Newsmap

Dictionary expansion technique for geographical classification of very short longitudinal texts

Kohei Watanabe
20/03/2015

Newsmap was created to classify short news texts according to their geographical association. The classification is performed by automatically constructing a large dictionary using a type of lexicon expansion technique based on a predefined dictionary that contain names of countries and cities and their demonyms. It aims to overcome the shortcoming of the widely-used simple keyword matching approach: simple keyword matching can achieve high precision in classification, but its recall tends to be very low due to the small number of entry words.

The dictionary expansion technique presented in this paper offers a way to overcome the weakness of simple keyword matching. Additional resource needed for this technique is only sizeable online news corpus, and machine learning is performed without costly manual supervision. The lack of human involvement in the lexical expansion allows us to update the dictionary frequently to adapt to longitudinal change in the text generation process.

Owing to the larger vocabulary, the expanded dictionary is able to classify subunits of documents (paragraphs and sentences) precisely without compromising recall. Such a subunit classification capability widen the range of application of this technique to analysis of other types of documents such as social media posts and diplomatic cables.

The discussion in this paper will be initiated by the statement of problems that this dictionary expansion technique aims to solve, and then followed by description of the algorithm and experiments using large human coded test data.

1 Problems

In geographical classification of documents, predefined lexicons have been commonly used. Researchers sometimes create own keyword dictionaries that contain hundreds of names of countries, and major cities (Blondheim, Segev, and Cabrera 2015; Watanabe 2013; Zuckerman 2008) to classify international news stories. Large databases of geographic information[1], called gazetteers, are maintained by government agencies such as the United States National Geospatial-Intelligence Agency's GEOnet Names Server (NGA) and Geological Survey’s Geographic Names Information System (GNIS), and used by the Global Data on Events, Language, and Tone (GDELT) system to monitor occurrence of events across the world.

Those two types of geographic lexicons are very different in size but sharing the same problem for

1 The Gazetters contain over 5 million location names (Leetaru 2012).

social scientific research: they only contain names of places and lacking other location indicators such as names of people or institutions. If those lexicons are used for geographical classificaiton, documents that are not always explicitly referring to places of names fall out from analysis. News stories usually contain explicit reference to names of places in the beginning, and the this problem is minimal, but it is more serious in analysis of less formal documents in which government officials and agencies are mentioned without reference to names of countries they are serving to.

Absence of non-place names in geographic lexicons can also be a problem in analysis of formal documents, when the unit of analysis is not whole documents but sentences are paragraphs, in which place-name location indicators less likely to occur. For example, the unit of sentiment analysis is usually sentence or paragraph, and exclusion of units that are not containing place names would reach different conclusions.

One can compile a geographical lexicon that contains non-place names, but it is extremely labour intensive, especially research projects are longitudinal and focusing on multiple countries, due to the sheer diversity of names of people or organizations, and dynamics in association between those names and locations: there are numerous important positions, and the officeholders change periodically or unexpectedly; previously unknown groups of people suddenly become important after significant events; well-established organizations cease to exist as a result of social changes; people and organizations simply move from one country to another. Researchers of the projects have to response to all those events and amend the lexicon to achieve high classification accuracy.

Supervised document classifiers, such as the Naive Bayes filter, is sometimes used to reduce the human involvement, and they can be seen as an automated method to construct geographical dictionaries. However, creation of training set is particular difficult when the number of potential classes are larger and units are not uniformly distributed across classes. For instance, in the classic human-coded benchmark dataset, Reuters-21578, all the 21,578 documents falls into 175 location classes (countries), despite the fact that there are over two-hundred countries in the world at the time. In supervised methods, classes do not appear in training data are never discovered in test data. Preclusion of low frequent classes may not be acceptable in social scientific research.

2 Solution

The solution to the absence of non-place-name location indicators in geographical dictionaries is lexicon expansion. In dictionary expansion, we utilize small pre-defined lexicon and a news corpus to extract named entities to construct a large dictionary for more accurate geographical classification. The dictionary is expanded based on co-occurrences of words in the pre-defined lexicon and named entities in the news corpus.

This approach solve the above-mentioned problems in creation of geographical dictionary: (1) the expanded dictionary contains multiple times more words than manually compiled original lexicon, (2) the expanded dictionary not only contain names of places but of names actors, (3) choice of words in the expanded dictionary is fully automated and objective, (4) the expanded dictionary can be updated without human involvement, (5) very larger training data can be used to extract words for all the possible classes.

2.1 Algorithm

The technique used for the dictionary expansion is very similar to supervised machine learning in that the algorithm estimated words association using class labels given to the documents in training data, but substantially different in that the classes are not assigned by human coders. The assignment of documents in the training data in classes is performed by the predefined lexicon that contains only names of countries and major cities, and demonyms. Since there is no human involvement in training, this process can be repeated for every single day during the research period to maximize the classification accuracy.

Since we are only interested in extracting names of places and actors, named-entity recognition was first performed in the tokenization stage and all the non-name tokens were removed. Note that the total number of words in document is therefore the total number of noun tokens in word frequency normalization.

