Melbourne Language Technology Group Microblog Track Report

Bo Han, Marco Lui and **Timothy Baldwin**
NICTA Victoria Research Laboratory
Department of Computing and Software Systems
The University of Melbourne
hanb@student.unimelb.edu.au saffsd@gmail.com tb@ldwin.net

1 Introduction

This report outlines the TREC 2011 microblog track submission of the Language Technology Group at The University of Melbourne. The microblog track is an ad–hoc retrieval task over Twitter data with temporally-specified queries, and the requirement that all results must predate the query. Our objective is to establish baseline results for the task and study the relative impact of various factors on microblog retrieval. Twitter messages are authored in many different languages (Hong et al., 2011), but the queries were all monolingual English and assessors where instructed to base their judgements on only the English content of tweets. As such, we first conduct language identification to filter out non–English tweets (Baldwin and Lui, 2010). Next, we lexically-normalise tweets, to remove typos and phonetic substitutions, and deabbreviate common abbreviations (Han and Baldwin, 2011). Finally, we index the language-filtered, normalised documents using Indri,[1] apply dynamic lexical normalisation to the queries, and temporally filter the results relative to the query timestamp. Descriptions of each module in our system are presented in the following sections.

As we use language processing tools and a dictionary as part of the lexical normalisation, our submission is classified as making use of external evidence.

2 Language Identification

For language identification, we used `langid.py`, a language identification toolkit developed at The University of Melbourne (Lui and Baldwin, 2011).[2] `langid.py` combines a naive Bayes classifier with cross-domain feature selection to provide domain-independent language identification. It is available under a FOSS license as a stand-alone module pre-trained over 97 languages. In in-house evaluation over short text messages, we found that `langid.py` was much faster than competing automatic language identification systems without any loss in accuracy.

We apply `langid.py` to a combined crawl of 15,198,435 tweets based on the official crawling tool. `langid.py` returns a monolingual prediction of language content for a given document. All documents which are predicted to be non-English were removed from the dataset, resulting in 5,478,459 (putatively) English tweets.

3 Lexical Normalisation

Lexical normalisation is potentially relevant to the retrieval task, since noisy tokens are prevalent in microblogs, and tend not to be picked up on by standard token normalisation techniques such as stemming. Types of noisy tokens which we target in lexical normalisation are typos (e.g. *earthquak* "earthquake"), abbreviations (e.g. *lv* "love"), phonetic substitutions (e.g. *b4* "before") and vowel lengthening (e.g. *goooood* "good"). We suggest that lexical normalisation is particularly pertinent for recall, but note that the evaluation metric of choice for the microblog track is precision-based, meaning that its impact on our official results may be slight.

Ultimately, we are interested in performing context-sensitive lexical normalisation á la Han and Baldwin (2011). For the purposes of the microblog track, however, we chose to go for a high-precision, low-recall approach and avoid over–normalising correct unknown words. Therefore, we utilise the dictionary lookup method of Han and

[1] http://sourceforge.net/projects/lemur/

[2] http://www.csse.unimelb.edu.au/research/lt/resources/langid

Baldwin (2011) to substitute noisy tokens with high confidence (e.g. *u* to *you*). For our pre-filtered English tweets, 1,279,169 tokens were normalised, and 868,993 tweets were influenced by normalisation, indicating the prevalence of noisy tokens in microblog data.

4 Data Processing, Indexing and Querying

We index the data with Indri, based on the TREC data format. Each document contains a single text-based tweet, with timestamp in the form of an unsigned integer. For instance, *Jan 31 05:11:48* is mapped into *131051148*, where month is placed in the first digit, followed by a direct conversion of the date to digits. We preserve case information in the queries, and used the Indri Retrieval Model (Strohman et al., 2005) with TREC-format outputs. The results are ordered by timestamp in decreasing order before the query time. Any results which a timestamp later than the query are removed from the result set according to the task guidelines.[3]

5 Discussion

Precision@N was selected as the primary evaluation metric for the microblog task, using $N = 30$ to determine official results. We compare the precision of our method with and without both language identification and lexical normalisation in Figure 5, at different cutoff points in the result ranking.

