Kohei Watanabe

Newsmap: semi-supervised approach to geographical news classification

Article (Accepted version) (Refereed)

Original citation:
Watanabe, Kohei (2017) *Newsmap: semi-supervised approach to geographical news classification*. Digital Journalism . ISSN 2167-0811
DOI: 10.1080/21670811.2017.1293487

Newsmap: semi-supervised approach to geographical news classification

Kohei Watanabe

London School of Economics and Political Science

K.Watanabe1@lse.ac.uk

http://koheiw.net

FUNDING

This work was supported by the Murata Science Foundation under Research Grant [H26-020].

ABSTRACT

This paper presents the results of an evaluation of three different types to geographical news classification methods: (1) simple keyword matching, a popular method in media and communications research; (2) geographical information extraction systems equipped with named-entity recognition and place name disambiguation mechanisms (Open Calais and Geoparser.io); (3) semi-supervised machine learning classifier developed by the author (Newsmap). Newsmap substitutes manual coding of news stories with dictionary-based labelling in creation of large training sets to extracts large number of geographical words without human involvement, and it also identifies multi-word names to reduce the ambiguity of the geographical traits fully automatically. The evaluation of classification accuracy of the three types of methods against 5,000 human-coded news summaries reveals that Newsmap outperforms the geographical information extraction systems in overall accuracy, while the simple keyword matching suffers from ambiguity of place names in countries with ambiguous place names.

Newsmap: semi-supervised approach to geographical news classification

Keywords: content analysis; digital methods; international news; geographical classification; machine learning; digital methods

In recent years, there has been a growing interest in so-called 'digital methods' that aim to answer important questions in media studies by collecting and analysing data available on the internet (Rogers, 2013). In this context, social media such as Twitter, Facebook, and Instagram are attracting much attention, but the scope of digital methods is not limited to user-generated content; some of the researchers collected news stories online to study mainstream news media (Blondheim, Segev, & Cabrera, 2015; Watanabe, 2013; Zuckerman, 2003, 2008). Zuckerman, in his pioneering Global Attention Profile project, gauged news media's attention to foreign countries by automatically searching the websites of the *New York Times*, *Washington Post*, BBC, and CNN. More recently, Watanabe (2013) collected news stories from newspapers and online portals (Google News and Yahoo News) in India and the United States to detect cultural biases in the news portals run by American IT giants. Blondheim *et al* (2015) collected economic news stories from 35 online news sites between 2012 and 2015 to study changes in the prominence of certain countries in the world economy.

It is relatively easy to construct a very large dataset of news stories through machine-readable pages such as RSS (Rich Site Summary) feeds, and the number of people who collect news stories online for research purposes is expected to grow. In those studies, it is very common for the researchers to apply computer-assisted content analysis due to the size of the datasets, which are too large for manual content analysis. In fact, in the abovementioned studies, classification of news stories in terms of their geographical focus was performed by constructing a dictionary and searching for keywords in the documents. This is a widely-used method in computer-assisted content analysis in media studies, but there are at least two alternatives: geographical information extraction systems and machine learning techniques. Geographical information extraction systems are a set of natural language processing technologies that recognize entities (e.g. places, people and organizations) in documents to associate them with locations in knowledge databases. Machine learning is a data-driven approach to geographical classification of documents, where algorithms discover association between words and locations in training data.

In this paper, I will present the results of my systematic evaluation of the three approaches to geographical news classification with 5,000 manually coded news summaries to highlight their strengths and weaknesses. A keyword dictionary is constructed from existing sources (gazetteers), and information extraction systems are those publicly available through Web APIs (Open Calais and Geoparaser.io), but I have developed a new machine learning technique, Newsmap, to overcome the short comings of the other two methods.[1] In this new technique, a geographical classifier is constructed from a corpus of news stories with weak supervision given by a small manually compiled dictionary. The classifier recognizes not only names of places but people and organizations, and scores words according the levels of association with countries. Despite the minimal human input, this technique archives high classification accuracy, particularly in recall, thanks to the richness of information in the training data.

The result of the evaluation will show that the large dictionary created from gazetteers is confused by place names that appear in more than one country, most often in the United States and United Kingdom, and names of people that are also used in names of places. The

[1] R package is available at https://github.com/****/newsmap.

geographical information extraction systems, however, perform very well in term of precision but not in terms of recall, because their knowledge databases have only a limited amount of information on people and organizations associated with locations. Newsmap performs as well as geographical information extraction systems in precision, and outperforms these in recall, thanks to the large amount of information extracted from the training data.

The basic unit of geographical classification in this paper is nation for consistency with earlier studies of foreign news, although I am fully aware that this approach has been criticised for being 'methodological nationalism' (Wimmer & Glick Schiller, 2002). Focus on nations seems particularly odd in studies of international news because these studies are often motivated by today's rapid political, economic and cultural globalization, but I argue that nation is still a suitable unit for foreign news classification, because news stories often mention names of countries to locate foreign events. When we focus on regional or supra-national levels, results of the national-level classification can be aggregated into groups of countries.

