Multilingual Language Identification: ALTW 2010 Shared Task Dataset

Timothy Baldwin and Marco Lui
NICTA VRL
Department of Computer Science and Software Engineering
University of Melbourne, VIC 3010, Australia
tb@ldwin.net, saffsd@gmail.com

Abstract

While there has traditionally been strong interest in the task of monolingual language identification, research on multilingual language identification is underrepresented in the literature, partly due to a lack of standardised datasets. This paper describes an artificially-generated dataset for multilingual language identification, as used in the 2010 Australasian Language Technology Workshop shared task.

1 Introduction

Language identification is traditionally defined as the task of determining the unique language a given document is authored in, under the assumption that all documents are monolingual (Baldwin and Lui, to appear). In contexts such as the web, however, multilingual documents are commonplace, suggesting the need for language identification research to move towards a more realistic task setting where a document can be authored in one or more languages (Hughes et al., 2006). This paper describes such a dataset, based around the task of multilingual language identification, where the task is to determine which one or two languages a given document is authored in. This dataset formed the basis of the 2010 Australasian Language Technology Workshop shared task.

Multilingual language identification is relevant in a number of contexts. "Word spotting" of foreign words in multilingual documents has been shown to improve parsing performance (Alex et al., 2007), and multilingual language identification is a first step in this direction. It can also be used as part of a linguistic corpus creation pipeline for low-density languages, e.g. to determine the language used in interlinear glossed text (IGT) embedded in language documentation (Xia et al., 2009; Xia and Lewis, 2009).

The ideal vehicle for multilingual language identification research would be a dataset genuinely representative of the true multilingualism of resources such as the web. Creating such a resource, however, would require: (a) a multilingual crawl without language bias; and (b) a large-scale document collection with gold-standard annotations over the full range of languages extant on the web, including sub-document extents for the individual languages contained in a document. While we would ultimately like to generate such a dataset for general usage, in this paper we describe a more modest effort to artificially generate a dataset for multilingual language identification purposes. Our basic approach is to: (1) select a language bias-preserving set of primary documents; (2) select a comparable document for each in a second language based on translation links; and (3) concatenate sections of the two documents together to form a single multilingual document. In this paper, we detail the methodology for generating the dataset, and outline baseline and benchmark results over the dataset to calibrate future efforts.

2 Dataset Synthesis

The dataset for the task was prepared from database exports of the various language Wikipedias provided by the WikiMedia Foundation.[1] The WikiMedia Foundation carries out an ongoing export of the databases of each of the language-specific Wikipedias, and makes these exports available for download. The exports that we utilized are dated between 9 June 2008 and 1 August 2008. We downloaded all the Wikipedias that exceeded 1000 articles, which at the time numbered 75 (as of October 2010, this number is now almost 200). Of these, the file for the Spanish (es) Wikipedia failed to download correctly,

[1] http://download.wikimedia.org/backup-index.html

Timothy Baldwin and Marco Lui. 2010. Multilingual Language Identification: ALTW 2010 Shared Task Data. In *Proceedings of Australasian Language Technology Association Workshop*, pages 4−7

Lang code	Language name	No. Docs 1°	2°
af	Afrikaans	9	1
an	Aragonese	8	1
ar	Arabic	71	24
ast	Asturian	5	2
az	Azerbaijani	8	2
be	Belarusian	10	0
bg	Bulgarian	57	39
bn	Bengali	24	6
bpy	Bishnupriya	8	10
br	Breton	8	3
bs	Bosnian	26	4
ca	Catalan	105	62
ceb	Cebuano	15	0
cs	Czech	80	37
cy	Welsh	12	4
da	Danish	72	27
de	German	747	1327
el	Modern Greek (1453-)	31	7
en	English	3330	3774
et	Estonian	52	7
eu	Basque	19	2
fa	Persian	53	12
fi	Finnish	154	88
fr	French	747	1084
gl	Galician	27	3
he	Hebrew	122	83
hi	Hindi	22	2
hr	Croatian	43	10
ht	Haitian	11	0
hu	Hungarian	82	38
id	Indonesian	95	31
io	Ido	4	0
is	Icelandic	23	3
it	Italian	384	505
ja	Japanese	442	552
jv	Javanese	8	1
ka	Georgian	25	8
ko	Korean	72	26
ku	Kurdish	11	0
la	Latin	21	1
lb	Luxembourgish	18	1
lt	Lithuanian	57	12
lv	Latvian	19	5
mk	Macedonian	16	5
mr	Marathi	22	1
ms	Malay (macrolanguage)	35	9
nap	Neapolitan	13	0
nds	Low German	9	1
new	Newari	33	4
nl	Dutch	330	419
nn	Norwegian Nynorsk	37	9
no	Norwegian	156	80
oc	Occitan (post 1500)	15	1
pl	Polish	335	340
pms	Piemontese	11	0
pt	Portuguese	413	410
ro	Romanian	92	63
ru	Russian	376	437
scn	Sicilian	23	0
sh	Serbo-Croatian	21	9
sk	Slovak	61	17
sl	Slovenian	52	7
sq	Albanian	18	0
su	Sundanese	11	0
sv	Swedish	220	136
ta	Tamil	11	5
te	Telugu	27	6
th	Thai	50	21
tl	Tagalog	11	0
tr	Turkish	111	34
uk	Ukrainian	106	41
vi	Vietnamese	54	16
wa	Walloon	13	0
zh	Chinese	181	125

