 7 categories (6 categories plus "others") (Table 2).

2 categories	4 categories	6 categories
1) the union of needs 1, 2, 3, 4 and 5 2) need 7	1) the union of needs 1, 2, 3 2) need 4 3) need 5 4) need 7	Need 1 Need 2 Need 3 Need 4 Need 5 Need 7

Table 2: Categories used

Table 3 presents the percentage of correct classifications for all the classifiers and Figure 2 shows precision and recall results divided by needs.

Number of categories	Percentage of correct
2 categories	90.93%
3 categories	76.97%
4 categories	65.06%
5 categories	56.56%
6 categories	52. 01%
7 categories	45.32%

Table 3: Percent of corrects using 10 fold cross-validation

The classification with 2 categories decides whether a page gives any kind of information about a topic or gives access to a service online. The corresponding resulting tree, which uses 10 features, is shown in Figure 1.

```
feature25 <= 2.578269
|  feature34 <= 0.453858
|  |  feature33 <= 0.053419
|  |  |  feature22 <= 0.041494
|  |  |  |  feature6 <= 4.481243: Need7 (16.0)
|  |  |  |  feature6 > 4.481243: Need12345 (2.0)
|  |  |  feature22 > 0.041494: Need12345 (3.0/1.0)
|  |  feature33 > 0.053419: Need7 (33.0)
|  feature34 > 0.453858: Need12345 (3.0)
feature25 > 2.578269
|  feature9 <= 11.322034
|  |  feature14 <= 0.451467
|  |  |  feature28 <= 0.287356
|  |  |  |  feature31 <= 0.613027
|  |  |  |  |  feature43 <= 0: Need12345 (8.0)
|  |  |  |  |  feature43 > 0: Need7 (11.0/1.0)
|  |  |  |  feature31 > 0.613027: Need12345 (24.0)
|  |  |  feature28 > 0.287356
|  |  |  |  feature14 <= 0.344828: Need7 (14.0/3.0)
|  |  |  |  feature14 > 0.344828: Need12345 (2.0)
|  |  feature14 > 0.451467: Need12345 (25.0)
|  feature9 > 11.322034: Need12345 (297.0/2.0)
```

Figure 1: J48 tree to classify in 2 categories

	Precision	Recall
2 categories		
Need12345	0.94	0.951
Need7	0.739	0.699
3 categories		
Need12345	0.866	0.901
Need7	0.636	0.671
Others	0.426	0.315
4 categories		
Need123	0.737	0.781
Need4	0.556	0.479
Need5	0.431	0.384
Need7	0.692	0.74
5 categories		
Need123	0.663	0.63
Need4	0.57	0.671
Need5	0.278	0.274
Need7	0.553	0.644
Others	0.35	0.288
6 categories		
Need1	0.395	0.428
Need2	0.446	0.452
Need3	0.478	0.438
Need4	0.632	0.589
Need5	0.358	0.329
Need7	0.671	0.699
7 categories		
Need1	0.409	0.521
Need2	0.411	0.411
Need3	0.507	0.493
Need4	0.577	0.562
Need5	0.361	0.301
Need7	0.606	0.589
Others	0.296	0.288

Figure 2: Accuracy by class for the 6 classifications

The classification with 4 categories differentiates among information about something, someone or some company/institution/organization, news, and online services. Finally, the classification with 6 categories is the full one we have presented in Section 2.1, excluding type 6 that can be of any type of text or genre.

The class "others" contains text types like blogs, jokes, poetry, etc. Although it makes the classification task harder, it cannot be ignored, as is often done in works dealing with classifiers for closed domains or those not dealing with real world applications. Since we are going to use this work in *Linguarudo*, we will be dealing with many different texts that are not from the seven users' needs types considered in its dialogue interface. Then, examples from those different types should be used during classifier training to be able to reliably identify the seven types vs. the others not catered for by *Linguarudo*.

Using a cross-validation strategy we obtain worse but more reliable figures. For example, for seven categories, using 90% of the corpus for training and 10% for testing we got 49.42% of correct results against 45.32% using cross-validation.

