Detecting Anomalous Documents in a Corpus-driven Language Model

Kristin Yancey
Elizabeth Hohman

Naval Surface Warfare Center
Computational Mathematics and Statistics Branch, Code Q21
18444 Frontage Road, Suite 327
Dahlgren, VA 22448-5161
kristin.yancey@navy.mil

Abstract

Given a corpus of documents purporting to be in a single language, we develop a methodology for detecting documents written in a different language. Unlike previous work in language identification, the methodology does not assume any prior knowledge specific to either language. Information theoretic methods are used to determine key terms that are highly predictive of language identity, and a language model is then built from the keywords and compared to models developed from each document using rank order statistics, cosine dissimilarity, and other approaches. We evaluate the different methods of constructing and ranking the models on a corpus of multilingual Wikipedia articles.

Keywords: text processing, language identification, anomaly detection

1 Introduction

Existing language identification systems generally rely on the fact that the possible languages are known in advance and the language models can thus be predefined with as much information as is available about the languages. As well as the information that can be mined from the training corpus, some systems may use stop-word or phrase lists, lexicons, thesauri, stemmers, decompounders, part-of-speech taggers, and so on to aid in identification. In general, the purpose of language identification is to select the correct automatic translator for a given piece of text.

The problem we wish to solve is different: better described as a problem of outlier detection than one of language identification. We have a corpus of documents, such as a Wikipedia, which purports to be in a given language, such

as Zulu. We have no knowledge of Zulu, nor do we wish to obtain knowledge, lexicons, etc. for each of the approximately 265 Wikipediae which we wish to process. Thus we have a corpus containing primarily Zulu documents, a small number of which may not be valid Zulu articles.

These anomalous articles may be one of several things. They may be no more than messages to the effect that the article is a stub, an invalid redirection, a candidate for deletion, etc., and such messages often appear in English rather than the language of the Wikipedia, since English as the originating Wikimedia language has more or less become the lingua franca of Wikimedia, but the messages also appear in common European languages like French, Italian, Spanish, or may appear in several languages at once. The anomalous article may also be a valid article on the subject at hand but provided in a language other than that of the Wikipedia which contains it. Because anyone may edit Wikipediae without immediate oversight (although the rules regarding this are becoming stricter), the purportedly Zulu article may also contain, in lieu of valid text, vandalism consisting of words one or more languages and even slang or nonsense words.

Vandalism may also occur in the correct language (that of the containing Wikipedia). For instance, articles could contain text unrelated to the subject indicated by the title or grossly misrepresenting that subject. However, it is beyond the scope of our current effort to identify vandalism of this kind, as this task would require natural language understanding and is clearly a major task in its own right.

Of course, anomalous articles of the kinds described above should, according to Wikimedia's rules, be remediated or deleted. However, there is no guarantee that all articles are clean at the time a snapshot is made, and in fact message to the effect that an article has been deleted may still appear in the area designated for article text.

In fact, although we are interested in developing a method to address the problem described above for the purposes of corpus data cleaning, such a technique would also be of use to the maintainers of online wiki-style repositories, enabling them to identify spam and vandalism automatically, without having to wait for human users to stumble across the faulty articles.

2 Language Models

For the purposes of language identification, language models in the literature generally consist of one of the following: (1) short words (Grefenstette [1995]), (2) n-Grams (Cavnar and Trenkle [1994]), or (3) frequent words (Souter et al. [1994]). These standard approaches of language identification make implicit use of the morphology of languages.

Different models have been produced by varying the maximum number of letters defining a short word (Prager, 1999), specifying whether n-gram lengths should be different or fixed, which values of n should be used, varying the number of most frequent terms, n-grams, etc. to use in a model, and using different ways of computing the term frequency (e.g. whether and how to normalise).

Overall, n-gram models have been shown to perform as well as if not better than short word models for various more or less fusional European languages by authors such as Grefenstette [1995]. For Asian languages such as Chinese and Japanese which contain no character separators, n-grams are essential to separate the text into lexemes, even though these languages are relatively analytic, and in general would be good candidates for whole word models (McNamee and Mayfield [2004]).

Since we wish to be able to handle an extremely broad selection of languages, ranging from analytic to highly synthetic, we hypothesize that n-gram-based models will be best suited to our task, but for the purpose of anaylsing possible dissimilarity metrics, we examine a variety of language models. Once a dissimilarity metric has been selected, we can in future work return to the problem of choosing the best composition and size for the language model.

3 Selecting a Dissimilarity Measure

Given a language model and the set of document models from which it was constructed, we can treat the models either as term-frequency vectors (including null or zero entries) or as lists of frequent terms ranked from most to least frequent. From Zipf's Law, we know that there is a predictable relationship between the ranking and frequency of words in a corpus, so the two are analogous ways of computing a dissimilarity measure. We define the frequency of a term to be the number of times it appears in the corpus. Zipf's law states that in a population of N documents, the frequency of a term is defined by the following function (Zipf [1932]):

$$f(k; s, N) = \frac{\frac{1}{k^s}}{\sum_{n=1}^{N} \frac{1}{n^s}} \tag{1}$$

where k is the rank of the term, and s is the shape parameter, discussed below.

Regarding the models as vectors of term frequencies, we can compute standard measures of dissimilarity such as the cosine, L2, and Kullback-Leibler distances. If instead we use the ranking of the terms in the two models, their ranks can be compared by measures from the literature such as the rank-order distance (Cavnar and Trenkle [1994]) or the Spearman rank correlation coefficient (Callan et al. [1999]).

References

Jamie Callan, Margaret Connell, and Aiqun Du. Automatic discovery of language models for text databases. In *Proceedings of the 1999 ACM SIGMOD International Conference on Management of Data*, pages 479–490. ACM Press, 1999.

William B. Cavnar and John M. Trenkle. N-gram-based text categorization. In *Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval*, pages 161–175, 1994.

Gregory Grefenstette. Comparing two language identification schemes. In *Proceedings of the Third International Conference on the Statistical Analysis of Textual Data (JADT'95)*, pages 263–268, 1995.

Paul McNamee and James Mayfield. Character n-gram tokenization for european language text retrieval. volume 7, pages 73–97. Springer Netherlands, January 2004.

Clive Souter, Gavin Churcher, Judith Hayes, John Hughes, and Stephen Johnson. Natual language identification using corpus-based models. *Hermes Journal of Linguistics*, 13:183–203, 1994.

G. K. Zipf. *Selected studies of the principle of relative frequency in language.* Harvard University Press, Cambridge, MA, 1932.