Definition of Geographical Focus of News Stories

Although there are several studies on news coverage of foreign countries, definitions of geographical focus have not been clearly stated in the literature. Therefore, geographical focus in this research is defined as *locations of events or issues that stories are mainly concerned with,* as it seems the most widely agreed. Nonetheless, there are stories that cannot be classified based on this definition. In those cases, the following criteria are applied in order. If there are no events or issues in the story, classification is performed based on the main actors' association with countries; when news stories do not concern countries, they are classified as regions (Africa, America, Antarctic, Arctic, Asia, Europe, Oceania); when news stories have no information on location or no association with locations (e.g. stories on space science), they are treated as unclassifiable.

Challenges in Geographical Classification of News

Automated identification of places in documents has been studies in computer science, and it is becoming increasingly important for geographical information retrieval systems that interpret user queries and return information on specific locations (Buscaldi, 2011; Martins & Calado, 2010; Zaila & Montesi, 2015). The main challenge in the development of geographical information retrieval systems is the ambiguity of place names (or toponyms). For example, "nice" can be either Nice in France or an English adjective (geo/non-geo ambiguity), and "London" refer to either the UK's capital or a city in Canada's Ontario (geo/geo ambiguity). To solve these ambiguities, knowledge-based, map-based and data-driven methods have been developed for geographical information extraction systems (Zaila & Montesi, 2015). In the knowledge-based system, it is assumed that the places with larger population or physical areas are more likely to occur in documents. In the map-based disambiguation, places in closer proximity to unambiguous or disambiguated locations in the same document are chosen over others. Data-driven methods extract association between place names from data with or without human supervision. These techniques are utilized in geographical information extraction systems such as the Edinburgh Geoparser, CLAVIN, Open Calais and Geoparser.io.[2]

[2] The Edinburgh Geoparser: https://www.ltg.ed.ac.uk/software/geoparser/
CLAVIN: https://clavin.bericotechnologies.com
Open Calais: http://www.opencalais.com
Geoparser.io: https://geoparser.io

Nevertheless, the challenges in geographical classification of news stories are different from those in geographical information retrieval for at least three reasons. These differences demand development of a specialized system for geographical news classification, but also make creation of it relatively easy. First, systems should determine the single most relevant location to a document in classification tasks, while documents are given multiple geographical tags in information retrieval applications. Second, important geographical traits in news stories are not only names of placed, but also names of widely recognized people (e.g. Barak Obama) or organizations (e.g. the Pentagon), because locations associated with these actors are often not explicitly mentioned in news articles. Third, geographical traits in news stories are less ambiguous than other types of documents, because it is very unlikely that news worthy events to occur concurrently in places with the same; even if it happens, professional journalists distinguish between places very clearly in news articles (Smith & Crane, 2001).

With the absence of specialized tool, geographical news classification was often performed by simple keyword matching in earlier studies texts (Blondheim et al., 2015; Watanabe, 2013; Zuckerman, 2008), but it is difficult determine the most strongly associated counties based on word counting, particularly in short texts, and they usually lack information on people and organizations associated with particular locations, which also offer important information on stories geographical focus (Roberts, Bejan, & Harabagiu, 2010). One can create a geographical dictionary that also contains names of people and organizations, but it is challenging to maintain such a dictionary, especially when the research project spans a long period concerning multiple countries, because there are many key figures and organizations whose association with locations changes regularly. For example, occupiers of influential public offices change constantly; previously unknown groups or individuals suddenly attract the attention of the public; organizations merge with others and change names; people and organizations simply move from one country to another. Those who maintain large geographical dictionary must respond to all these occurrences.

Fully supervised machine learning techniques such as the naive Bayes classifier have been used for various document classification tasks, but creation of a training set is particularly difficult when the number of potential classes is larger and documents are not uniformly distributed across the classes. For instance, in the classic human-labelled benchmark dataset, Reuters-21578, all the 21,578 documents fell into 175 location classes (countries), even though there were 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 (Buscaldi, 2011).

Newsmap: semi-supervised geographical news classifier

Newsmap is a data-driven approach to geographical news classification. It constructs a classifier from a news corpus with weak supervision given through a manually complied small dictionary. This “semi-supervision” frees us from the burden of manually classifying thousands of news stories to train a model. The advantages of this technique are the following: (1) the classifier recognizes not only names of places, but also names of people and organizations, (2) the classifier can identify the most relevant countries based on continuous scores attached to words in the model, (3) the classifier can be constructed and updated with minimal human involvement, and (4) very large training data can be used to train the classifier for infrequent classes.

This section explains the procedure to construct a Newsmap classifier in detail. The examples presented to enhance readers’ understanding are taken from the experiment. The

training set is news summaries collected from Yahoo News website in 2014 as detailed in the beginning of the Experiment section.