The overall Precision@30 was 0.2565 (normalised) and 0.2571 (original) on English-only tweets. The four run settings yielded almost identical results for Precision@30. However, over fewer results, language identification improves precision by about 2% relative to the full document set. We also notice that lexical normalisation generally doesn't boost precision, and actually degrades precision slightly when using all tweets.[4] The only exception is at Precision@20, when lexical normalisation delivers marginally better results both with and without language identification. We further investigate these observations from the viewpoints of the retrieval model, language identification and normalisation.

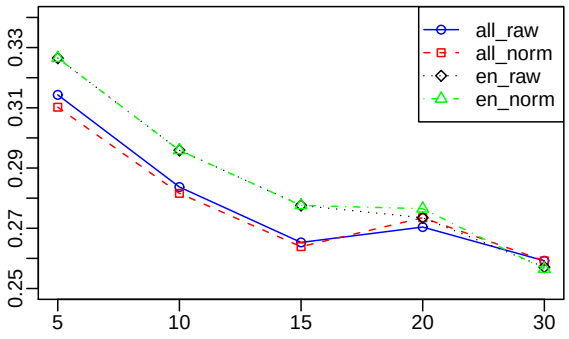

Figure 1: Average Precision@N across all topics for varying $N \in \{5, 10, 15, 20, 30\}$ and with different settings (all_raw = all tweets; all_norm = lexical normalisation only; en_raw = English tweets; en_norm = English tweets with lexical normalisation)

In our approach, we don't specifically tune the system parameters or perform query expansion, but instead consider the Indri Retrieval Model as a black box, combining factors such as a language model, term vectors and smoothing. As a result, we hypothesise that the setup is tweaked in favour of long documents, and that further improvements should be possible by tweaking the underlying IR engine. We confirm this hypothesis by manually checking the retrieval results. Over the English-only document collection, the highly-ranked documents are generally highly readable and well structured, regardless of relevance. For results on all tweets, some non-English tweets are present in the higher reaches of the document ranking, due to the occurrence of only one of the query terms. Language identification filters out most of these, but as N increases, more high-quality irrelevant tweets find their way into the results, pulling down precision.

We further manually compare results from the original and normalised English tweets at Precision@30 (hereafter denoted as en_raw and en_norm). We find that en_raw returns two additional relevant tweets for both topic 8 (`phone hacking British politicians`) and topic 14 (`release of The "Rite"`), while en_norm has four additional relevant tweets for topic 14 that are ranked in the top-20 results, which explains the prominence in the Figure 5 at Precision@20. While

[3] https://sites.google.com/site/microblogtrack/2011-guidelines

[4] Although less than we might expect given that the method is applied to all documents, the majority of which are non-English.

we hypothesised that lexical normalisation would predominantly impact on recall, there were isolated instances of it being able to reduce false positives in the retrieval results. For instance, in en_raw, some occurrences of *rite* are actually typos for *right*, but those tweets are selected as valid results. However, in en_norm, the standalone word *rite* is normalised to *right*, while quoted occurrences of *"rite"* are preserved as valid results.

In the future, we plan to enhance our lexical normalisation method to boost the ranking of poorly structured but relevant tweets, so that the document ranking is not dominated by high quality (but potentially irrelevant) tweets.

References

Timothy Baldwin and Marco Lui. 2010. Language identification: the long and the short of the matter. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 229–237, Los Angeles, USA.

Bo Han and Timothy Baldwin. 2011. Lexical normalisation of short text messages: makn sens a #twitter. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, pages 368–378, Portland, USA.

L. Hong, G. Convertino, and E. H. Chi. 2011. Language matters in Twitter: a large scale study. In *AAAI Conference on Weblogs and Social Media (ICWSM'11)*, pages 17–21, Barcelona; Spain.

Marco Lui and Timothy Baldwin. 2011. Cross-domain feature selection for language identification. In *Proceedings of the 5th International Joint Conference on Natural Language Processing (IJCNLP 2011)*, pages 553–561, Chiang Mai, Thailand.

Trevor Strohman, Donald Metzler, Howard Turtle, and W. Bruce Croft. 2005. Indri: a language-model based search engine for complex queries. In *Proceedings of the International Conference on Intelligent Analysis*.