Reconsidering Language Identification for Written Language Resources

Baden Hughes, Timothy Baldwin, Steven Bird, Jeremy Nicholson and Andrew MacKinlay

Department of Computer Science and Software Engineering
The University of Melbourne
Parkville VIC 3010, Australia
{badenh,tim,sb,jeremymn,amack}@csse.unimelb.edu.au

Abstract

The task of identifying the language in which a given document (ranging from a sentence to thousands of pages) is written has been relatively well studied over several decades. Automated approaches to written language identification are used widely throughout research and industrial contexts, over both oral and written source materials. Despite this widespread acceptance, a review of previous research in written language identification reveals a number of questions which remain open and ripe for further investigation.

1. Introduction

As evidenced by the literature, written language identification has received less focus than spoken language identification. Numerous reasons for this have been proposed by earlier researchers: Muthusamy and Spitz (1996, p 275) summarised it thus:

> Judging by the results, it appears that language ID from character codes is a less hard problem than that from speech input. This makes intuitive sense: text does not exhibit the variability associated with speech (e.g., speech habits, speaker emotions, mispronunciations, dialects, channel differences, etc.) that contributes to the problems in speech recognition and spoken language ID.

In contrast with this apparent simplicity, the task of identifying the language in which a given document is written has been relatively well studied over several decades. Automated approaches to language identification are now used widely throughout research and industrial contexts. Despite this widespread engagement, a review of previous research in written language identification reveals a number of questions which remain open and ripe for further investigation.

In this paper we review a number of classes of methods for enabling language identification for written language resources, observing their relative strengths and weaknesses from research published over several decades - modeled on a similar survey by Sibun and Reynar (1996). Our motivation is to consider the remaining open questions in the area of language identification for written language resources, a number of which are expressed in Section 3. Finally we draw conclusions based on the findings of our survey and motivate future work in a variety of areas.

2. Data Resources and Tools

A variety of written language identification tools are in circulation in the language technology community. Arguably the best known is van Noord's TextCat,[1] an implementation based on character n-gram sequences. Other well known implementations include BasisTech's Rosette Language Identifier,[2] and a number of web based language identification services such as those by Xerox[3] and Ceglowski[4]. While this paper does not specifically catalogue and review these tools, it is important to note that they are freely available and commonly used.

On the other hand, one significant shortcoming of written language identification research has been that there is no common data set on which evaluations can be based. Most of the existent research is based on relatively small ad-hoc collections drawn from larger sources, and almost without exception, these data sets are not made available to other interested researchers. This is not to say that standardised data sets are not available: multilingual corpora are commonplace, but they have not been specifically utilised by researchers working on the written language identification problem.

3. Approaches and Methods

In arguably the seminal work in the area, Gold (1967) construed language identification as a closed class problem: given a list of possible languages, a subject is provided with an exemplar and in a finite period asked to classify the exemplar. The experiments were constrained by all languages having a common orthographic representation, and facilitated both by a generation mechanism (randomised selection of strings from a given text) and an informant mechanism (a given string is nominated as being from a given language).

Subsequent work was dominated by the use of various *feature-based models*. Cavnar and Trenkle (1994) generated task specific statistical models of character co-occurrence; Dunning (1994) used Bayesian models for character sequence prediction. Darnashek (1995) applied models based on dot products of frequency vectors of words in a corpus. Souter et al. (1994) also derived task specific and corpus specific models, but demonstrated considerable performance in open domains. More recently, McNamee and Mayfield (2004) applied character n-gram tokenisation as the basis for language identification in cross language text retrieval contexts.

[1] http://odur.let.rug.nl/ vannoord/TextCat/

[2] http://www.basistech.com/language-identification/

[3] http://www.xrce.xerox.com/competencies/content-analysis/tools/guesser

[4] http://languid.cantbedone.org

More generally, explorations of *similarity based classification and categorisation* have also yielded interesting results. Aslam and Frost (2003) demonstrate that an information theoretic measure of document similarity can have suprisingly good performance in multilingual classification.

An interesting theme of research closely related to language identification has been conducted in the area of optical character recognition, specifically in the linguistic *classification of scanned data*. Work by Sibun and Spitz (1994) has shown that language identification can be performed reliably on individual segments of scanned text documents. Abstraction of characters into approximate character shapes has shown promising results in the OCR context (Sibun and Reynar, 1996). Lee et al. (1998) demonstrate that document level language classification can be achieved reliably even without considering the character sequences which make up a text.

