Manuscript

Mapping international news

Evaluation of common lexicon-based and new dictionary-based methods for geographical classification of news texts

21/03/2015

This paper presents a new dictionary-based technique for geographical classification of news texts, which was developed to overcome shortcomings of the widely-used lexicon-based method. With this new technique, a very large geographical dictionary is automatically constructed by extracting words strongly associated with locations from a corpus of international news stories. Classification accuracy of both the dictionary-based and the lexicon-based techniques is evaluated using 5,000 human-coded news summaries to reveal the weakness and the strength of the two methods. The result shows that the lexicon-based classification is sufficiently accurate only in limited circumstances, while the dictionary-based classification is more accurate even when stories are very short or complex.

1 Problems

In geographical classification of textual data, use of manually compiled lexicons has been a common approach among the communication and international relations scholarship. Some researchers have created own lists of keywords that contain hundreds of names of countries and cities to identify geographical focus of international news stories (Blondheim, Segev, & Cabrera, 2015; Watanabe, 2013; Zuckerman, 2008). The Global Data on Events, Language and Tone (GDELT) system utilizes large databases maintained by the US government agencies,[1] which register millions of place names, to monitor occurrences of events across the world (Leetaru, 2012). These two types

[1] The United States National Geospatial-Intelligence Agency's GEOnet Names Server (NGA) and Geological Survey's Geographic Names Information System (GNIS)

of geographical lexicons are very different in size but sharing the same problem in social scientific application. Namely, these lexicons have no information about names of people and institutions that are related to particular places.

One can compile a geographical lexicon that also contains names of people and institutions, but it is very challenging, especially when the research is longitudinal and interested in multiple countries, because of the sheer number of key figures and institutions, and dynamics in association between names and locations: there are numerous influential public positions whose occupants change regularly; previously unknown groups of people suddenly emerge as security threats; companies merge with others and change names over night; people and organizations simply move from one country to another. Those who include names of people and institutions into geographical lexicons have to be always aware of those events and constantly update the words.

Supervised document classifiers, such as the Naïve Bayes filter, which can be seen as a type of automated dictionary construction techniques, are sometimes used for machine content analysis, but creation of training set is particular difficult when the number of potential classes are larger and units are not uniformly distributed across numerous classes. For instance, in the classic human-coded benchmark dataset, Reuters-21578, all the 21,578 documents fall into 175 location classes (countries), despite the fact that there are over two hundred countries in the world at the time. Supervised geographical classifiers never correctly classify stories about countries that are missing in the training set.

2 Solution

The solution to the lexicon-based geographical classification is creation of a dictionary comprised of a diverse set of proper nouns with varying class association scores. Such a dictionary can be constructed by a so-called lexicon expansion technique, in which a pre-defined lexicon and a large corpus of international news are employed to estimate words' association with countries. The advantages of this technique includes (1) the numbers of words in dictionaries are multiple times larger than manually compiled lexicons, (2) dictionaries are not only comprised of names of places but also of names people and organizations, (3) selection of words for dictionaries is fully automated and objective, (4) classifiers can identify the most relevant countries based on association scores, (5) dictionaries can be updated without human involvement, and (6) very larger training data can be used to extract words for infrequent classes.

2.1 Algorithm

The lexicon expansion technique is similar to supervised machine learning techniques, such as the naïve Bayes filter, in that computer programs estimate levels of association between words and classes using class labels given to documents in a training set, but it is substantially different in that the class labels are not assigned by human coders. In the lexicon expansion, assignment of documents into classes is performed by keywords in a pre-defined geographical lexicon.

2.1.1 Named-entity recognition

In the beginning of construction of geographical dictionaries, named-entity recognition is performed to identify proper nouns and eliminate all other types of words. In the current system developed by the author, named-entity recognition is based on capitalization of words in a news corpus i.e. frequencies of capitalization are compared with non-capitalization, and words are treated as proper nouns if capitalization is more frequent. However, such a simple method does not always function, because some names (multi-part names) are composed of multiple parts, which include very common words (e.g. New York, High Court and Geneva Motor Show). In order to handle multi-part names properly, a sophisticated mechanism that concatenates components of the names is utilized before the capitalization-based named-entity recognition.