2.1.1 Named-entity recognition

Named-entity recognition was performed simply based on capitalization of words. The frequency of words appearing with their first letter being capitalized was compared with their frequency being all lower-cased, and they are identified as nouns if capitalization is more frequent. However, such a simple method does not work with names constituted of more than one part (multi-part names). If one of the par of the name is a very common as general words (e.g. New York, High Court and Geneva Motor Show), it is separated from other parts and wrongly tokenized. To tokenize multi-part names properly, a more sophisticated concatenater that statistically estimates association between components of multi-part of names was implemented and utilized before the simple 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 strongly significantly ($p<0.001$) associated sequences are identified and concatenated as multi-part names. Such concatenation of strongly associated words also increases independence of word occurrences, which is a usually-violated assumption of bag-of-words text analyses.

2.1.2 Original lexicon

The original lexicon[2] to be expanded is created by the author and it contains names of countries and major cities of 239 countries and wildcard expressions that match their demonyms. For example, dictionary entries are {UK, United Kingdom, Britain; British, Briton*, Brit*; London} for the United Kingdom, {Turkey; Turk*; Ankara, Istanbul} for Turkey, and {India; Indian*; Mumbai, New Delhi} for India. The names of the cities can be added, but restricted to the capital or the largest cities of the respective countries. The total number of words in the lexicon is 799, averaging 3.3 words for each country.

2.1.3 Word scoring

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

This lexicon expansion algorithm not only extract words but assign scores for each word, and therefore its product is a called 'dictionary'. The word scoring is extremely simple: first, individual text units (news stories) are labelled by search for words in the original lexicon; second, by this labelling, word frequencies are obtained as shown in the contingency table presented below: c_j is a country of interest and $\acute{c}_i$ is all other countries, and 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 categories.

	c_j	$\acute{c}_i$
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 the association between w_i and c_j subtracted by the association between w_i and $\acute{c}_i$ measure by normalized frequency:

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

2.1.4 Classification

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

$$\hat{c} = \underset{j}{argmax} \sum_i \sum_j \hat{s}_{ij} f_i$$

3 Experiment

3.1 Training data

The expansion of manually original lexicon is achieved by utilizing rich and abundant online texts. Yahoo News US edition offers a large number of news stories produced by international news agencies (mainly AP, AFP, and Reuters) in RSS feeds. The author has been subscribing to the news feed and downloading the texts to a database since 2011. The total amount of news stories collected in 2014 was 157,005 items (430/day).

The advantages of using the news agency stories collected online for lexicon expansion are that (1) news agencies tend to cover wider range of countries (Watanabe 2013), and that (2) subscription to RSS feeds allows to sample stories without any filtering. The wide variety in the online news stories helps to construct a geographical dictionary that cover countries rarely mentioned. The stories are not full-texts, but contain headings and lead sentences and, on average, 32-words length.

The geographical dictionary was updated every day using the news stories collected on the present day and past 7 days to accurately estimate words' association with countries. This approach is similar to the k-nearest neighbours 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 do not affect the outcome.

3.2 Test data

Test data was also created using news stories collected online in 2014, but from different sources: *The Times* (UK), *The New York Times*, *The Australian*, *The Nation* (Kenya), and *The Times of India*. From the news stories, a balanced sample of 5,000 was randomly taken and classified accordingly to geographic association by human coders. The sample needed to be this large because international news coverage has a power-law distribution, in which internationally less-influential countries appear only very infrequently. The choice of sources was also meant to include as many countries as possible across regions.

The manual coding of the sample was performed using an Oxford-based online recruiting platform, Prolific Academic[3]. The sample was divided into 20 subset each containing 250 items. Participants were asked to choose countries most strongly associated with news items focusing on the location of the events that the stories concern[4]. Classification is single-membership and regional and 'I do not know' categories were also allowed to use if necessary. The coders performance were constantly monitored by the gold-standard answers created by the authors and subsets that less than 70% agree with the gold-standard were rejected. Eventually, same items were coded at least by three coders, and inter-coder agreement measured by Fleiss' multi-coder Kappa was $\kappa=0.75$. After disagreement among coders were settled by the majority rule, coders agreement with the gold standard was $\kappa=0.88$. The main causes of the disagreement were (1) the difficulty in identifying the most strongly associated counties in international stories, and (2) the lack of coders' knowledge about differences between countries with similar names (e.g. Congo Republic and The Democratic Republic of Congo). The imperfect human coding limit precision and recall to the level even if the machine classification is all correct from experts' point of view, but it was treated as true answers in the experiment.

1.1 Measurements

Measurements of classification accuracy were micro-average precision and recall, unless stated otherwise. The classification is single membership, and only the most strongly associated class was adopted. If an item is found to be associated with more than one country to the same extent, it is assigned to multiple classes effectively increasing the total number of items, because it allows to measure erroneous classification by the method.