Another theme of research has revolved around the *detection of the character encoding* of a given document, and hence deducing the language in which the document is written. In the leading work in the area, Kikui (1996) applies the encoding detection problem to textual resources on the internet, and determines language of a text on this basis across 9 languages and 11 coding systems. This method has two obvious shortcomings: there is no 1:1 correspondence between an encoding and a language, even if there was there is no guarantee that the declared encoding actually has a strong positive correlation to the language of a text. However, it is one of the first papers to consider web-sourced data, a point to which we will return later.

The application of *support vector machines* and *kernel methods* to the language identification task has been considered relatively recently by a number of researchers. Teytaud and Jalam (2001) investigated kernel methods based on n-grams derived from inverse document frequency indices. Lodhi et al. (2002) propose a method using the character sequence as opposed to words as the nexus for kernel creation, and show promising results for discrimination between texts of different languages and for clustering based on string kernels. More recently Kruengkrai et al. (2005) have revisited the language identification task and show state of the art results using string kernels with very small amounts of training data.

More general *statistical methods* have been tested — Poutsma (2001) used Monte Carlo based sampling to generate large random feature models which are then used to identify a language based on the occurrence of the features. Elworthy (1998) demonstrated the use of confidence limits to ensure efficient termination based on iterative evidence acquistion, and used this approach to simplify statistical approaches by leveraging confidence intervals prior to consideration of fine grained distinctions which significantly increase the complexity of purely statistical methods.

The use of *linguistically grounded models* has also been considered. Grefenstette (1995) used correlated word and part of speech (POS) co-occurrence as the basis for determining if two given text samples were from the same, or different languages. In a similar vein, Giguet (1995) considered tokenisation patterns across languages, and derived a cross-language feature model for tokenisation which was used to identify languages based on their tokenisation similarity. Considering smaller linguistic units, Giguet (1996) also considered error analysis from both a linguistic and statistical perspective. Lins and Goncalves (2004) considered the use of syntactically derived closed grammatical class models, matching syntactic structure rather than words or character sequences as many previous works.

Another approach, with strong mathematical foundations, was demonstrated by Beesley (1988). Drawing on core ideas from cryptanalysis, he demonstrated the use of letter and sequence probabilities as a method for written language identification. Importantly too, this paper contributed a typology of a number of problems we will consider later, including language identification for closely related languages.

4. Work in Spoken Language Identification

With reference to spoken language, a significant amount of research has been conducted in automatic language identification, particularly oriented at discovering efficient mechanisms to trigger language model or grammar switching for application instances. An overview of this research can be found in Muthusamy and Spitz (1996). The majority of the speech-oriented research has been focused on language detection in data streams (as opposed to discrete documents typically used in the written language identification tasks). A key enabler in the advances in spoken language identification domain has been the long availability of standardised data sets (for example the OGI Multilanguage Telephone Speech Corpus (Muthusamy et al., 1992) has been available for almost 15 years) which have allowed for the shared evaluation task model to be adopted for advancing the research agenda.

5. Outstanding Issues

Having considered this range of prior research, we turn to the identification of open questions in the domain of automatic language identification for written language resources. We offer to the language resource creation and evaluation community a number of areas of research which we believe are not adequately addressed in published research to date.

5.1. Supporting Minority Languages

The majority of published research is focused on languages which are spoken by large numbers of speakers, or are well resourced in terms of written language resources, or a combination of both. Very few published results are available which include languages which are not in these categories. It has often been seen elsewhere that approaches which perform well for major languages often do not scale to smaller ones, and the lack of experimental evidence to date is not sufficient to conclude that language identification for minority languages approximates the performance demonstrated for major languages. How well therefore, do existing techniques support language identification for languages which form the bulk of the more than 7000 languages identified in the Ethnologue (Gordon, 2005)?

5.2. Open Class Language Identification

Most work to date has focused on the classification of a given document according to a pre-determined list of candidate languages, and hence assessed performance on how well a given identification or classification approach handles this task. A different task is to allow for open class language identification whereby a text can be classified as being in unspecified language(s).

5.3. Sparse or Impoverished Training Data

There is little consideration given as to the performance of the variety of systems in environments when the amount of gold standard data for training is small. In our context, we define poor as being less than several thousand words of correctly identified text - all of the previous work assumes collections of minimally tens, and frequently hundreds of thousands of words of gold standard data for language identification. How well do existing methods language idenfitication work when the only sample accessible is 50/100/250 words or 50/100/250 characters?