The algorithm of the multi-part name concatenater is based on the Blaheta and Johnson’s (2001) phrasal verb identifier. The identifier estimates significance of n-tuples of binary variables by n-way interaction in the log-linear model. From the training corpus, all the sequences of capitalized words are extracted and assessed if they are multi-part names based on the level of statistical association.[2]

2.1.2 Geographical lexicon

A geographical lexicon, which contains names of 239 countries, and their major cities as well their demonyms, is created by the author.[3] For example, keywords registered to the lexicon for the United Kingdom are {UK, United Kingdom, Britain; British, Briton*, Brit*; London}. Similarly, keywords for Turkey and India are {Turkey; Turk*; Ankara, Istanbul} and {India; Indian*; Mumbai, New Delhi}. Names of cities in the lexicon are restricted to the capital and the largest cities, so the total number of keywords for all the 239 countries is 799 words, on average, only 3.3 words for one country.

2.1.3 Word scoring

Words extracted for the dictionary are given continuous scores indicating strength of association with classes (countries) using a very simple algorithm. First, individual text units are labelled using the keywords in the geographical lexicon; second frequencies of words are aggregated by the class labels. In the contingency table presented below, c_j is a country of interest and $\bar{c}_j$ is all other countries; w_i is the word for which scores are calculated and $\acute{w}_i$ is all other words; Fs are all raw frequency counts of words in respective classes.

	c_j	$\bar{c}_j$
w_i	F_{11}	F_{01}
$\acute{w}_i$	F_{10}	F_{00}
$w_i + \acute{w}_i$	$F_{1\cdot}$	$F_{0\cdot}$

The estimated score $\hat{s}$ of word $w_i$for a country c_j is calculated as association between w_i and c_j subtracted by association between w_i and $\bar{c}_j$:

[2] The threshold for the assessment is p<0.001. Concatenation of strongly associated words also increases independence of word occurrences, which is a usually-violated assumption of bag-of-words text analysis.

[3] The seed dictionary is made available online: http://koheiw.net/wp-content/uploads/2015/03/Newsmap_seed_v1.txt

$$\hat{s}_{ij} = log\frac{F_{11}}{F_{1\cdot}} - log\frac{F_{01}}{F_{0\cdot}}$$

2.1.4 Classification

Classification of texts is achieved by finding a country that gains the largest total scores $\hat{s}$ weighted by normalized frequency of word f_i in documents:

$$\hat{c} = \underset{j}{argmax} \sum_{i} \sum_{j} \hat{s}_{ij} f_i$$

3 Experiment

3.1 News corpus (training set)

In this experiment, a geographical dictionary was constructed from a corpus of online news texts. The author has been subscribing to Yahoo News US edition, which continuously supply news stories produced by international news agencies (mainly AP, AFP, and Reuters) via a RSS feed, and a total of 157,005 news texts were collected in 2014 (on average 430 items per day). The news stories are not full-text but contain both headings and lead sentences (on average, 32-words length). Use of news agency stories collected online is advantageous in making of geographical dictionaries, because (1) news agencies tend to cover wider range of countries than the retail media (Watanabe, 2013), and (2) subscription to RSS feeds allows us to sample stories without any pre-filtering.

The geographical dictionary is updated every day using the corpus of news texts collected on the present day and last 7 days to accurately estimate association between words and geographical locations. This approach is similar to the k-nearest neighbour algorithms in that estimates are local, but different in that training is retrospective. The length of training period is arbitrary, but small changes in the length have only little impact on the outcomes.

3.2 News stories (test set)

The geographical dictionary is utilized to classify a set of manually coded news stories to test its accuracy. This dataset is also comprised of news texts collected online in 2014, but they were sourced from different outlets: *The Times* (UK), *The New York Times*, *The Australian*, *The Nation* (Kenya), and *The Times of India*. From the collected news texts, a balanced sample of 5,000 was randomly taken and classified by human coders in terms to their geographic association. The dataset had to be this large because international news coverage typically follows a power-law distribution, in which internationally uninfluential countries appear only very infrequently. The motivation behind the choice of news sources was also to include counties that are under-represented in the western news media.

Manual coding of news stories was performed using an Oxford-based online recruiting platform, Prolific Academic.[4] The dataset was divided into 20 subsets, each containing 250 items, and participants were asked to choose countries most strongly associated with the news items focusing on the location of the events that the stories mainly concern (single-membership).[5] The coders' performance was constantly monitored using gold-standard answers created by the author and coded subsets that did not achieved more than 70% agree with the gold-standard were discarded. Eventually, the same items were coded by at least three different coders, and inter-coder agreement measured by Fleiss' multi-coder Kappa was $\kappa = 0.75$. After disagreement among coders was settled by the majority rule, coders' agreement with the gold standard becomes $\kappa = 0.88$. The main causes of the disagreement were (1) the difficulty in identifying the most strongly associated counties in international stories, and (2) coders' lack of knowledge about differences between countries with similar names (e.g. Congo Republic and The Democratic Republic of Congo). Imperfection of human coding imposes a ceiling on measured accuracy of the classifier even if the classification were perfect from experts' point of view, but their coding was treated as true answers in the experiment.

