Using Linguistic Data for Genre Classification

Peter WASTHOLM
Dept. of Linguistics and Philology
Uppsala University
Sweden
wastholm@stp.ling.uu.se

Annette KUSMA
Dept. of Linguistics and Philology
Uppsala University
Sweden
mak@stp.ling.uu.se

Abstract

Automatic categorization of texts into genres, rather than subject categories, is typically quite difficult. We have run a series of experiments on an annotated text corpus to determine whether the use of linguistic metadata (in this case, parts of speech) can be used to improve the performance of such categorizers. Our finding is that for this type of categorization, part of speech frequencies do indeed produce better results than the word frequencies normally used. Some genres, like reviews and imaginative prose, seem especially easy to recognize using this technique.

1 Introduction

Automatic text classification is a common task in the area of machine learning. The traditional approach is to have the system learn the statistical distributions of words in order to differentiate between the relevant categories. This works well where the subject matter varies significantly between the categories, such as when automatically filing incoming mail as “spam” or “ham” (where the latter, in most cases, is unlikely to contain references to cheap medications, easy ways of obtaining credit, and other matters that appear frequently in the former). However, the technique does not do an equally good job of determining genres of text, where more or less the same words can appear, but instead the language styles vary.

We have studied how genre classification can be improved by using linguistic metadata. To this end, we ran a series of experiments on the Stockholm–Umeå Corpus (SUC) (Gustafsson-Capková, 2004), which is both part-of-speech tagged and divided into genres1, and evaluated the results.

2 Method

2.1 The Corpus

The SUC consists of 500 texts, each around 2,000 words long, in nine main genres and 47 subgenres. The word “text” is used quite loosely; some are excerpts from longer texts (novels, in some cases), while others are aggregates of several shorter texts. As mentioned, the corpus is also part-of-speech tagged (the version we use is tagged with Parole tags (Ejerhed and Ridings, 2004)), letting us use and refine the categorizer to study and to running experiments.

The texts in the SUC are organized in a two-level genre hierarchy where the top level contains main genres such as “Press: Reportage,” “Press: Editorials,” “Skills, Trades and Hobbies,” and “Religion.” The second level further refines these genres into subgenres; “Reviews,” for instance, has subgenres for “Books,” “Films,” “Art,” etc. Thus, some genres and subgenres are subject-oriented while others denote different types of texts, which is primarily what we were interested in. We considered flattening the structure by merging subject categories but leaving “true” genres separate, but found no way of doing this without substituting one form of arbitrariness for another and instead decided to use only the top-level genres. Before further refining our genre recognizer we suggest that some resources be put into defining a more complete and uniform set of genres.

2.2 The Categorizer

A widely used algorithm for text categorization is the so-called naive Bayes categorizer. Its fundament is Bayes’ theorem (Russell and Norvig, 2003), which can be used to calculate the statistical probability that a hypothesis is true — in this case, that a particular categorization of a particular text is the correct one — based on observations one has made — essentially, the char-

1 The division follows closely that of the LOB and Brown corpora, which calls for a comparative study of the English texts.
acteristics of the text and how frequent these characteristics are in the various categories. The “naive” epithet reflects the fact that, in order to be able to make the calculations, one assumes that these characteristics (usually word frequencies) are independent of each other. Although this is not necessarily the case, naive Bayes text categorizers often work quite well.

We used a naive Bayes categorizer written in Perl and a few simple scripts that extracted information from the corpus and trained and tested the categorizer. We settled on four attributes to study: words, parts of speech, parts of speech plus subcategories (as defined by Parole), and complete Parole word classifications. For completeness, we also decided to run experiments on uni-, bi- and trigrams of each attribute.

2.3 The Experiments
We conducted our machine learning experiments in the typical fashion, using ten-fold cross-validation and separate test data (Witten and Frank, 2000). First, we set aside 75 of the texts (or 15%) for testing purposes. We then divided the remaining 425 texts into ten groups, each having either 42 or 43 members, and used cross-validation to improve and refine our categorizer. (All such grouping was done randomly, without stratification, and by automatic means.) Finally, we used the test data to determine the categorizer’s overall performance.

3 Results
3.1 The Big Picture
As mentioned, we examined the use of four different attributes: word lemmas (Lem), parts of speech (PoS), parts of speech plus subcategories (Sub), and complete Parole tags (Tag). For each of these attributes, we made separate experiments in which we used unigram, bigram and trigram frequencies \( (n = 1, 2, \text{ and } 3, \text{ respectively}) \). For \( n > 1 \), we included sentence beginning and end markers as tokens in our \( n \)-grams. Figure 3.1 shows the overall accuracy achieved in these experiments.

As can be seen, overall accuracy when using linguistic data was better than the best word frequency result in most cases, and better than the worst word frequency result in all cases. The highest overall accuracy was achieved by using subcategory bigrams. Our interpretation is that, for this corpus size, this represents the right trade-off between meaning and frequency: these bigrams contain enough information to mean something, but are still observed a sufficient number of times to be statistically significant. For a corpus of a different size, the outcome might have been somewhat different.

3.2 Further Analysis
This paragraph looks at detailed results, that is, it presents reflections on what went wrong or right and why. To our help, we have run a second set of tests, asking the categorizer to take care of not main genres but subgenres. Caution is to be used when working with the details of our experiments: there were very few texts of some categories in our test suite and some of the subcategories were not even represented at all. Obviously, it is hard to make generalizations based on figures of size one or two. Nevertheless, just for curiosity’s sake, let us assume for the moment that the detailed results are significant per se.