Feature selection

Newsmap identifies proper names solely based on capitalization of words following the earlier work (Smith & Crane, 2001; Wacholder, Ravin, & Choi, 1997). This method is not comparable to syntactical named-entity recognition, but the purpose here is to select features that the classifier should extract from the corpus. This simple approach works sufficiently well with single-word names, but identification of multi-word names requires (e.g. New York or Prime Minister David Cameron) a different mechanism. Wacholder et al (1997) created a system to automatically identify multi-word names in the corpus based on pre-defined rules, which requires considerable manual inputs, but it is based on statistical estimation of association between words in Newsmap, exploiting today's greater computational capacity.

Newsmap identifies multi-word names by estimating strength of contiguous collocations of capitalized words based on an algorithm proposed by Blaheta and Johnson's (2001). The system first extract all the sequences capitalized words (8,321 unique combinations were found in the training corpus) from the training corpus.[3] Then, it performs pair-wise comparison of sequences in terms of the occurrences of the same words in the same positions. Here an absence of trailing capitalized words (show by □ signs), which suggests the combinations are semantically complete, are also considered as a match. For example, "British Prime Minister David Cameron" and "British Prime Minister Tony Blair" are sequences of n = 6 words and have m = 4 matches:

Sequence 1	British	Prime	Minister	David	Cameron	□
Sequence 2	British	Prime	Minister	Tony	Blair	□
Match	TRUE	TRUE	TRUE	FALSE	FALSE	TRUE

In the next example, where the lengths of sequences are different, the number of matches is m = 2:

Sequence 1	British	Prime	Minister	David	Cameron	□
Sequence 2	British	Airways	□	□	□	□
Match	TRUE	FALSE	FALSE	FALSE	FALSE	TRUE

After comparing the sequences with all other sequences, we obtain odds ratios λ of perfect or no matches over partial matches in a log-linear model for each of the sequences:

$$\lambda = (n-1)\log C_0 - \sum_{m=1}^{n-1} \log C_m + \log C_n$$

where C_m denotes the number of sequences that have m matches. To test statistical significance of the odds ratios, their standard errors are obtained by taking squares of the sum of inverse of the counts:

[3] Sequences that appear less than 10 times and more than five word long are ignored to limit the number.

$$\sigma = \sqrt{\frac{(n-1)^2}{C_0} + \sum_{m-1}^{n-1} \frac{1}{C_m} + \frac{1}{C_n}}$$

The threshold for the statistical significance is set to z-score $\lambda/\sigma > 3.09$, which is a 99.9% confidence level. Continuing with the above example, "British Prime Minister David Cameron" was found insignificant in this test due to occurrences of "British" in many other sequences, but "Prime Minister David Cameron" was found significant.

Seed dictionary

In Newsmap, seed dictionary is the only manual input to the system, and serves as semi-supervision. The seed dictionary that I created contains names of 239 countries and their major cities, as well as their demonyms.[4] For example, the keywords registered to the seed dictionary for Ukraine and Iraq are only {Ukraine, Ukrainian*, Kiev} and {Iraq, Iraqi*, Baghdad}. Names of cities in the seed dictionary are restricted to the capital and the largest cities; therefore, the total number of keywords for all 239 countries is 800 words, on average, only 3.3 words per country.

Word scoring

Newsmap calculates associations scores of words solely based on co-occurrences of words, therefore not requires costly syntactical analysis of a large corpus. Firstly, the system searches individual documents for keywords in the seed dictionary (simple keyword matching) and gives them class labels (countries); secondly, the system aggregates the frequency of words according to the class labels to create contingency tables. 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 the association between w_i and c_j subtracted by the association between w_i and $\bar{c}_j$:

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

Table 1 shows scores given to the words most strongly associated with Ukraine and Iraq. The keywords for Ukraine in the seed dictionary only match "Ukraine", "Ukrainian" and "Kiev", but many new words are identified based on cooccurrences: Mariupol, Lugansk and Slovyansk are small but important cities in the Ukraine; President Viktor Yanukovich, Prime Minister Arseny Yatseniuk and President Petro Poroshenko are the leaders of the country before and after the Euromaidan revolution in early this year. Similarity, the

[4] The seed dictionary is available at https://github.com/***/***

keywords in the seed dictionary for Iraq only match “Iraq”, “Baghdad”, “Iraqi” and “Iraqis”, but the classifier discovered names of smaller cities (Anbar, Ramadi, Fallujah, Kirkuk and Tikrit, including their spelling variants), political leaders (Saddam Hussein and Prime Minister Nuri) and key ethnic groups (Iraqi Kurds and Iraq’s Sunni).