[4] https://prolificacademic.co.uk

[5] Regional and 'I do not know' categories were also allowed to use if necessary. The coding instruction is available online: http://koheiw.net/wp-content/uploads/2015/02/Newsmap_coding_04_online.pdf.

3.3 Measurements

Measurements of classification accuracy in this experiment were micro-average precision and recall,[6] which are the standard measurements in computer scientific literature, unless stated otherwise. Since classification was single membership, only the most strongly associated classes were taken into account, Yet, if items had the same level of association with more than one country, they were assigned into multiple classes, effectively increasing the total number of items in the output, to measure the indetermination of the classifier. Along with classification by the dictionary, a simple keyword matching classification was performed using the original geographical lexicon to construct a benchmark, which simulates the classification accuracy of earlier studies (Blondheim et al., 2015; Watanabe, 2013; Zuckerman, 2008).

Finally, the manually coded dataset was used in two ways: with headings and without headings. Headings were included to the test accuracy of the classification algorithm when the amount information given to the classifiers was exactly the same as to the human coders, but headings were removed to measure the accuracy when they were applied to classify less informative texts, in which fewer location indicators can be found.

3.4 Results

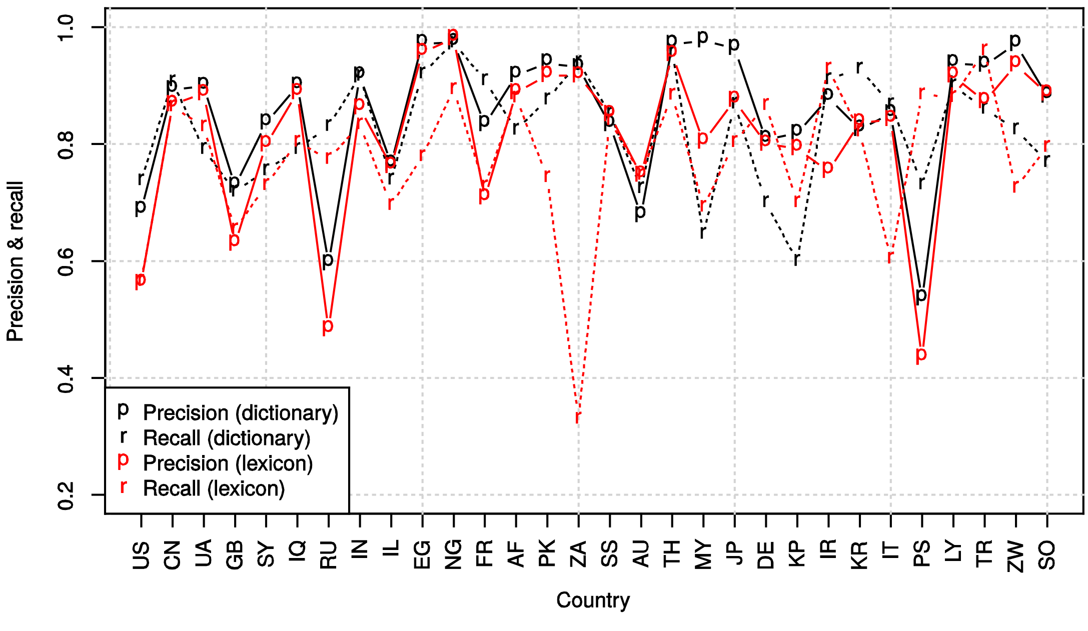
The result of the experiment shows that the overall classification accuracy of the dictionary was 0.82 in micro-average precision and 0.81 in micro-average recall, while they were 0.77 and 0.76 in the lexicon. Figure 1 clearly shows higher precision and recall of the dictionary in the top-30 most frequent countries, and it was outperforming the lexicon in most of the classes in both precision (25 classes; 83%) and recall (19 classes; 63%). The lexicon performed as good as the dictionary in many

[6] 'Precision' is an ability of classifiers to retrieve ONLY relevant items, while 'recall' is an ability to retrieval ALL the relevant items. There is usually a trade-off relationship between the two abilities, and a high precision often leads to a low recall, and vice-versa. Micro-average precision and recall are calculated by pooling of classification results of in all the classes.

classes, but its accuracy was sometime very poor: the lowest precision and recall were 0.43 and 0.33 in the lexicon, while they were 0.53 and 0.60 in the dictionary.