Not unexpectedly, a comparison of the lemma-based versus the tag-based test results reveals a clearly distinctive pattern. For the lemma-based tests in many cases, a wrong prediction was one of genre K (“Imaginative Prose”), and peeping at the results for subgenres, nearly all guesses became K — and more so, they were explicitly of genre KK (“General Fiction”); considerably many wrong guesses were

---

\(^2\) We tried to keep the original proportions.

\(^3\) This implies some missing information, namely that of the performance of the categorizer on texts of that kind, which in turn makes the overall accuracy figures non-applicable to some full-fledged test suite, or say, to real life situations (assuming the corpus presented a reasonable account for the statistics of all possible genres which it obviously does not) but it does not falsify the recognition results on the texts actually encountered.
also made on genre J (“Learned and Scientific Writing”). There were 10 texts of category K in the test suite, all of which were guessed correctly (speaking in terms of main genres). The other main genre that improved the overall accuracy of the lemma-based tests was H (“Miscellaneous”), which despite of the many occurrences in the test suite scored low on mis-categorizations; again, looking at the subgenre tests, four HB texts (“Municipal Publications”) were categorized mainly as HC (“Financial Reports, Business”) and eight HA (“Federal Publications”) were all correctly categorized. As a categorization of texts into genres based on contents seems rather pointless, nothing more is said about the lemma-based texts here.

The three tag-based experiment sets show the following similarities:

Genres A (“Press, Reportage”) and E (“Skills, Trades and Hobbies,” the subgenres being for example “Society Press” or “Occupational and Trade Union Press”) get mixed up with each other in considerably many cases. This is not surprising since the division seems based on the type of source (newspaper versus journal or periodical type), rather than on type of writing style. We doubt that a human being could (or would) distinguish these presumptive genres without knowing the source of the texts.

Genre B (“Press, Editorials”), with only two examples in the testsuite, was misguessed as E (“Skills, Trades and Hobbies”) in often. This seems strange since articles of that kind are reasonably easily identified by a human. We give them a high frequency of questions and a special sort of information structure (a repetition of a fact-question sequence) as well as a considerable frequency of first person pronouns as their main characteristics. Questionmarks, however, do not have a distinguished tag from fullstops or exclamationmarks which rules out the detection of two of those characteristics.

Only three texts of genre C (“Reviews”) are found in the testsuite. In all non-lemma-based experiments except for complete tag trigrams, all three were predicted correctly as C. Looking at the subgenre-based tests, a film review was categorized as a book review and music reviews were classed as film or show reviews. Other subgenres of C were theatre, art and radio/TV reviews. Intuitively one would expect subcategorization\(^4\) to improve if lemma and tag-frequency statistics were used at the same time since the subcategorization seems to be of content type. One would like to look at the ‘profile’ that the categorizer assigns to this genre (or others, of course), but time limits have made it impossible for us to modify the program to this end. The fact that one text can consist of around five to ten entities of type review should be accounted for in some way or other within the tag-statistics, being for an extensive use of proper noun tags.

\(^4\)This sentence being, of course, a sideline from our study of main genres of the corpus.

4 Discussion

Our experiments show that categorization performance can be improved by using linguistic data instead of, or in addition to, the more traditional approach of using word frequencies. We have not used any machine learning tricks (boosting, bagging, etc.) in our experiments, as we felt this would shift our focus away from the distinction between using vs. not using linguistic data, which was what we had set out to study. Naturally, the use of linguistic data does not preclude the use of such performance-enhancing methods.

It might be interesting to take a look at the statistical profiles, so to say, that the categorizer assigns to every genre after the training process. Apart form their computational linguistic relevance, they might give us some figures to back up theories about certain genres containing certain proportions of grammatical constellations.

As always, time was limited, and there are many things we could have done had we had more time. The most obvious additional activity would have been to try more types of linguistic attributes. For instance, it seems plausible that the average lengths of paragraphs, sentences, and words; the types of proper nouns used (places vs. people, for example); and the frequencies of headings and subheadings may differ between different genres of text. It would also have been interesting to run the corpus texts through a chunker or parser to be able to study the complexity of sentences (e.g., the average number of prepositional phrases or subordinate clauses per sentence). It should be noted, though, that since we included sentence beginning and end markers in our \(n\)-grams when \(n > 1\), and since all texts are roughly the same length (around 2,000 words), average sentence length does affect our categorizer’s choices.

The recognition of genres such as poetry (which is missing from the SUC) might make use of other length comparisons, such as aver-
age line length, that is, the number of words or characters between newline codes. The lengths (not just frequencies) of noun phrases, adjective phrases and subordinate clauses might also be used as a distinguishing property, perhaps for very formal genres such as publications from authorities, legal documents, and financial reports.

There may, of course, also be other characteristics for each genre that might be detectable by means of some algorithm and whose detections might aid our part of speech based frequency counts. One thing that comes to mind is categorizing words into levels of formality/triviality or into functional groups such as discourse markers. This obviously requires either some extension of the tags used in the corpus or, even more usefully (and re-usably), a word list for each level, group, etc. These, of course, being projects of dissertation size in themselves, were not considered for this work.

5 Acknowledgements

Our thanks go to Beáta B. Megyesi for helping us come up with such an interesting subject for this assignment and for providing us with a copy of the SUC.

References


