A Dissimilarity Approach to Detecting Out-of-Language Documents

Yancey, Kristin
Marchette, David J.
Naval Surface Warfare Center, Computational Mathematics and Statistics Branch
18444 Frontage Road, Suite 327, Dahlgren, VA, 22448-5161, USA
E-mail (corresponding author): kristin.yancey@navy.mil

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: information that can be mined from the training corpus, stop-word or phrase lists, lexicons, thesauri, stemmers, decompounders, taggers, etc. 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. We wish to find these anomalous articles and remove them from our corpus.

Anomalous articles may be Wikimedia messages (stub, invalid redirection, candidate for deletion, etc.), which often appear in one or more common languages (English, French, Spanish, etc.) rather than the language of the Wikipedia which contains them. An anomalous article may also be an instance of vandalism (often slang or nonsense), or it may be written in a different language whether or not it contains relevant subject matter. Anomalous articles like these should according to Wikimedia's rules be remediated or deleted, but this requires a massive amount of human oversight. In fact, rather than using the method we propose as described above, the method could be used by the maintainers of on-line databases to quickly and automatically identify potential instances of spam and vandalism.

For the purposes of language identification, a language model, $\mathscr{L}$, generally consists of one of the following: short words (Grefenstette [1995]), n-grams (Cavnar and Trenkle [1994]), or frequent words. Since we wish to be able to handle an extremely broad selection of languages, ranging from analytic to highly synthetic, we hypothesise that n-gram-based models will be best suited to our task, but for the purpose of analysing 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, size being the number of terms it contains.

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 wherein k is the rank of the term, and s is the shape parameter:

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

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, we can use measures from the literature such as the rank-order distance (Cavnar and Trenkle [1994]) or the Spearman rank correlation coefficient (Callan et al. [1999]).

Instead of comparing each document model d with the much larger language model $\mathscr{L}$, we compare each d with $\hat{\mathscr{L}}$, $|\hat{\mathscr{L}}| = |d|$. To construct $\hat{\mathscr{L}}$ this we chose p $\hat{\mathscr{L}}$'s for each document d, where p is a reasonably large number. A uniform sampling from $\mathscr{L}$ will not accurately represent an "average" document, so we use Zipf's distribution to sample from $\mathscr{L}$ and obtain $\hat{\mathscr{L}}$.

In the smaller Wikipediae, many of the articles are very short. Due to the low number of valid documents and the resulting sparsity of the language model, the Zipf function is an imperfect fit for many of the languages we wish to analyse, thus we choose a generic shape parameter $s* = 1$. We can now sample terms from $\mathscr{L}$ according to the Zipf distribution with $s = s^*$ to obtain $\hat{\mathscr{L}}$.

For each document d and each dissimilarity measure $dist$, we have can compute the distance $dist(\mathscr{L}, d)$ for all dissimilarity measures under consideration and thus:

$$c = \frac{\sum_d (dist(\mathscr{L}, d) - \mu)}{\sum_d \left(dist(\mathscr{L}, \hat{\mathscr{L}}) - \mu \right)} \text{ where } \mu = \frac{\sum_{\hat{\mathscr{L}}} dist(\mathscr{L}, \hat{\mathscr{L}})}{|\{\hat{\mathscr{L}}\}|} \tag{2}$$

A small value of c indicates a dissimilarity measure that minimises the apparent difference between the actual documents and the $\hat{\mathscr{L}}$s, so we select the measure which results in the lowest value for c. We see from the results that the preferred dissimilarity measure is the Kullback-Leibler distance, which we will use in subsequent work.

Distance Measure	c
Kullback-Leibler	17.51
Euclidean Distance	119.69
L1 Distance	1097.54
Spearman Rank	24291.81
Cosine Dissimilarity	27106.78
Rank-Order	21296698.60

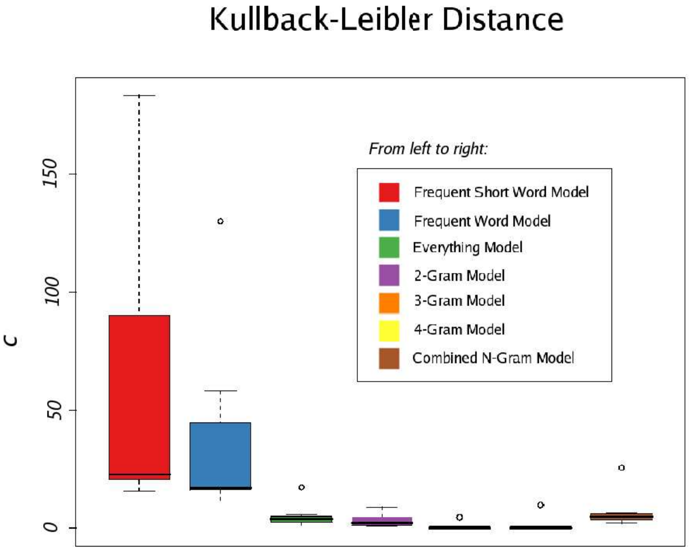

Figure 1: The table, above, contains the c values for all dissimilarity measures, averaged over the set of languages tested and all model types. The plot at right shows the c values of the Kullback-Leibler distance for all languages tested, separated by model type.

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.