Nevertheless, both the lexicon and the dictionary shared the same tendency that their precision was low in texts associated with the United States (US), Russia (RU), and Palestine (PS). The low precision for the United States and Russia was explained by the greater chance that those influential states were covered in multi-national stories, which are more difficult to classify correctly. The poor performance in Palestine (PS) was due to stories about an unusual event (i.e. a suspected terrorist attack against Palestine embassy in Plague in Czech Republic), for which the classifier's failed to discover the association between the country and the city.

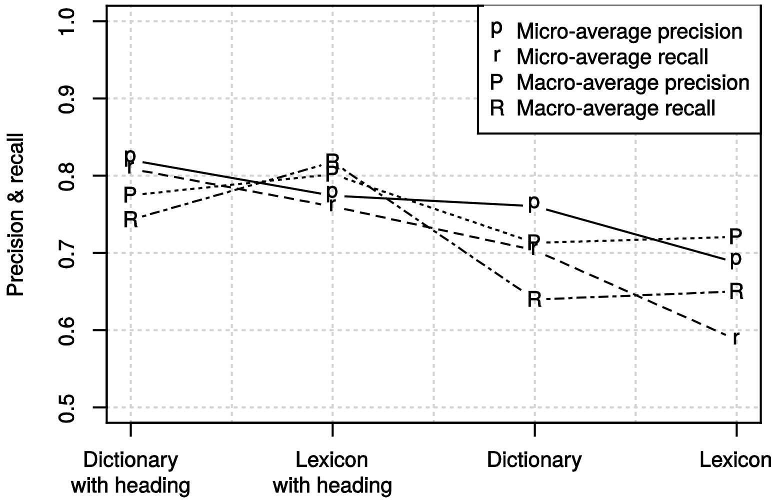
Figure 1: Recall and precision in top-30 most frequent classes



The difference in classification accuracy between the dictionary and the lexicon increases when the amount of information in news texts was reduced. When news texts were supplied with headings (on average 31.4 words), the difference in both micro-average precision and recall were 0.05, but the gaps expanded to 0.07 in precision and 0.12 in recall when headings were removed (on average 23.7

words). The higher macro-average precision and recall of the lexicon indicates its better performance in infrequent classes, in which names of countries and cities are often explicitly mentioned.

Figure 2: Micro and macro-average precision and recall



	p	r	P	R
Dictionary (with heading)	0.82	0.81	0.77	0.74
Lexicon (with heading)	0.77	0.76	0.80	0.82
Dictionary	0.76	0.70	0.71	0.64
Lexicon	0.69	0.59	0.72	0.65

The higher micro-average precision and recall of the dictionary can be partially explained by its larger vocabulary. The number of types of tokens in the lexicon was, on average, only 625, but it was as high as 4,010 in the dictionary. As illustrated in Table 1, while words in the lexicon were only the names of the countries and cities, the dictionary also contained names of British public figures such as BRITISH-PRIME-MINISTER-DAVID-CAMERON, RUPERT-MURDOCHS, KATE and ASSANGE for the United Kingdom as a result of lexicon expansion. The dictionary also included name of the British institutional actor, MI6. Similarly, words in the lexicon were only IRAQ, IRAQI, IRAQIS and BAGHDAD for Iraq, but names of cities (TIKRIT, KIRKUK and MOSUL) and important political figure (IRAQI-PRIME-MINISTER-NURI and SADDAM-HUSSEIN) were also added to the dictionary.