Along with the classification by the expanded dictionary, simple keyword matching classification was performed using the original lexicon to create a benchmark, that simulate the classification accuracy in the previous studies (Blondheim, Segev, and Cabrera 2015; Watanabe 2013; Zuckerman 2008).

The test data was used in two ways: with headings and without headings. The headings are included

3 https://prolificacademic.co.uk/

4 The coding instruction is available online: http://koheiw.net/wp-content/uploads/2015/02/Newsmap_coding_04_online.pdf.

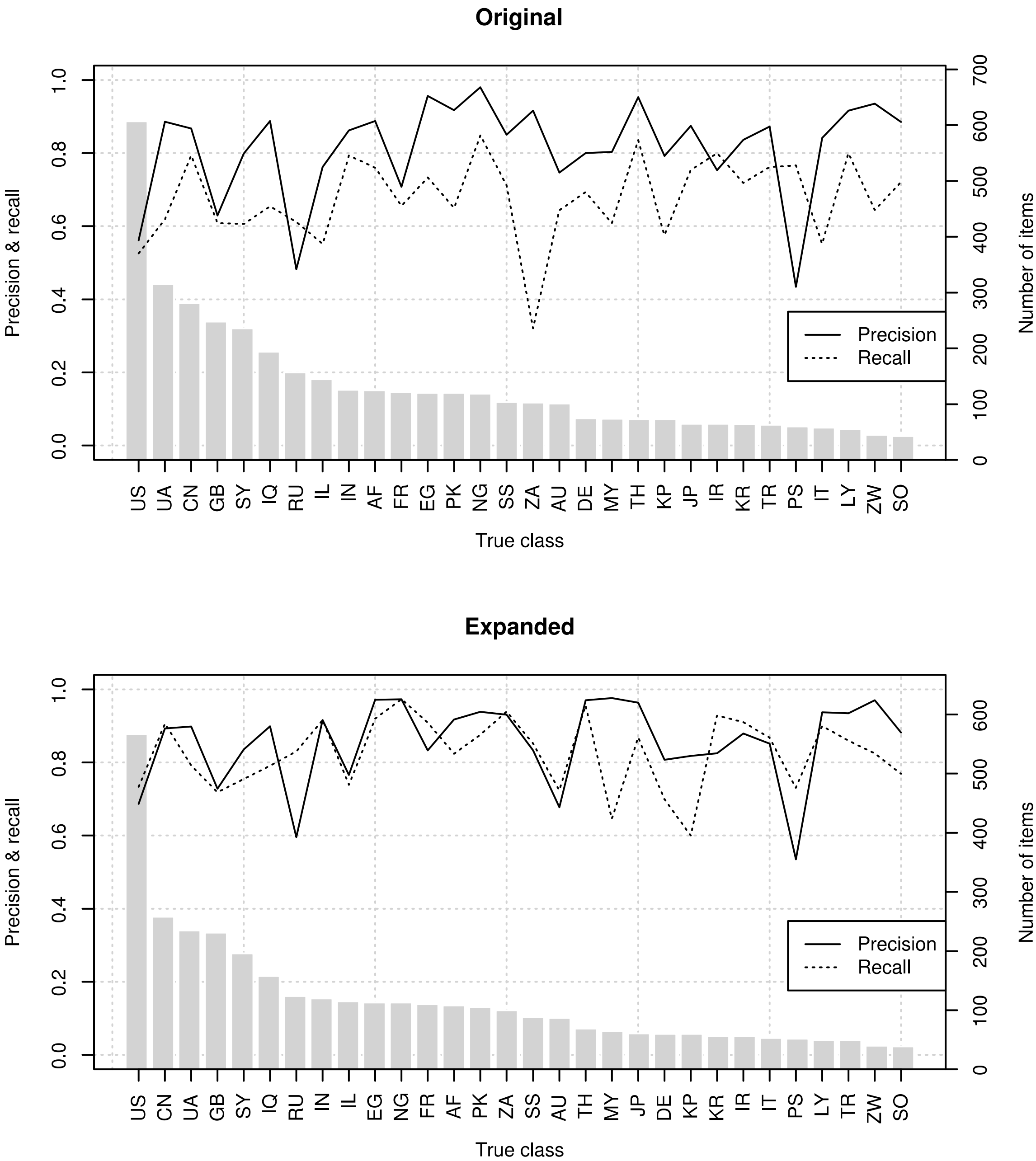
to test the performance of the classification algorithm when the same amount information is given as to human coders. The headings were removed, however, to show performance in more realistic research settings, in which sentence or paragraph level classification is needed.