Table 1: Top 50 features for Ukraine and Iraq in Newsmap

	Ukraine	Score	Iraq	Score
1	Ukraine	11.84	Iraq	11.58
2	Ukrainian	10.36	Baghdad	10.56
3	Kiev	10.34	Iraqi	10.39
4	Ukrainians	7.94	Iraqis	8.15
5	Ukrainian President Petro Poroshenko	7.64	Anbar	8.14
6	Mariupol	7.15	Ramadi	7.55
7	President Petro Poroshenko	6.94	Fallujah	7.51
8	Prime Minister Arseny Yatseniuk	6.92	Iraqi Kurdish	7.50
9	Natalia Zinets	6.84	Kirkuk	7.42
10	Lugansk	6.72	Tikrit	7.36
11	Pavel Polityuk	6.71	Falluja	7.32
12	Donetsk Ukraine	6.63	Maliki’s	7.25
13	Slovyansk	6.61	Arbil	7.20
14	Ukrainian President Viktor Yanukovich	6.54	Iraqi Kurdistan	7.12
15	Slaviansk	6.48	Iraqi Kurds	7.08
16	Richard Balmforth Kiev	6.46	Prime Minister Nuri	6.97
17	Petro Poroshenko	6.38	Irbil	6.90
18	Crimean Tatars	6.36	Isabel Coles	6.74
19	Kharkiv	6.12	Iraqi Prime Minister Nuri	6.73
20	Yanukovich	6.05	Baiji	6.71
21	President Viktor Yanukovich	6.04	Samarra	6.68
22	Slavyansk	6.03	Amerli	6.56
23	Andriy Lysenko	5.98	Levant	6.54
24	Andriy	5.98	Iraq PM	6.53
25	Ukraine PM	5.98	Sistani	6.51
26	Hrabove	5.97	Raheem Salman	6.47
27	Kramatorsk	5.89	Diyala	6.40
28	Pavel Polityuk Kiev	5.86	Baquba	6.32
29	Kiev’s	5.85	Sinjar	6.29
30	Poroshenko’s	5.84	Arbil Iraq	6.15
31	Ukrainian President Viktor Yanukovych	5.82	Sadr	6.15
32	Poroshenko	5.81	Mosul Dam	6.15
33	Natalia Zinets Kiev	5.80	Iraq’s Sunni	6.13
34	Tatars	5.75	Baghdad’s	6.11
35	Luhansk	5.75	Jurf	6.10
36	Donetsk	5.74	KRG	6.08
37	Vitaly Klitschko	5.73	Mosul	6.03
38	Ukraine’s Poroshenko	5.68	Saddam Hussein	6.03
39	Right Sector	5.64	Basra	6.00
40	Richard Balmforth	5.64	Iraqi PM	6.00
41	Volodymyr	5.62	Ahmed Rasheed	6.00
42	President Viktor Yanukovich’s	5.59	Raheem Salman Baghdad	5.97
43	Anton Zverev	5.57	Grand Ayatollah Ali	5.94
44	Oettinger	5.54	Kurdistan Regional Government	5.91
45	Serhiy	5.52	Ned Parker	5.91
46	Interior Minister Arsen Avakov	5.49	Ahmed Rasheed Baghdad	5.91
47	East Ukraine	5.49	Michael Georgy Baghdad	5.88
48	Vyacheslav Ponomaryov	5.43	Maliki	5.85
49	Donetsk People’s Republic	5.43	Sinjar Mountain	5.82

50	Ukraine's Moscow	5.40	Sadr City	5.76

Classification

Newsmap predicts countries most strongly associated with documents in the classification stage simply by finding a country that yields the largest total scores $\hat{s}$ weighted by the normalized frequency of word f_i in documents:

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

Experiment

There are several geographical information extraction tools available both in the forms of open source software packages and Web APIs, but I have chosen only the latter type to compare with Newsmap, because the former type are hardly accessible to media and communications scholars as they demands users advanced knowledge of Java and operation systems, whereas the online interfaces of Open Calais and Geoparser.io require only a short block of code in widely-used programing language (e.g. Python or R). They are commercial services but offers free access through non-commercial user accounts. Open Calais is provided by Thomson Reuters, one of the largest company in the media and information industry and the owner of Reuters news agency, which made us to expect that Open Calais is optimized for extracting geographical information from news articles. Dlugolinsky et al (2013) have reported that it recognizes 39 types of entities and its performance in identification of geographical locations in microblog posts are the best among six well known knowledge extraction tools. Geoparser.io is an online service started in 2016, specialising in geographical information extraction from natural texts. The founder of the service is one of the original developer of CLAVIN, who aims to improve the accessibility of geographical information extraction tools by providing the online interfaces (Greenbacker, 2017).

In addition to these geographical information extraction tools, I constructed a geographical dictionary, which contains 27,678 place names in 255 countries to replicate the methodologies in earlier studies (Blondheim et al., 2015; Zuckerman, 2003, 2008). The dictionary combines the list of the names of countries, administrative districts and cities with a population larger than 15,000. The gazetteers were created originally for the United States National Geospatial-Intelligence Agency's GEOnet Names Server (NGA) and Geological Survey's Geographic Names Information System (GNIS), but made available at GeoNames.[5] These gazetteers are also used by the above mentioned geographical information extraction systems. In applying the dictionary, English stopwords were removed from the documents and case-sensitive matching was performed to minimalize false positive matches.