Table 1: A sample of entry of words registered to the dictionary

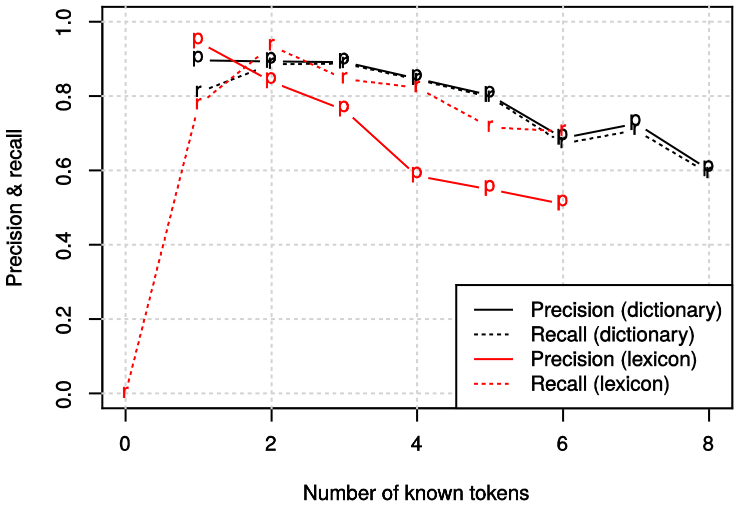
	UK	Score	Iraq	Score
1	LONDON	20.25	IRAQ	21.77
2	BRITISH	18.67	BAGHDAD	19.36
3	BRITAIN	18.55	IRAQI	18.99
4	BRITAINS	16.53	LEVANT	17.85
5	UK	16.11	IRAQS	17.79
6	ROLF-HARRIS	14.72	TIKRIT	15.99
7	REBEKAH-BROOKS	14.72	KIRKUK	15.47
8	BRITISH-PRIME-MINISTER-DAVID-CAMERON	14.72	ISLAMIC-CALIPHATE	15.06
9	ANDY-COULSON	14.50	MALIKI	14.42
10	SERENA-WILLIAMS	14.38	MOSUL	14.15
11	YPRES	14.38	PRIME-MINISTER-NURI	14.01
12	COULSON	14.11	IRAQIS	13.85
13	WATCH	14.11	ABU-BAKR	13.85
14	CENTRE-COURT	13.80	ARBIL	13.67
15	LVIV	13.80	SUNNIS	13.67
16	HAGUE	13.63	JEDDAH	13.48
17	FTSE	13.44	ALBU-KAMAL	13.48
18	WIMBLEDONS	13.23	HOLMES	13.48
19	KATE	13.23	SUNNI-ARAB	13.04
20	ASSANGE	13.23	BARZANI	13.04
21	DAVID-CAMERON	12.99	SAUDI-KING-ABDULLAH	12.77
22	BRITAINS-DAVID-CAMERON	12.99	IRAQI-PRIME-MINISTER-NURI	12.77
23	RUPERT-MURDOCHS	12.99	SIEG	12.77
24	WSI	12.99	IRAQI-TV	12.46
25	CAMERONS	12.99	SADDAM-HUSSEIN	12.46
26	LISICKI	12.72	ABU-GHRAIB	12.46
27	PARTON	12.72	SPA	12.10
28	PRIME-MINISTER-DAVID-CAMERON	12.72	RAHEEM	12.10
29	MARIA-SHARAPOVA	12.72	JARBA	12.10
30	FRENCH-OPEN	12.72	BAGHDADS	12.10
31	BRITISH-COLUMBIA	12.72	BAIJI	12.10
32	MI6	12.72	NED	12.10
33	BOUCHARD	12.42	ISRA	12.10
34	WONGA	12.42	IRAQI-PRIME-MINISTER-NOURI	12.10
35	VENUS-WILLIAMS	12.42	SUNNI-ISLAMIST	11.65
36	INDYK	12.42	JOHNSSON	11.65
37	SHARAPOVA	12.42	IRBIL	11.65
38	JOHN-SAWERS	12.42	SAMARRA	11.65
39	VENUS	12.42	SADR	11.65
40	MERS	12.42	NURI	11.65
41	GIBRALTAR	12.42	PRINCE-KHALED	11.65
42	BROOKS	12.42	RAQA	11.65
43	AINSLIE	12.42	ALARMING	11.65
44	LABOUR	12.42	MASSUD	11.65
45	EUROPEAN-COUNCIL	12.42	MUHAMMAD	11.65
46	POMFRET	12.05	MALIKIS	11.65
47	MURDOCH	12.05	FAO	11.65
48	MILIBAND	12.05	DZIADOSZ	11.65
49	UKS	12.05	NOURI	11.65
50	ACRON	12.05	CIA	11.65

The better classification accuracy of the dictionary was not only a product of the larger vocabulary, but also of continuous scores attached to the words, which permits subtle judgement in identifying the most strongly associated countries. This become clear when classification accuracy was measured separately by the number of known tokens in news texts (Figure 3).[7] The increase in known tokens

[7] ‘Known tokens’ refers to words included in the lexicon/dictionary

from 1 to 2 improves recall in both the dictionary and the lexicon, but the change decreases precision of the lexicon sharply, from 0.95 to 0.84. While deterioration in precision was consistent in the lexicon, reaching to 0.51 when the number became 6, classification accuracy remains relatively high until the end in the dictionary, because the dictionary-based classifier is able to identify the single most strongly associated country accurately by the continuous scores even when texts contain multiple location indicators.