3.3 Results

The overall performance of the classification accuracy of the expanded dictionary is 0.82 in micro-average precision and 0.80 in micro-average recall, while they were 0.77 and 0.67 in original lexicon. The precision is still high in the original lexicon, but recall is considerably lower. The charts in Figure 1[5] clearly show higher precision and recall of the expanded dictionary method in the top-30 most frequent classes when stories are given with headings. We can also observe higher stability in recall in the expanded dictionary: the lowest precision among the top-30 classes is Palestine (PS) in both cases, but score is 0.10 point higher in expanded dictionary (0.53) than the original lexicon; the lowest point of precision is 0.60 in North Korea (KP) in the former, while it is as low as 0.32 in South Africa (ZA) in the latter. Nevertheless, a similar tendency can be observed in the both methods that the precision is low in stories associated with the United States (US), and Russia (RU). This is can be explained by the greater chance that those internationally influential states are covered in relation to many other countries, and those international national stories are often more difficult to classify correctly. The low precision in both methods in Palestine (PS) is due to an unusual event that a suspected terrorist attack against Palestine embassy in Plague in Czech Republic, and the classifier failed to discover the association between the embassy and the city.

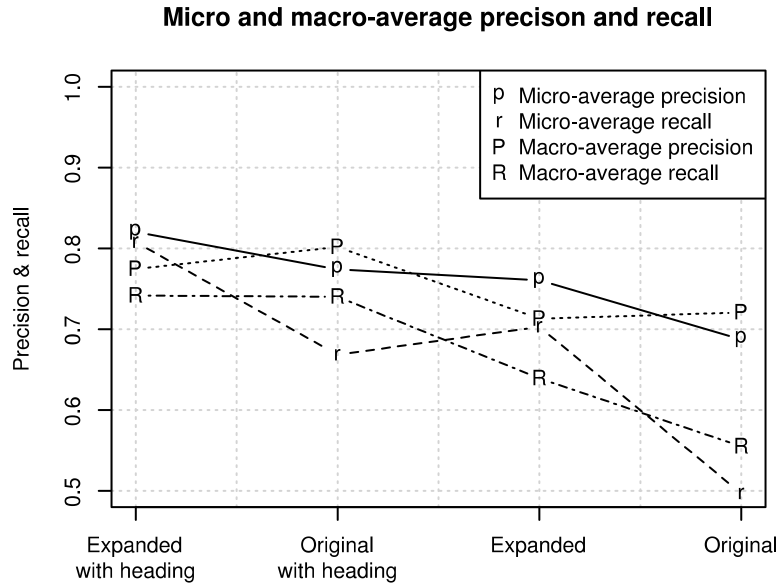
5 The difference in number of items is due to the multiple coding by the original dictionary.

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



The difference in performance between the original lexicon and expanded dictionary increases when the amount of information in each item is reduced. When the units contain headings (on average 31.4 words), the difference in micro-average precision and recall are 0.05 and 0.14, but it expands to 0.07 and 0.21 when headings are removed (on average 23.7 words). Macro-average precision and recall are better in the original lexicon when items contain headings, indicating its better performance in infrequenct classes than the expanded dictionary, but its macro-average recall also sharply falls to 0.56 when headings are omitted.

Figure 2: Performance in micro and macro-average



	p	r	P	R
Expanded (with heading)	0.82	0.81	0.77	0.74
Original (with heading)	0.77	0.67	0.80	0.74
Expanded	0.76	0.70	0.71	0.64
Original	0.69	0.50	0.72	0.56

The precision and recall of the expanded dictionary can be explained by its greater vocabulary. The number of types of tokens that match words in the original lexicon is, on average, only 625, but it is 4,010 in the expanded dictionary. Table 1 is an example of words in the expanded dictionary taken from July 1st. The words for the United Kingdom were only UNITED-KINGDOM, UK, BRITISH, BRITAIN, BRITAINS, BRITONS, BRITISH-BORN, BRITAIN-CHINA and LONDON in the original lexicon, but the expanded dictionary contains names of key figures such as BRITISH-PRIME-MINISTER-DAVID-CAMERON and RUPERT-MURDOCHS. It also includes name of the key British institutional actor, MI6, along with other high profile figures such as KATE and ASSANGE. For Iraq, words in the original lexicon were only IRAQ, IRAQI, IRAQIS and BAGHDAD, but the expanded dictionary contains names of cities such as TIKRIT, KIRKUK and MOSUL, and important political figure IRAQI-PRIME-MINISTER-NURI and SADDAM-HUSSEIN can also be found.

Table 1: Words in the expanded 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 larger number of entry words in the expanded dictionary, however, does not mean that the text units have to contain larger number of words that are also in the dictionary. The truth is quite the opposite as we can see in the Figure 3. For the plot, units without headings were classified by the expanded dictionary, but precision and recall reaches the peak levels (0.85 and 0.79) only with a single known token in each unit, and maintains the same level until the number of known tokens becomes four.