5.4. Multilingual Documents

The vast majority of research to date involves handling written language resources in a single language. Given the increasing prevalence of multilingual documents, we see the need for systematic exploration of the language identification task both at finer granularity (eg sentence, paragraph, section) within a multilingual document and in quantitative terms (eg a document is 3% French, 95% English and 2% Italian).

5.5. Standard Evaluation Corpora

One of the criticisms of research in language identification to date is that there is still, despite several decades of research, no standard evaluation corpus on which the variety of systems developed can be reported and hence evenly compared. Such a corpus must necessarily be multilingual, but should also be representative of linguistic diversity. We propose therefore that in order to effectively evaluate existing language identification systems a corpus is required on which systems can be run and comparable results reported.

5.6. Performance Evaluation Criteria

The traditional information retrieval evaluation models dominate. While precision, recall and F-measure remain the metrics of choice, binarised relevance judgements in language identification (ie it is language X or it is not language X) are perhaps inappropriate owing to situations such as multilingual documents or for tasks such as coarse grained classification. We propose therefore that judgements should be made on a graduated scale, allowing for language identification to occur in a manner akin to the linguistic continuum, and that binary assessments should be compared to graduated assessments across a range of approaches and techniques to determine performance of language identification systems. Any such annotation schema should be to be able to handle multilingual documents more robustly as well as being able to express the relative degree of certainty about a given classification (not X or Y, not A or B, possibly C or D).

5.7. Effects of Preprocessing

It is common for text classification tasks construed within an IR paradigm to be preceded by activities such as stemming, stop word removal, case folding, and other kinds of normalisation. While most language identification research has not to date demonstrated this tendency, we propose that the preprocessing phase itself may be of considerable interest in the language identification task, particularly for low density languages. However, a wide range of experiments need to be conducted to determine the downstream effects of pre-processing on language identification performance.

5.8. Non-Roman Script / Multi-script

Almost without exception, work on written language identification to date has been focused on languages which are written using a romanised script, with any non-Roman script language identification being reduced to an encoding detection exercise. However this approach can be considered immature for a variety of reasons. Some languages can be rendered in more than one script (eg Uighur can be written in either the Cyrillic alphabet or native Mongolian script); some languages employ multiple scripts, with discrete roles for each (eg Japanese and the hiragana, katakana and kanji scripts); and some non-Roman scripts can be used to write multiple languages (eg the Cyrillic script can be used to write Russian, Uzbek and Macedonian). These orthographic issues represent new complexity in the written language identification task which has previously been largely ignored through selection of single script data for analysis.

5.9. Legacy and Non-Standard Encodings

While encoding detection has been considered as a partial solution to aspects of the written language identification task, there remain open issues. In the first place, issues similar to those discussed above in the context of languages with multiple scripts also pervade: there is no one to one relation between a language and an encoding. In addition, there are many written documents which use legacy encodings, or modified standard encodings which are not simply able to be accurately classified based on encoding alone. This issue is further complicated by the emergence of Unicode, where a single encoding may in fact be common, but inferences about the language expressed in that single encoding are even less valid.

5.10. Document vs Text Selection

Almost all previous work is performed at a document level. With the emergence of highly multilingual documents (either by design, or by accident eg through lexical borrowing) and more advanced document markup technologies, we propose that a new focus is required on language identification at a finer granularity within the document itself (at the level of character and/or word sequence extents).

5.11. Exploiting the Linguistic Content of Documents

To date, most work has assumed that understanding the semantic properties of a given document is not necessary

for language identification to be performed robustly. However, recent research has shown that the actual document content can be a useful additional resource to enhance discriminative capacity in language identification. There is the potential to exploit the occurrence of specific lexical items (eg named entities such as persons or places) within a given document text for greater precision in language identification task; this is particularly true for low-density language materials drawn from sources such as the web.

6. Conclusion

In conclusion, from even the high level survey undertaken earlier in this paper, it is clear that written language identification has been the topic of significant, and varied research over several decades. For the most part, this research activity has been separated from the written language resource creation and curation community on purely functional grounds: curated corpora have a fixed set of languages for which they are relevant. It is our view that addressing these open questions is of significant interest to the written language resources community owing to the increasing prevalence of highly multilingual resources and resources collated from open collections such as the web. We offer these challenges to the language resources and language technology communities, with an open invitation for collaboration.