Table 2 summaries the different approaches to geographical classification of news in Newsmap, Open Calais, Geoparser.io and the. Newsmap as a semi-supervised machine learning technique is ignorant about syntactical structure of documents, and named-entity recognition is solely based on capitalization of words; it requires training sets to be taken from the same time period as the test set to learn time-dependent association between names and places; it downplay ambiguous names that appear in different countries by assigning small weights.[6] Open Calais and Geoparser.io utilize syntactical analysis in named-entity

[5] GeoNames: http://www.geonames.org

[6] For example, names of athletes who travel frequently and appear in many different countries in the training set gain only small weights for all the countries. Therefore, their impact on the overall classification is small.

recognition, and lookup gazetteers for place names identified; place names are disambiguated exploiting the contextual information (map and knowledge-based). The dictionary as the simplest method performs neither named-entity recognition nor place name disambiguation, merely searching the list of place names for the words in documents.

The differences in the ways these methods detect geographical traits result in how they determine the focus of news stories. Newsmap ranks countries based on continuous scores assigned to names, but Open Calais only gives entities (places, persons or organizations) discreet relevance scores ranging from 0.0 to 1.0 in five steps, which does not always help identifying the single most important countries; Geoparser.io and the dictionary only report the positions and number of countries discovered. Therefore, when Open Calais gives the same total relevance scores to more than one country, identification of the top country becomes random among these; when Geoparser.io or the gazetteer find multiple counties with the same frequency, documents are classified into the country first mentioned.

Table 2: Characteristics of the geographical classification methods

	Type	NE recognition	PN disambiguation	Relevance scoring
Newsmap	Machine learning	Capitalization	Temporal	Continuous
Open Calais	Web API	Syntax	Contextual	Discreet
Geoparser.io	Web API	Syntax	Contextual	None
Gazetteer	Dictionary	None	None	None

Dataset

Training set

In this experiment, the Newsmap classifier was constructed from a corpus of online news. I have been subscribing to the Yahoo News US edition, which continuously supplies news stories produced by international news agencies (mainly AP, AFP, and Reuters) via a RSS feed, and a total of 156,980 items were collected in 2014 (approximately 13 million words). These texts are not complete news articles but summaries containing both headings and lead sentences. Use of newswires collected online is advantageous in the construction of a geographical news classifier because (1) news agencies tend to cover a wider range of countries than the retail media (Watanabe, 2013), and (2) subscription to RSS feeds allows users to sample stories without any pre-filtering. The Newsmap classifier constructed with the seed dictionary and the corpus contains 68,755 words for 227 countries. This training process takes only less than five minutes to complete on a laptop computer.

Test set

The classification accuracy of the four methods were measured by a set of manually coded news stories. This dataset is also comprised of news summaries collected online in 2014 but from different outlets: *The Times* (UK), *The New York Times*, *The Australian*, *The Nation* (Kenya), and *The Times of India*. From the collected news summaries, a balanced sample of 5,000 was randomly taken and classified by human coders in terms of their geographic association. The dataset needs be this large because international news coverage typically follows a power-law distribution, in which countries with little influence internationally appear only very infrequently. The motivation behind the choice of news sources was also to include countries 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.[7] 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 concerned (single-membership).[8] The coders' performance was constantly monitored using gold-standard answers created by the author, and coded subsets that did not achieve more than 70% agreement with the gold standard were discarded. Eventually, the same items were coded by at least three different coders, and the inter-coder agreement measured by Fleiss' multi-coder Kappa was $\kappa = 0.75$. After disagreement among coders was settled by the majority rule, the coders' agreement with the gold standard became $\kappa = 0.88$. The main causes of disagreement were (1) the difficulty in identifying countries most strongly associated with international stories, and (2) the coders' lack of knowledge about differences between countries with similar names (e.g. Congo Republic and The Democratic Republic of Congo). Such imperfection of human coding imposes a ceiling on the accuracy found in the experiments, even if the classifications were perfect from the experts' point of view, but their coding was treated as true answers.

Measurements

To simulate the most common setting in media studies where researchers select news stories based on their geographical focus, the classification is single membership in this experiment. Measurements of the classification accuracy are micro-average precision and recall, which are the standard measures for document classification tasks in computer science literature. 'Precision' is the ability of classifiers to retrieve ONLY relevant items, while 'recall' is the ability to retrieve ALL the relevant items. There is usually a trade-off relationship between the two abilities, and high precision often leads to low recall, and vice-versa. Low precision indicates many false positive cases, and low recall indicates many false negative cases. Micro-average precision and recall are calculated by pooling the classification results of all the classes, while macro-average precision and recall are the average of precision and recall separately calculated for each class.