Figure 3: Number of known tokens

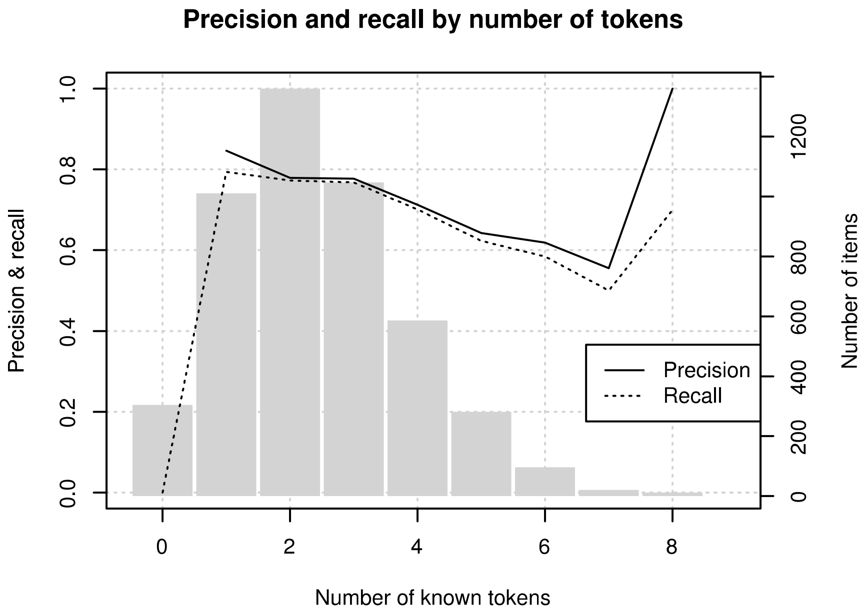
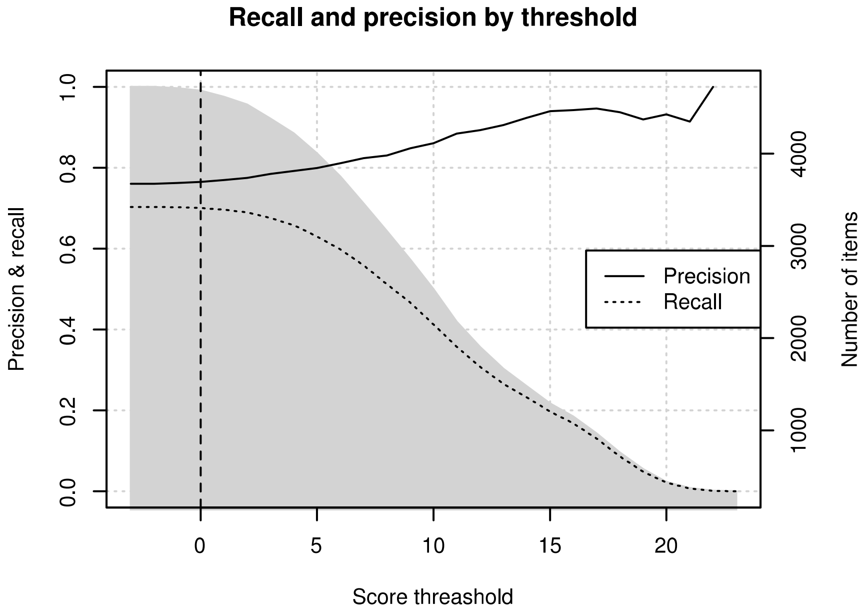