7. References

J.A. Aslam and M. Frost. 2003. An information-theoretic measure for document similarity. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2003)*.

K.R. Beesley. 1988. Language identifier: A computer program for automatic natural-language identification of online text. In *Languages at Crossroads: Proceedings of the 29th Annual Conference of the American Translators Association*.

W. Cavnar and J.M. Trenkle. 1994. N-gram-based text categorization. In *Proceedings of Symposium on Document Analysis and Information Retrieval*.

M. Darnashek. 1995. Gauging similiarity with n-grams: Language independent categorization of text. *Science*, 267:843–848.

T. Dunning. 1994. Statistical identification of language. Technical Report CRL MCCS-94-273, Computing Research Lab, New Mexico State University, March 1994.

D. Elworthy. 1998. Language identification with confidence limits. In *Proceedings of the 6th Annual Workshop on Very Large Corpora*.

E. Giguet. 1995. Categorisation according to language: A step toward combining linguistic knowledge and statistical learning. In *Proceedings of the 4th International Workshop on Parsing Technologies*.

E. Giguet. 1996. The stakes of multilinguality: Multilingual text tokenisation in natural language diagnosis. In *Proceedings of the PRICAI Workshop on Future Issues for Multilingual Text Processing*.

E.M. Gold. 1967. Language identification in the limit. *Information and Control*, 5:447–474.

R.G. Gordon, Jr, editor. 2005. *Ethnologue: Languages of the World, Fifteenth Edition*. SIL International.

G. Grefenstette. 1995. Comparing two language identification schemes. In *Proceedings of Analisi Statistica dei Dati Testuali (JADT)*.

G-I. Kikui. 1996. Identifying the coding system and language of on-line documents on the internet. In *Proceedings of the 16th International Conference on Computational Linguistics (COLING'96)*.

C. Kruengkrai, P. Srichaivattana, V. Sornlertlamvanich, and H. Isahara. 2005. Language identification based on string kernels. In *Proceedings of the 5th International Symposium on Communications and Information Technologies (ISCIT-2005)*.

D-S. Lee, C.R. Nohl, and H.S. Baird. 1998. Language identification in complex, unoriented, and degraded document images. In *Document Analysis Systems*, chapter 2, pages 17–39. World Scientific.

R.D. Lins and P. Goncalves. 2004. Automatic language identification of written texts. In *Proceedings of the 2004 ACM Symposium on Applied Computing (SAC 2004)*.

H. Lodhi, J. Shawe-Taylor, N. Cristianini, and C.J.C.H. Watkins. 2002. Text classification using string kernels. *Journal of Machine Learning Research*, 2:419–444.

P. McNamee and J. Mayfield. 2004. Character N-gram Tokenization for European Language Text Retrieval. *Information Retrieval*, 7:73–97.

Y.K. Muthusamy and A.L. Spitz. 1996. Automatic language identification. In *State of the Art in Human Language Technology*, pages 273–285. Cambridge University Press.

Y.K. Muthusamy, R.A. Cole, and B.T. Oshika. 1992. The OGI multi-language telephone speech corpus. In *Proceedings of the 1992 International Conference on Spoken Language Processing (ICSLP 1992)*.

A. Poutsma. 2001. Applying Monte Carlo Techniques to Language Identification. In *Proceedings of Computational Linguistics in the Netherlands (CLIN) 2001*.

P. Sibun and J.C. Reynar. 1996. Language identification: Examining the issues. In *Proceedings of the 5th Symposium on Document Analysis and Information Retrieval*.

P. Sibun and A.L. Spitz. 1994. Language determination: Natural language processing from scanned document images. In *Proceedings of the Fourth ACL Conference on Applied Natural Language Processing (ANLP)*.

C. Souter, G. Churcher, J. Hayes, J. Hughes, and S. Johnson. 1994. Natural Language Identification using Corpus-based Models. *Hermes Journal of Linguistics*, 13:183–203.

O. Teytaud and R. Jalam. 2001. Kernel-based text categorization. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN'2001)*.

Acknowledgements

The research in this paper has been supported by the Australian Research Council through Special Initiative (E-Research), grant number SR0567353 "An Intelligent Search Infrastructure for Language Resources on the Web".