Figure 3: Recall and precision by number of known tokens

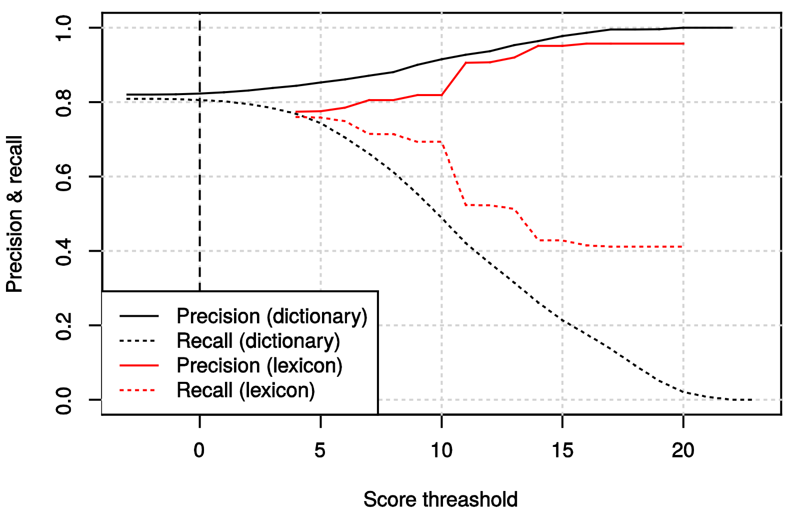


Classification of documents with continuous scoring also gives desirable properties to the dictionary-based method. As shown in Figure 4, its precision linearly improved as the score threshold increased. The scores also have a theoretical threshold of $\hat{s} = 0$, at which likelihoods that a text belong to class c_j or $\acute{c}_j$ becomes equal. This theoretical threshold can be used for exclusion of low-confident classification results and also for multi-membership classification.[8] The lexicon, however, has neither a linear relationship between scores and classification accuracy nor a theoretical threshold for high-confidence classification.[9]

[8] In multi-membership classification, if text units were given total scores greater than zero for multiple countries, they were associated with all those countries.

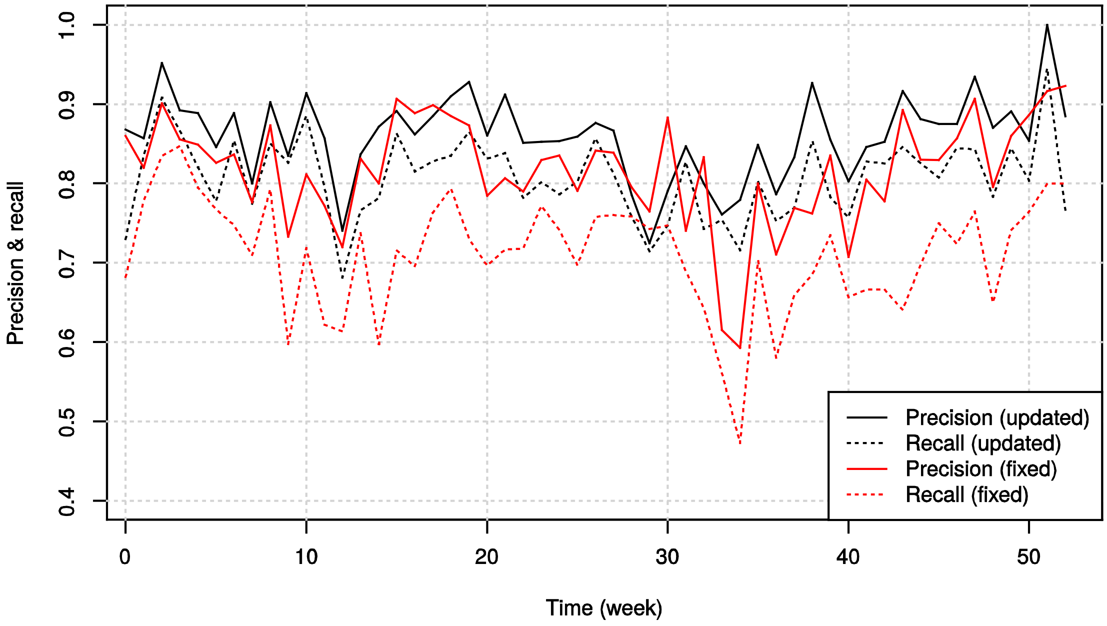
[9] In the lexicon-based classification, scores were normalized frequency of keywords (i.e. dividing the number of matched keywords by the total number of words in each text), that which range between 0 and 1. They were multiplied by a factor of 20 for comparison.