Figure 4: Distribution of item scores

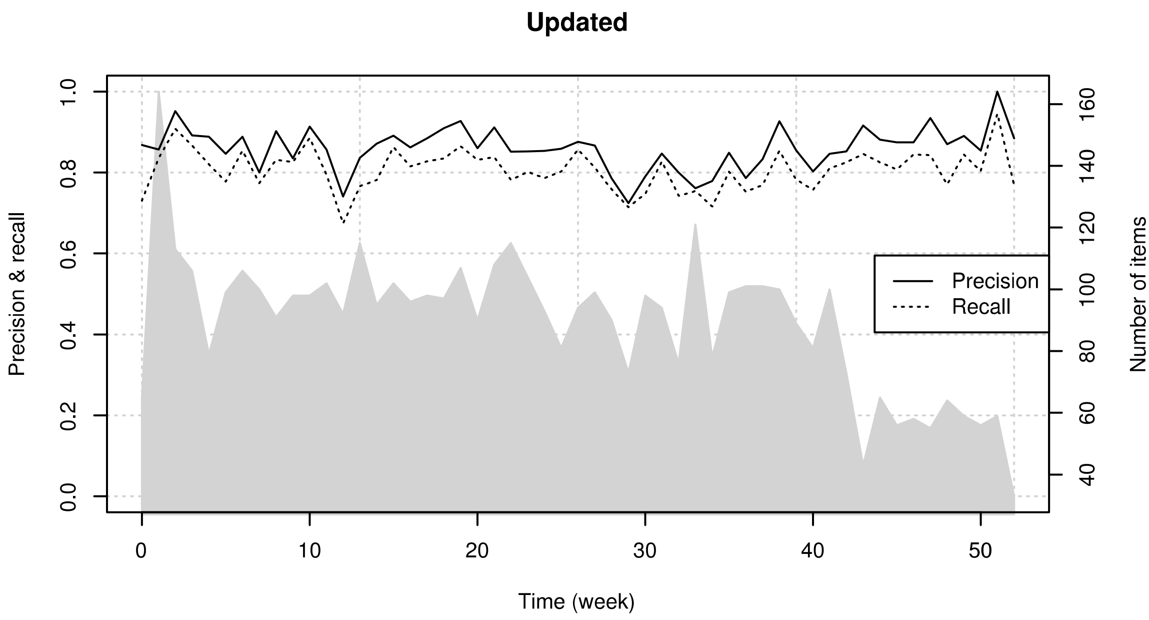


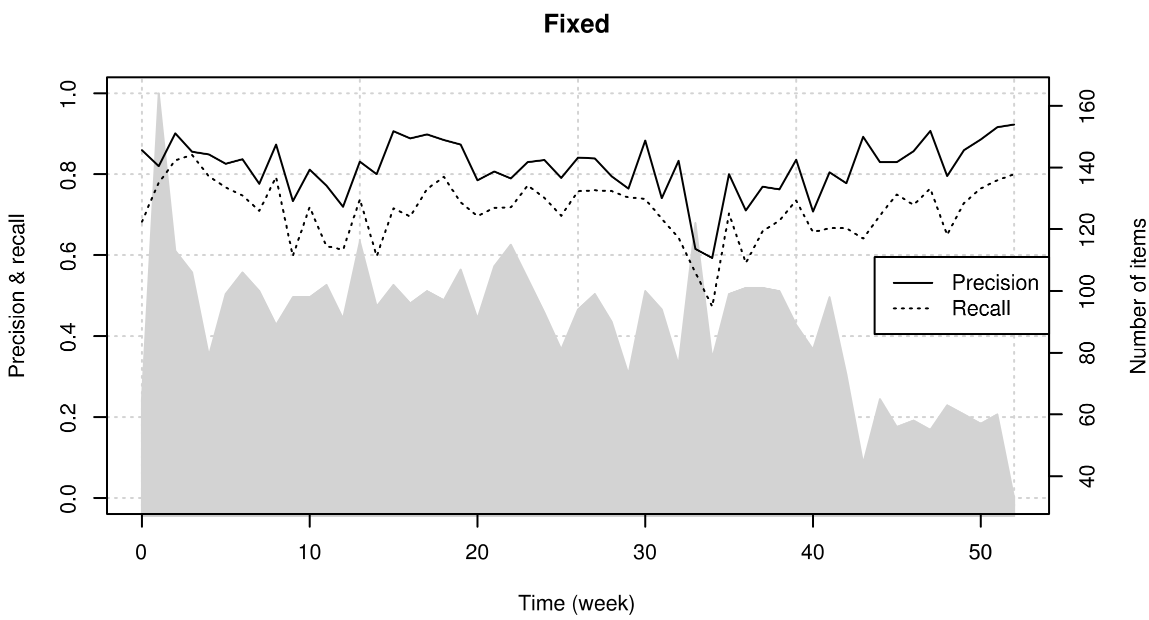
In addition to the high precision and recall, the relative frequency-based scoring grants desirable properties to the classifier. As shown in Figure 4, the precision almost linearly glows as the score threshold glows. The scores are also have a theoretical threshold of $\hat{s}=0$, at which likelihoods that texts belong to c_j and $\acute{c}_j$ becomes equal. This theoretical threshold can be used exclude low-confident classification.

The higher performance of the expanded dictionary is not only due to the size of vocabulary, but the temporal locality. As already explained, the dictionary was updated every day though the period using stories published on the present and the past 7 days to reflect that changes in association between named entities and the classes. This is particularly effective when the named entities are mobile actors, who travel between countries. The first plot in Figure 5 is showing weekly precision and recall of the expanded dictionary with daily update, and we can confirm consistency in its performance over weeks in 2014[6]. However, if the dictionary is created on the first day of the year, and never be updated, the performance considerably worsens to 0.77 in precision and 0.67 in recall. As shown in the second plot, fixed dictionary's recall falls to below 0.8 in 5 weeks, and never recover until the end of the year; both precision and recall plunge between week 30 and 40.

6 The decrease in the number of items in the end of the year was simply due to decline in number of items collected from the Kenyan newspaper.