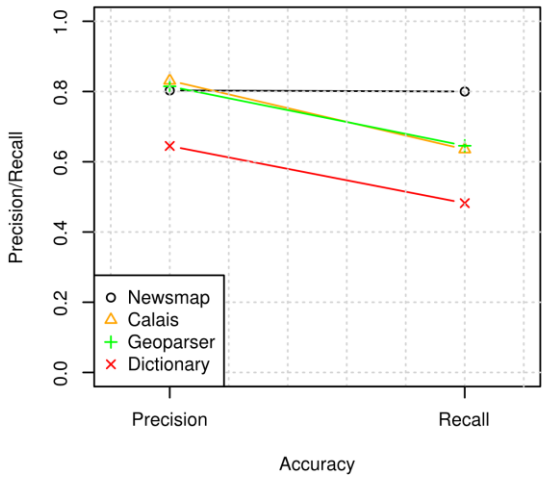
Results

The results of the experiment show that the overall classification accuracy of Newsmap is 0.80 both in precision and recall, while they are 0.83 and 0.63 in Open Calais, 0.81 and 0.64 in Geoparser.io and 0.64 and 0.48 in the dictionary (Figure 1). The harmonic means of precision and recall (F1 scores) are 0.80, 0.72, 0.72 and 0.55, respectively.

Figure 1: Overall classification accuracy

[7] Prolific Academic: https://prolificacademic.co.uk

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



In Figure 2, we find roughly the same level of precision across the most frequent countries in Newsmap, Open Calais and Geoparser.io. Exceptions are the Newsmap's relatively poor performance in the United Kingdom (GB), China (CN) and Malaysia (MY), and Open Calais's lack of precision figure in South Sudan (SS). There is no precision figure for Open Calais, because it does not distinguish between Sudan and South Sudan. The dictionary's precision varies and its precision is as high as other methods in many of the countries, but very low in some of the countries for the ambiguity of place names: its precision is low in the United Kingdom (GB) because the country shares many place names with the United States (US); stories about gay rights campaigns were misclassified into Russia (RU) because the country has a city called "Gay"; many of the stories about the former Egyptian president Mohamed Morsi were classified into India (IN) because of the Indian city of "Morsi"; many stories on Nigel Forage, a British Politician, was classified into South Africa (ZA) because of his first name; Many stories on social cohesion are classified into South Sudan (SS) for a state named "Unity"; a large number of stories about the United States were classified into Japan (JP) because of a small costal city called "Obama".

Figure 2: Precision in the top 20 countries

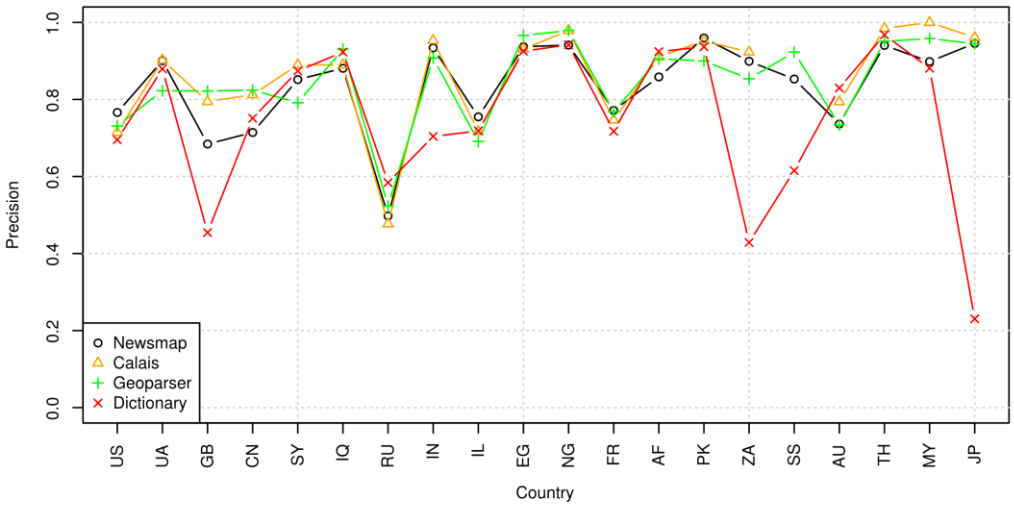
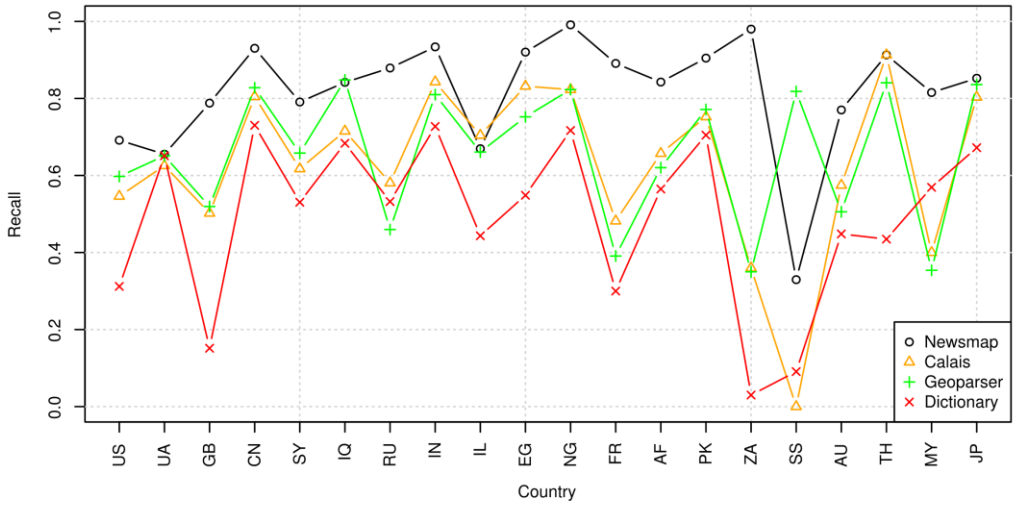