Figure 4: Recall and precision by score threshold



As already mentioned, the dictionary was updated every day throughout the period using news texts published on present and last 7 days to adapt to temporal changes in association between words and locations. The benefit of frequent updating of the dictionary is clear in Figure 5, where the black lines show weekly average precision and recall when the dictionary was updated every day, and red lines do when was not updated after the first day of the year. In the chart, we observe greater accuracy of the updated dictionary than the fixed dictionary throughout the year, which results in overall differences in precision and recall by 0.06 and 0.10 points.

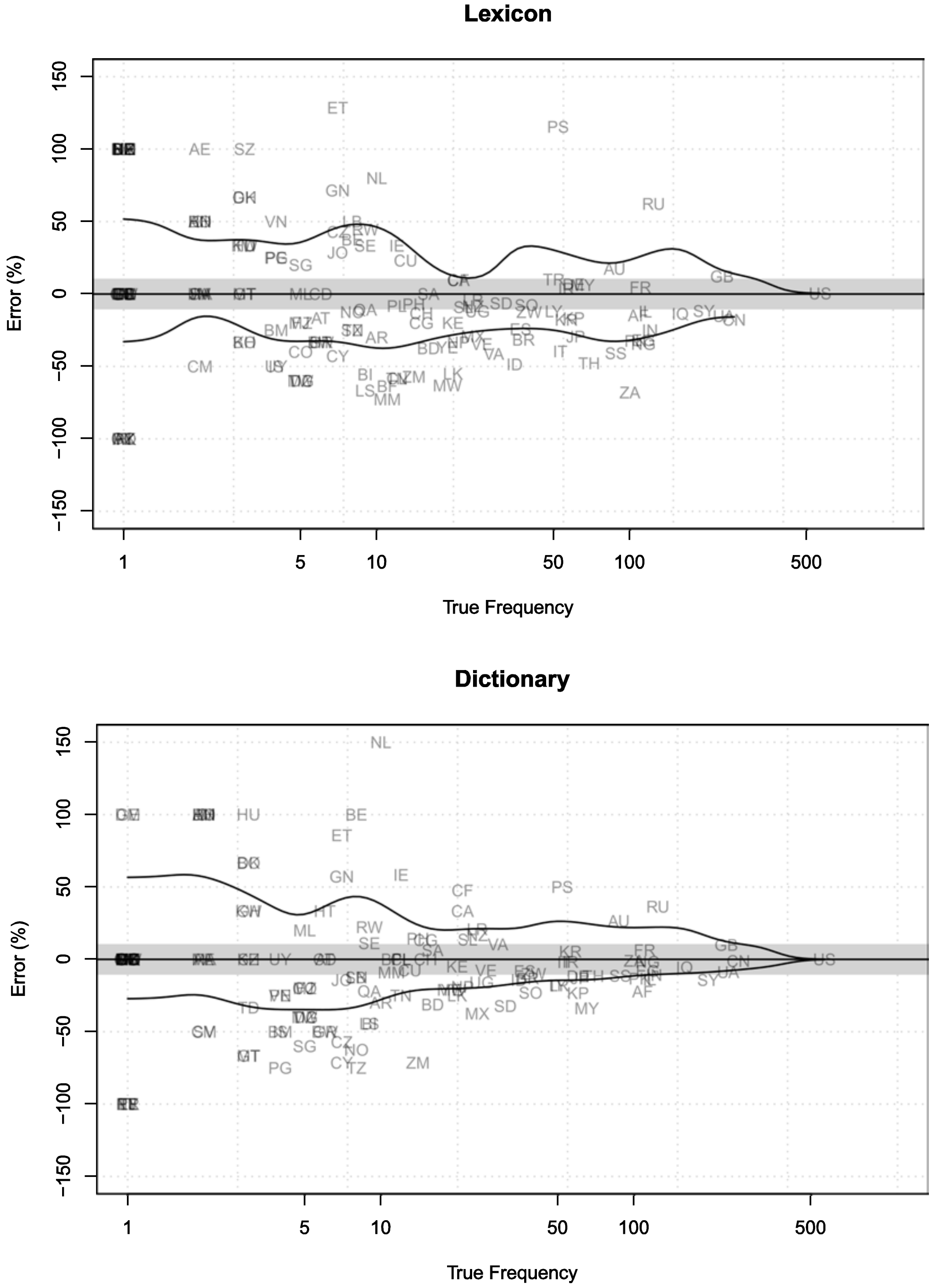
Figure 5: Classification accuracy of updated and fixed dictionaries through 2014