Figure 5: Classification performance over weeks in 2014

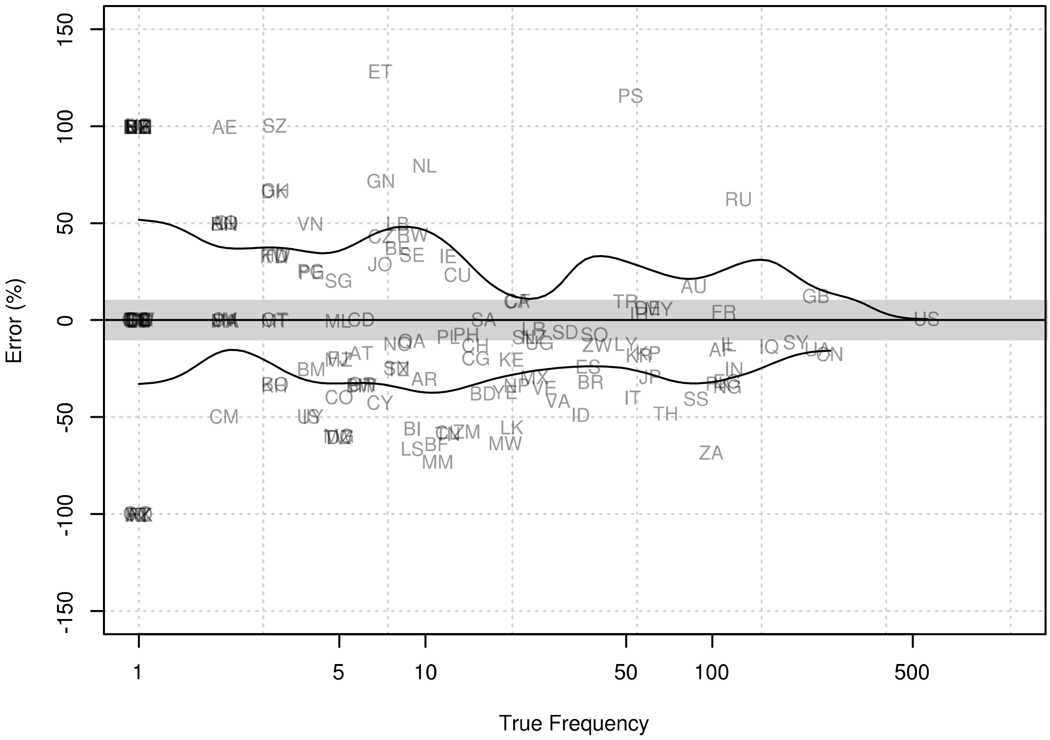




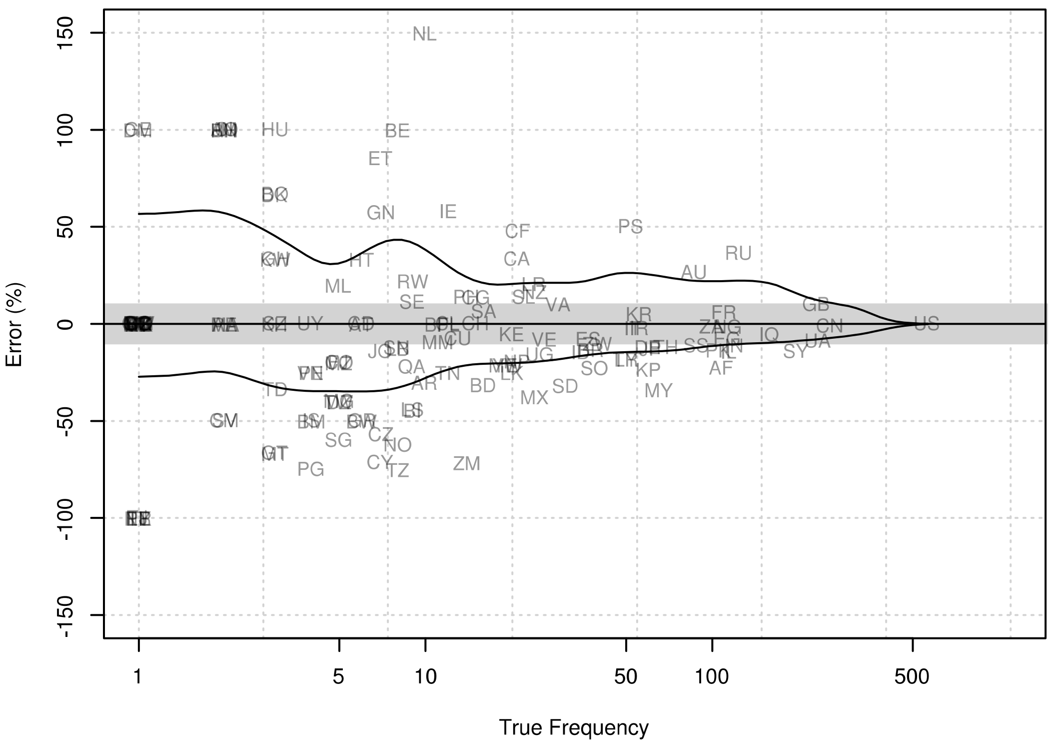
Finally, Figure 6 shows amount of errors in estimated class frequencies by the original lexicon and expanded dictionary. News stories are not containing headings, and items with score less than zero were filtered out for the plot. These plots are the more intuitive expression of the precision and recall, and low precision and recall results in over- and under-estimation of class frequencies. Curves in the plots are the kernel-smoothed positive and negative error, and grey rectangles are ±10% error range. Overall, not surprisingly, errors in the low-frequency classes (F<10) are larger than the high-frequency classes in both the original lexicon and the expanded dictionary. Nevertheless, the original lexicon has large negative errors across all classes; some of the countries are over −50%, and the negative average error is around −25%; positive errors are limited to around 10% in many of the frequent class, although Palestine (PS) and Russia (RU) are highly overestimated. In contrast, the negative errors in the expanded dictionary is smaller, particularly in the high frequency classes, and the ranges way less then −25%; overestimation can be found in Palestine and Russia in the expanded dictionary too, but it is +50% or less. Importantly, the errors in the expanded dictionary become consistently smaller as the frequencies of classes increase.