Figure 3 shows that Newsmap's recall is higher than or equal to other methods in nearly all the countries. Open Calais and Geoparser have similar levels of recall, but they are very different in South Sudan (SS). Since Open Calais classifies none of the story about the country correctly, its recall is zero, but it is over 0.8 in Geoparser.io. Newsmap also suffers in South Sudan, although it is still better than the dictionary. The dictionary's recall figures are much lower than other methods in United States (US) and the United Kingdom (GB), reflecting the misclassification caused by the above-mentioned ambiguity; the very low recall in South Africa (ZA) and South Sudan (SS) is due to Cameroon's "South" region.

Figure 3: Recall in the top 20 countries



Discussion

In the experiment, I have shown that Newsmap's precision is as high as Open Calais and Geoparser.io, both of which are equipped with syntactical named-entity recognition and contextual place name disambiguation mechanisms. This result suggests that weighting geographical traits based on the levels of ambiguity found in a concurrent corpus is an effective way of avoiding errors, and that ranking countries by total scores of geographical words is very accurate in determining the most strongly associated countries. The very low precision of the dictionary in several countries highlights the importance of place names disambiguation mechanisms in methods based on unweighted word counts as in Geoparser.io. Nonetheless, the dictionary's precision was as high as other methods in countries that have less ambiguous names (e.g. Ukraine, Syria, Iraq, Greece and Nigeria).

The much higher recall of Newsmap, Open Calais and Geoparser.io than that of the dictionary suggests that names of people and organizations are very important geographical traits in classification of news, in which widely known entities appear without explicit mentions of their home countries. Surprisingly, Newsmap's recall was even considerably higher than Open Calais and Geoparser.io. One would argue that this result is produced by the particular design of the experiment, but this is not the case. The figures are close to those shown in an earlier study conducted by Dlugolinsky et al (2013): the precision and recall was respectively 0.80 and 0.67 in recognizing locations. There is no surprise that Geoparser.io has roughly the same precision and recall figures as Open Calais's because they utilize very similar technologies.

The high precision of Newsmap is largely owing to its ability to recognize multi-word names, because the ambiguity of names is much higher when they are separated into single words. While the ability of the other methods to recognize multi-word names are constrained by the size of knowledge databases and gazetteers, Newsmap extracted various multi-word names from the training corpus. Although Buscaldi (2011) noted that the data-driven methods are uncommon in geographical information extraction systems due to the lack of manually labelled data despite their technological advantages, Newsmap has shown that the lack of training data is not an unsurmountable problem: simple keyword matching can be a substitute for manual coding. With the only 800 manually chosen keywords, Newsmap extracted 68,755 words from the Yahoo News corpus, which is an 86-time increase in the size of geographical vocabulary.

Simple keyword matching has been the popular approach to geographical news classification in media studies, but the dictionary's poor performance highlights the limitation of word counting-based classification, not only in geographical classification of news but in other applications. One might argue that its poor performance is due to the inappropriate size of vocabulary (either too small or too large) for the task, but determining the optimal size of vocabulary is very difficult: its recall will increase but its precision will decrease if the dictionary is larger containing names of smaller cities; its precision will increase but its precision will decrease, if the dictionary is smaller only including the name of the countries and administrative districts. Additionally, the heuristic rule to determine the most relevant country in the dictionary method is clearly not the reason of the its poor performance, because Geoparser.io performed very well based on the same rule.

Semi-supervised machine learning was utilized only for the geographical classification in this study, but nothing seems to limit the scope of its application. In general terms, the advantages of the semi-supervised document classification are the high degree of control given by a manually compiled seed dictionary, and the high reliability of parameters estimated in large training data. These features would solve challenges researchers face when they apply unsupervised (e.g. topic models) and fully-supervised (e.g. naive Bayes) document classifiers. The benefit of semi-supervised learning becomes greatest in multiclass classification tasks, but it is also useful in binary classification tasks.

