MLADENIC D. (1998), "Turning Yahoo to Automatic Web-Page Classifier", in *Proc. European Conference on Artificial Intelligence*: 473-474.

RAYSON P. and GARSIDE R. (2000), "Comparing corpora using frequency profiling", in *Proc. of the Comparing Corpora Workshop at ACL 2000*,Hong Kong: 1-6.

REHM G. (2002), "Towards automatic web genre identification – a corpus-based approach in the domain of academia by example of the academic's personal homepage", in *Proc. of the Hawaii Internat. Conf. on System Sciences*.

SANTINI M. (2007), *Automatic Identification of Genre in Web Pages*, PhD thesis, University of Brighton.

SEBASTIANI F. (2002), "Machine learning in automated text categorization", in *ACM Computing Surveys*, n° 1, vol. 34.

SHAROFF S. (2006), "Creating general-purpose corpora using automated search engine queries", in Baroni M. and Bernardini S. (eds), *WaCky! Working papers on the Web as Corpus*, Gedit, Bologna, http://wackybook.sslmit.unibo.it.

SINCLAIR J. (1996), *Preliminary recommendations on corpus typology*, Expert Advisory Group on Language Engineering Standards document EAG–TCWG–CTYP/P, http://www.ilc.cnr.it/EAGLES96/corpustyp/corpustyp.html.

WITTEN I. and FRANK E. (2005), *Data Mining: Practical machine learning tools and techniques*, Morgan Kaufmann, San Francisco.

ZHAO Y. and KARYPIS G. (2004), "Empirical and Theoretical Comparisons of Selected Criterion Functions for Document Clustering", in *Machine Learning*, n° 3, vol. 55.

Identification of Languages and Encodings in a Multilingual Document

Anil Kumar Singh[1] and Jagadeesh Gorla[1]
Language Technologies Research Centre, IIIT, Hyderabad, India

Abstract

Text on the Web is available in numerous languages and encodings, often not according to any standards. The number of multilingual documents on the Web is also increasing. The problem of identifying the languages and encodings in a multilingual document and marking portions of a document with them has not been addressed so far. We present an exploration of this problem, the implied or required assumptions, and a solution. The problem can be divided into three parts: monolingual identification, enumeration of languages and identification of the language of every portion. For enumeration, we have been able to get a precision of 96.20%. We also experimented on language identification of each word. Given correct enumeration, we could obtain *type* precision of 90.91% and *token* precision of 86.80%. Finally, we show how precision is affected by language distance.

Keywords : multilingual, language identification, encoding, retrieval.

1. Introduction

One user one language and *one document one language* have been the assumptions on which much of the work on computers, the Internet and even Natural Language Processing (NLP) has been based. But as more and more people from around the world, especially from countries with many languages, have joined the community of computer and Internet users, the importance of accommodating bilingualism and multilingualism is gradually being realized.

Language identification becomes an important problem in the electronic world of many languages (Gordon 2005), even more so when multiple languages are mixed up in one document. Monolingual identification has been attempted by many researchers and it is now considered by many to be an almost solved problem. But multilingual identification has been rarely attempted. This is partly due to the fact that for a long time most of the documents on the Internet were monolingual. Multilingual documents are becoming more common now. Since it is very difficult to directly estimate the number of multilingual documents, we have used an indirect method as shown in Table 1.

[1] {anil,jagadeesh}@research.iiit.ac.in