Figure 6: Errors in estimated class frequency

Original dictionary



Expanded dictionary



4 Discussion

The result of the experiments indicate that the accuracy of keyword matching is sufficiently high when text units contain headings (0.77 in precision and 0.67 in recall), but it performance significantly deteriorates when headings are removed (0.69 in precision and 0.50 in recall). This endorses previous researchers use of the keyword matching as an international news classification method, but strongly deny its capability to perform subunit-level (sentence or paragraph) classification.

However, the dictionary expansion technique presented in this paper achieved dramatic improvement to the level (0.76 in precision and 0.70 in recall) that is equivalent to keyword-based classification of units with headings. The increased performance, recall in particular, is owing to the greater of vocabulary the expanded dictionary (over 40,00 words) that encompasses not only the names of places but of people and institutions. This improvement paves the way to automated subunit-level geographical classification.

The experiments showed that the sufficient number of words in each unit for accurate classification by the expanded dictionary was only one. This means that the classifier can correctly retrieve 79% relevant documents and 85% of the documents are correctly classified only having one of the four thousand words in the expanded dictionary. This ability to classify information-lean documents implies wider application of this technique beyond news classification.

The expansion technique presented in this paper does not require complex algorithms. More importantly, the additional data required for the dictionary expansion technique was news stories collected online. The training data does not need to be tagged by human coders unlike common supervised machine learning classifiers, and eventually almost eliminate human involvement in training of the classifier.

The distribution of news stories geographic association has long-tail distribution, and some of the classes appear very infrequently, but the dictionary expansion technique minimizes preclusion of some of the infrequent classes by using very large training data (157,005 items). Although classification into rare classes tend to be inaccurate, the expanded dictionary constantly had words for all the 239 classes.

More technically, the expanded dictionary assign continuous scores to each item, and the threshold of $\hat{s}=0$ can be used to eliminate low-confidence items. This is not usually possible in simple keyword matching which only produces discreet scores based on number of matches. The threshold can be also used for multi-membership classification, although this type of classification was not the focus of this paper: if text units are given scores greater than zero for multiple classes, they are associated with all those classes.

Another desirable property of the frequency-based estimator is its consistency. The errors in classification by the expanded steadily decreases as the frequency of the class increases. This seems to be indicating that the estimator is asymptotically unbiased and the error in classification becomes closer to zero as the number of observation glows.

Finally, the key for the high performance is the frequent update of the dictionary and use of large training data. This could be archived by eliminating human involvement in training classifier.

Considering the rapid changes in the text generation process, and amount of the data available online today, dictionary expansion seems to be one of the most viable approaches in large longitudinal text analysis projects. Further, the technique presented in this paper might be also applied to other purposes such as thematic classification.

2 Conclusion

The dictionary expansion technique presented in this paper achieved substantial improvement in classification performance from simple keyword matching, although the additional resource for the higher classification accuracy is only a news corpus collected online. This means that the additional human and financial cost for the improvement in classification performance using this technique is negligible, and makes it one of the strongest candidate for analytical methods of large longitudinal text data.

Considering the small amount of information necessary for the accurate classification, the application of the technique is not limited to international news stories, but also other types of documents, such as social media or diplomatic cables. In all types of documents, the unit of classification can be paragraphs or sentences rather than whole documents. Shift from analysis of whole document to subunits allow us to measure mentions to individual entities, such as political leaders and government institutions by automated methods, and consequently substitute manual content analysis.

Given the simple algorithm, the dictionary expansion technique may find other applications such as thematic classification. This possibility was not explored in this paper, but it presumably less challenging than classification of very short texts into over two hundred possible classes.

3 Bibliography

Blaheta, Don, and Mark Johnson. 2001. 'Unsupervised Learning of Multi-Word Verbs'. In *Proceeding of the Acl/Eacl 2001 Workshop on the Computational Extraction, Analysis and Exploitation of Collocations*, 54–60.

Blondheim, Menahem, Elad Segev, and María-Ángeles Cabrera. 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, Kalev H. 2012. 'Fulltext Geocoding Versus Spatial Metadata for Large Text Archives: Towards a Geographically Enriched Wikipedia'. *D-Lib Magazine* 18 (September/October). http://www.dlib.org/dlib/september12/leetaru/09leetaru.print.html.

Watanabe, Kohei. 2013. 'The Western Perspective in Yahoo! News and Google News: Quantitative Analysis of Geographic Coverage of Online News'. *International Communication Gazette* 75 (2): 141–56. doi:10.1177/1748048512465546.

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