4. Discussion and Further Work

The work reported here can be considered preliminary, but it is the first, as far as we know, that tried to automatically categorise, in terms of user needs, the texts in Portuguese on the Web. Our hypothesis behind this study was that it is going to be easier for an user to choose among types of needs than between genres or text types; this has to be confirmed later using a user-centred evaluation.

A lot of work still remains to be done, but already at this stage we can draw some conclusions.

As said before, the corpus was built by five different persons using websites which were already known to contain the sort of things we look for. The result was a corpus with texts classified in mutually exclusive categories. However, we know that a text can be equally appropriated to answer two different users' needs, for example, the same text can be an answer for both type 1 and type 2. Then, the corpus must cater for texts that belong to multiple types. We intend to analyse the texts we already have in our corpus to reclassify those which can answer to more than one type of need. Those texts will then be assigned to a class that represents texts that answer both types, for example, instead of being classified as type 1 or type 2, it will be classified as type 12.

Second, our corpus does not have enough texts to represent all range of variation that some categories may display. The corpus must thus be enriched and increased in size so that it may be employed later on also by other researchers in IR of Portuguese, following the general philosophy of Linguateca (www.linguateca.pt).

Third, considerable work should be devoted to finding more specific discriminating features. The ones we have used are too generic and neither have they been developed for the Web nor for the Portuguese language.

Nevertheless, it was shown that it is possible to discriminate reliably at least among some of the categories, and this should have a positive impact in the usability of a Web system. Just to separate between pages that give information and those that offer services (a task with a success rate of 90.95%) seems intuitively useful.

We plan, as future work, to perform a detailed study of the discrimination features. As can be seen in Figure 1, only 10 of the 46 features have been employed to distinguish between two categories. For the 7 categories classification 40 features were used. These 2 cases exemplify the importance of analysing the resulting rules and eliminating those features that have not been used.

It is also interesting to compare the results using simple features with a new study using also features depending on POS taggers or parsers after lemmatizing the corpus.

As concerns the training process, we want to investigate whether good results can be obtained by always classifying one class against all others, i.e. turning the classification into a set of binary ones.

Besides the experiments reported on previous sections we tried four more algorithms from the Weka package: NNGe, VFI, Multilayer Perceptron and Bagging. All of them have had better percentage of corrects than C4.5, respectively: 45.47%, 47.8%, 53.44% and 54.9%. As a further step, we have to evaluate how easy would be to use the classification scheme generate by them in our application.

Finally, a related research topic is the use of a more flexible classification in terms of axes such as formal/informal, short/elaborated, contextualized or not, involved/detached, etc. allowing customized choices.

Acknowledgements

We thank Akwan Information Technologies (www.akwan.com.br) for the TodoBr logs; Luiz Carlos Genoves Jr. and Marcos Felipe Tonelli de Carvalho for the script to calculate the features and Crislaine Aparecida Francisco, Vanessa Silva Marquiafável and Lucélia Helena de Oliveira for building the corpus. This work was supported by Fundação para a Ciência e Tecnologia through the grant POSI/PLP/43931/2001 and co-financed by POSI.

References

Aires, R.V.X. & Aluísio, S.M. (2003). Como incrementar a qualidade das máquinas de busca: da análise de logs à interação em Português. Revista Ciência da Informação, 32(1), 5--16.

Aires, R. & Santos, D. (2002). Measuring the Web in Portuguese. In Proceedings of the Euroweb 2002 Conference (pp. 198--199). Oxford, UK.

Biber, D. (1988) Variation across speech and writing. Cambridge University Press. Cambridge, UK.

Calado, P. (1999) The WBR-99 Collection: Description of the WBR-99 Web collection data-structures and file formats. LATIN - Laboratório para o Tratamento de Informação, Dep. de Computação, Universidade Federal de Minas Gerais, Brazil.

Karlgren, J. (2000) Stylistic Experiments for Information Retrieval. PhD Dissertation. Stockholm University, Department of linguistics.

Witten, I.H. & Frank, E. (2000) Data Mining: Practical machine learning tools with Java implementations. San Francisco: Morgan Kaufmann.