

What is my Style?

Using Stylistic Features of Portuguese Web Texts to classify Web pages according to Users' Needs

Rachel Aires^{1,2}
raires@icmc.usp.br

Aline Manfrin¹
aline@nilc.icmc.usp.br

Sandra Aluísio¹
sandra@icmc.usp.br

Diana Santos²
diana.santos@sintef.no

¹ NILC / ICMC - USP
Cx. Postal 668
3560-970 São Carlos - São Paulo, Brazil
<http://www.nilc.icmc.usp.br>

² Linguatca / SINTEF ICT
Pb 124, Blindern
NO-0314 Oslo, Norway
<http://www.linguatca.pt/>

Seven users' needs

This classification is the outcome of a qualitative analysis of two TodoBr 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
- 2) To learn how to do something or how something is usually done
- 3) A comprehensive presentation or survey about a given topic
- 4) To read news about a specific subject
- 5) To find information about someone or some company or organization
- 6) To find a specific web page that he wants to visit, but does not remember its URL
- 7) To find URLs where he can have access to a given online service

The corpus of 511 Web texts

In our experiment we created a corpus with texts written in Brazilian Portuguese extracted from the Web, 73 for each type of need (except for type 6) plus additional 73 texts that would not answer any of the six types used (we call it "others"). The resulting corpus has **640,630 words**.

Some of the 46 stylistic features

We selected some features easy to compute from non-annotated text that might be clues for style, from a literature overview and the consideration of the texts themselves. Some of them are:

- Word-based statistics

capital type token ratio
average word length in characters

- Text-based statistics

sentence count
text length in words

- Other statistics:

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")

number of prepositions

It has been demonstrated that stylistic variation can be used to characterize the genre of documents. Our focus is on the investigation of the use of stylistic features of Web texts in Portuguese to classify web pages according to users' needs, in order to decrease the user effort to find the information he is looking for. 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 study is part of a wider work - Linguarudo -, aimed at exploring the use of NLP in Portuguese-aware IR (<http://www.nilc.icmc.usp.br/nilc/projects/linguarudo.html>).

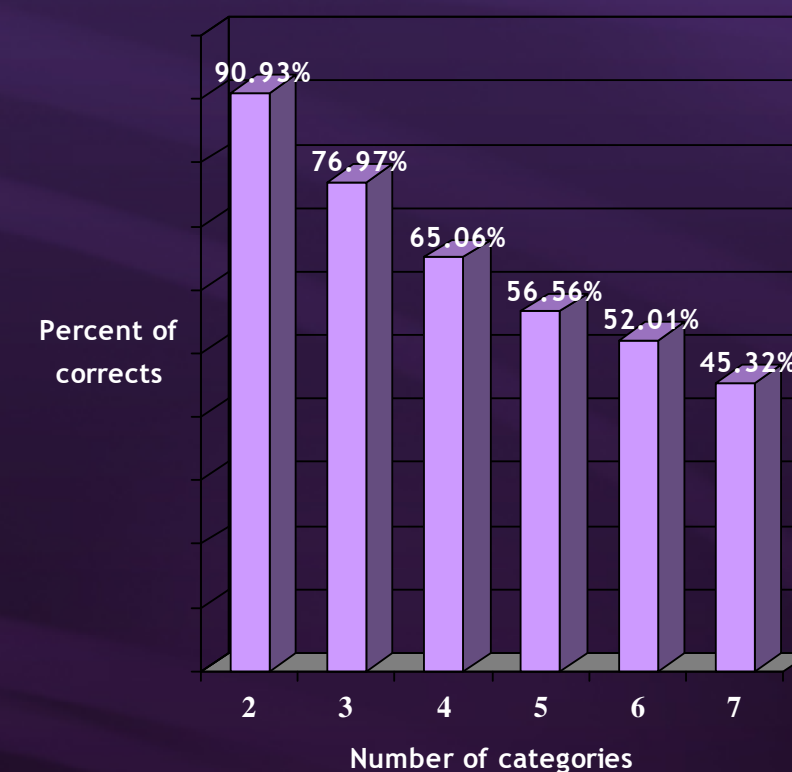
Results

We have trained seven classifiers using the J48 algorithm with 2 (the needs 1, 2, 3, 4 and 5 are considered together), 3 (2 categories plus "others"), 4 (the needs 1, 2 and 3 are considered together), 5 (4 categories plus "others"), 6 and 7 categories (6 categories plus "others").

To decide whether a page gives any kind of information about a topic or gives access to a service online we achieved 90.93% of corrects. To differentiate among information about something, someone or some company/institution/organization, news, and online services the percent of corrects was 76.97%. On the full classification scheme we achieved only 45.32%.

Further work

- To analyse the texts we already have in our corpus to reclassify those which can answer to more than one type of need.
- To enrich and increase in size the corpus, and make it publically available.
- To find 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.
- Concerning the training process, 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.
- To investigate whether the use of this classification improves user satisfaction.
- To try other classification algorithms.
- To have an alternative and more flexible classification scheme in terms of axes such as formal/informal, short/elaborated, contextualized or not, involved/detached, etc. allowing customized choices.



```
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)
```

J48 tree to classify in categories