Table 1: Composition of primary (1°) and secondary (2°) documents in the dataset for each language (based on ISO-639 language codes).

leaving us with data in 74 languages. All of the data is UTF-8 encoded, and the total volume of uncompressed data is almost 60GB.

For this task, we were interested in presenting a language identification challenge over largely bilingual documents. We assumed that Wikipedia documents were all monolingual, and that the language they were written in corresponded exactly to the Wikipedia they were located in. On the basis of these assumptions, we set out to build bilingual documents by combining portions of monolingual documents. Each document in our dataset is compiled from two source documents, which we will refer to as "primary" and "secondary". In addition to making our documents bilingual, we were interested in maintaining semantic linkage between the sections of the document in different languages. We did this by taking advantage of the fact that many Wikipedia documents contain links to a comparable document in another language. For example, the English Wikipedia document on *Natural language processing* contains a link to the equivalent document in a variety of languages, including the Italian *Elaborazione del linguaggio naturale* and French *Traitement automatique du langage naturel*. The links are of the form [[<language-prefix>:<page title>]], and thus can easily be parsed with a regular expression. For purposes of elaboration, we shall refer to this kind of link as a language-link. It is important to note that the language-linked documents are not translations, they are comparable documents, on the same topic in different languages.

To construct each bilingual document, we first selected the language of the primary document via a roulette-wheel approach, weighted according to the relative distribution of the number of pages for each language Wikipedia. From there, we randomly sampled a document (without replacement) from the primary language Wikipedia. We then selected a secondary document from the set of language-links in the primary document via the same routlette-wheel approach, again weighted by the global distribution of the languages present.

To each source document, we applied simple regular expression-based normalisation to remove redirects, language links and templates. We also replaced intra-wiki links with the anchor text of the link. We then chunked each of the two source documents into paragraphs by splitting on two consecutive newline characters. We select the first half of the paragraphs from the primary document and the second half of the paragraphs from the secondary document (rounding up in each case), and concatenate them together to form a single document. For example, if the primary document contained 5 paragraphs and the secondary contained 8 paragraphs, we would select the first 3 paragraphs from the primary document, and the last 4 paragraphs from the secondary document. If either of these sections falls below 1000 bytes, we reject this primary–secondary pair and start over.

3 Dataset Characteristics

The dataset contains 10000 documents, separated into three partitions: 8000 for training, 1000 for development and 1000 for test. All except three of the documents are multilingual. These three documents are caused by anomalies in the Wikipedia data, in that the primary document contained a language-link to a document in the same language; in two of these cases, the primary document contained the same content under different identifiers. As a result, the same secondary document was selected for both, resulting in two documents with identical content in the final dataset.

The language distributions of the primary and secondary document components are as detailed in Table 1.

In addition to the raw documents and language annotations, we have also made available an evaluation script. The full dataset is available from `http://www.csse.unimelb.edu.au/research/lt/resources/altw2010-langid/`.