Finally, two plots in Figure 6 illustrates amount of errors in estimated class frequencies by the lexicon and the dictionary.[10] Curves in the plots are kernel-smoothed positive and negative errors, and grey bands represent ±10% error ranges. Overall, not surprisingly, errors in the low-frequency classes (true frequency < 10) were larger than the high-frequency classes in both the lexicon and the dictionary. Nonetheless, the lexicon has large negative errors, that were sometimes over-50% underestimation, across all classes as a result of its low recall; positive errors were limited to around 10% in many of the frequent class, but Palestine (PS) and Russia (RU) were highly overestimated. In contrast, the errors in the dictionary is smaller, particularly in the high frequency classes, ranging mostly within −25% and +50%.

[10] News texts do not contain headings, and items with score less than zero were filtered out from the data for the plot.

Figure 6: Errors in estimated class frequency by lexicon and dictionary



Lexicon

Dictionary

4 Discussion

The better accuracy of the dictionary-based classification in the experiment was achieved by two factors. First, the large vocabulary of the dictionary, which encompasses over 4,000 names of places, people and organizations, allowed the system to discover implicit location indicators in texts, consequently leading higher classification accuracy, in recall in particular. The benefit of the larger vocabulary become clear when news texts were information lean i.e. the deterioration in precision and recall after the removal of headings was smaller in the dictionary (0.6 and 1.1 points) than the lexicon (0.8 and 1.7 points).

Second, continuous scores attached to words in the dictionary allowed the classifier to estimate geographical association more precisely, which consequently improved its precision, especially in classification of complex texts with reference to multiple countries. It was shown in the experiment that the precision of the dictionary-based classification was only marginally affected by the greater numbers of known tokens, and its precision sustained above 0.8 until the number became five, while precision of the lexicon-based classification quickly deteriorated.

These strengths of the dictionary-based classification conversely highlight the weaknesses of the lexicon-based approach i.e. the method is able to perform accurate classification only when texts mention locations of events explicitly and they relate to a single location. Although use of lexicon-based classifiers in the earlier international news studies is justified as their documents were news summaries, the weaknesses of the approach found in the experiment limits its scope of application. That is, the lexicon-based approach is not suitable for classification of (a) full-text news stories, whose content is more complex and greater number of location indicators are found, and of (b) user-generated content (e.g. social media posts), which tend to contain fewer location indicators for its size and informal nature.

5 Conclusion

The comparison of the lexicon-based and dictionary-based geographical text classification techniques in this paper clearly showed that the common lexicon-based classifier is only suitable for classification of relatively simple texts that explicitly mention names of places, but the new dictionary-based classifier can perform the task accurately even when reference to locations is implicit and texts are complex. The difference between the two approach leads us to a conclusion that lexicon-based methods can be used for classification of news summaries, but dictionary-based classifier should be used for full-text news stories or social media posts.

Bibliography

Blaheta, D., & Johnson, M. (2001). Unsupervised Learning of Multi-Word Verbs. In *Proceeding of the Acl/Eacl 2001 Workshop on the Computational Extraction, Analysis and Exploitation of Collocations* (pp. 54–60).

Blondheim, M., Segev, E., & Cabrera, M.-Á. (2015). The Prominence of Weak Economies: Factors and Trends in Global News Coverage of Economic Crisis, 2009–2012. *International Journal of Communication*, *9*(0), 22.

Leetaru, K. H. (2012). Fulltext Geocoding Versus Spatial Metadata for Large Text Archives: Towards a Geographically Enriched Wikipedia. *D-Lib Magazine*, *18*(September/October). Retrieved from http://www.dlib.org/dlib/september12/leetaru/09leetaru.print.html

Watanabe, K. (2013). The Western perspective in Yahoo! News and Google News: Quantitative analysis of geographic coverage of online news. *International Communication Gazette*, *75*(2), 141–156. http://doi.org/10.1177/1748048512465546

Zuckerman, E. (2008). *International News: Bringing About the Golden Age*. The Berkman Center for Internet and Society, Harvard University. Retrieved from http://cyber.law.harvard.edu/sites/cyber.law.harvard.edu/files/International%20News_MR.pdf