The fully automated multi-word feature identification algorithm used in Newsmap can also be adopted to other applications since the model was originally purpose for identification phrasal verbs (c.f. Blaheta & Johnson, 2001). Identification of multi-word features would improve the classification accuracy of other bag-or-words models, not only because it reduces ambiguity of single words but also because it alleviates violation of the independent assumption by frequently cooccurring words (McCallum & Nigam, 1998). The solution is identifying multi-word features by the algorithm and redefining their word boundaries (i.e. concatenating components of multi-word features) (Lewis, 1998). There are other association measures that can identifying occurring words (e.g. PMI or likelihood ratio), but only very few measures can be applied to sequences of words in variable lengths (Blaheta & Johnson, 2001).

Finally, it is worth mentioning problems readers would face in applying the geographical classification methods in actual research. As for Newsmap, it requires large corpus of news to train a classifier, but, unlike the experiment, where two separate datasets were used for training and testing, the same dataset can be used in practice as far as it is large enough to cover all the countries the researchers wish to select or exclude.[9] As for Open Calais, its Web API is much more accessible than the open source packages in Java, it is still not straightforward to extract geographical information from its output. It is not only due to

[9] If such a corpus is not available, the author is willing to provide the Yahoo News corpus on request.

the large size of its output in XML, but also to the variety of geographical labels attached to entities.[10] Further, Open Calais only allows non-commercial user to call the API 5,000 times a day with 1 to 3 second intervals. In these respects, Geoparser.io's Web API is much more user friendly, because it returns only geographical information in JSON format, all the countries being identified by the standard country code, and it does not impose strict limit on non-commercial users. As for the dictionary method, removal of English function words (stopwords) and case-sensitive match is necessary to reduced false positive classification. The precision figure can be 10 points lower without this precaution in this method.

REFERENCES

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.

Buscaldi, D. (2011). Approaches to Disambiguating Toponyms. *SIGSPATIAL Special*, *3*(2), 16–19. https://doi.org/10.1145/2047296.2047300

Dlugolinsky, S., Ciglan, M., & Laclavík, M. (2013). Evaluation of named entity recognition tools on microposts. In *2013 IEEE 17th International Conference on Intelligent Engineering Systems (INES)* (pp. 197–202). https://doi.org/10.1109/INES.2013.6632810

Greenbacker, C. (2017, January 26). API key for Geoparser.io [Email interview].

Lewis, D. D. (1998). Naive (Bayes) at Forty: The Independence Assumption in Information Retrieval. In *Proceedings of the 10th European Conference on Machine Learning* (pp. 4–15).

Martins, B., & Calado, P. (2010). Learning to Rank for Geographic Information Retrieval. In *Proceedings of the 6th Workshop on Geographic Information Retrieval* (p. 21:1–21:8). New York, NY, USA: ACM. https://doi.org/10.1145/1722080.1722107

McCallum, A., & Nigam, K. (1998). A comparison of event models for Naive Bayes text classification. In *AAAI-98 Workshp in learning for text categorization* (pp. 41–48). AAAI Press.

Roberts, K., Bejan, C. A., & Harabagiu, S. M. (2010). Toponym Disambiguation Using Events. In *FLAIRS Conference* (Vol. 10, p. 1).

Rogers, R. (2013). *Digital Methods*. Cambridge, Massachusetts: MIT Press.

Smith, D. A., & Crane, G. (2001). Disambiguating Geographic Names in a Historical Digital Library. In P. Constantopoulos & I. T. Sølvberg (Eds.), *Research and Advanced Technology for Digital Libraries* (pp. 127–136). Springer Berlin Heidelberg. https://doi.org/10.1007/3-540-44796-2_12

Wacholder, N., Ravin, Y., & Choi, M. (1997). Disambiguation of Proper Names in Text. In *Proceedings of the Fifth Conference on Applied Natural Language Processing* (pp. 202–208). Stroudsburg, PA, USA: Association for Computational Linguistics. https://doi.org/10.3115/974557.974587

[10] It required the author to classify labels manually. For example, North Korea has at least four labels: "North Korea", "N. Korea", "North Korean" and "North korean".

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. https://doi.org/10.1177/1748048512465546

Wimmer, A., & Glick Schiller, N. (2002). Methodological nationalism and beyond: nation–state building, migration and the social sciences. *Global Networks*, *2*(4), 301–334. https://doi.org/10.1111/1471-0374.00043

Zaila, Y. L., & Montesi, D. (2015). Geographic Information Extraction, Disambiguation and Ranking Techniques. In *Proceedings of the 9th Workshop on Geographic Information Retrieval* (p. 11:1–11:7). New York, NY, USA: ACM. https://doi.org/10.1145/2837689.2837695

Zuckerman, E. (2003). *Global Attention Profiles: First steps towards a quantitative approach to the study of media attention*. The Berkman Center for Internet and Society, Harvard University. Retrieved from http://cyber.law.harvard.edu/sites/cyber.law.harvard.edu/files/2003-06.pdf

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