Baseline	$\mathcal{P}_M$	$\mathcal{R}_M$	$\mathcal{F}_M$	$\mathcal{P}_\mu$	$\mathcal{R}_\mu$	$\mathcal{F}_\mu$
en	.011	.015	.012	**.701**	.350	**.467**
en+de	**.014**	**.030**	**.018**	.458	**.458**	.458

Table 2: Results for the different baseline strategies over the development documents

4 Baseline Results

As each document has two languages associated with it, three different baselines can be considered:

best-1 monolingual: the single most common language

best-2 monolingual: the two most common languages

best-1 multilingual: the most common language pair

The results for the different strategies are presented in Table 2, as trained over the training documents and evaluated over the development documents. In our case, the two most common languages are `en` followed by `de`, and it also happens that the most common language pair is `en-de`. As such, our latter two baselines are identical in behaviour, and are presented together in the final row of the table. Based on the evaluation scripts made available as part of the dataset, we evaluate the models using micro-averaged precision ($\mathcal{P}_\mu$), recall ($\mathcal{R}_\mu$) and F-score ($\mathcal{F}_\mu$), as well as macro-averaged precision ($\mathcal{P}_M$), recall ($\mathcal{R}_M$) and F-score ($\mathcal{F}_M$). The micro-averaged scores indicate the average performance *per document*, while the macro-averaged scores indicate the average performance *per language*.

5 Benchmark Results

To provide a minimal benchmark, we consider a prototype-based classifier based on skew divergence, with the usual mixing parameter $\alpha = 0.99$, based on the findings of Baldwin and Lui (to appear). The prototype is calculated as the arithmetic mean across all instances for each feature. We deal with the multiple-language labelling using three different methods:

single: a single prototype is learned for each language; any document containing the language is used in the calculation of this prototype.

Tokenisation	**Multiclass**	$\mathcal{P}_M$	$\mathcal{R}_M$	$\mathcal{F}_M$	$\mathcal{P}_\mu$	$\mathcal{R}_\mu$	$\mathcal{F}_\mu$
unigram	single	.440	.274	.295	.264	.132	.176
bigram	single	.540	.376	.413	.583	.291	.389
trigram	single	**.564**	.412	.453	.814	.407	.543
unigram	stratified	.412	.458	.414	.629	.622	.625
bigram	stratified	.460	.448	.435	.775	.768	.771
trigram	stratified	.497	.467	**.464**	**.833**	.826	**.829**
unigram	binarised	.115	**.786**	.155	.057	.878	.107
bigram	binarised	.171	.705	.221	.114	.885	.202
trigram	binarised	.227	.686	.292	.259	**.903**	.402

Table 3: Results for the benchmark methods over the development documents, for a nearest prototype learner in combination with different tokenisation and multiclass handling strategies

stratified: a single prototype is learned for each language pair.

binarised: a pair of prototypes is learned for each language, one from documents containing the language, and the other from documents that do not contain the language; a classification for each language is produced via this binarisation.

We combine these three strategies with three tokenisation strategies, based on byte unigrams, bigrams or trigrams.

We present results over the development documents in Table 3. In the shared task, the primary evaluation measure was micro-averaged F-score ($\mathcal{F}_\mu$), on the basis of which the best-performing benchmark method is nearest prototype with skew divergence on the basis of byte trigram tokenisation, and with stratified multiclass handling.

6 Conclusion

This paper has described a multilingual language identification dataset, as used in the 2010 Australasian Language Technology Workshop shared task. We outlined the methodology for constructing the dataset from Wikipedias for different languages, and detailed results for a series of baseline and benchmark methods.

Acknowledgements

NICTA is funded by the Australian Government as represented by the Department of Broadband, Communications and the Digital Economy and the Australian Research Council through the ICT Centre of Excellence program. This research was supported in part by a Google Research Award.

References

Beatrice Alex, Amit Dubey, and Frank Keller. 2007. Using foreign inclusion detection to improve parsing performance. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning 2007 (EMNLP-CoNLL 2007)*, pages 151–160, Prague, Czech Republic.

Timothy Baldwin and Marco Lui. to appear. Language identification: The long and the short of the matter. In *Proceedings of Human Language Technologies: The 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL HLT 2010)*, Los Angeles, USA.

Baden Hughes, Timothy Baldwin, Steven Bird, Jeremy Nicholson, and Andrew MacKinlay. 2006. Reconsidering language identification for written language resources. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*, pages 485–488, Genoa, Italy.

Fei Xia and William Lewis. 2009. Applying NLP technologies to the collection and enrichment of language data on the web to aid linguistic research. In *Proceedings of the EACL 2009 Workshop on Language Technology and Resources for Cultural Heritage, Social Sciences, Humanities, and Education (LaTeCH – SHELT&R 2009)*, pages 51–59, Athens, Greece.

Fei Xia, William Lewis, and Hoifung Poon. 2009. Language ID in the context of harvesting language data off the web. In *Proceedings of the 12th Conference of the EACL (EACL 2009)*, pages 870–878, Athens, Greece.