CLIFF-CLAVIN: Determining Geographic Focus for News Articles

[Extended Abstract]

Catherine D'Ignazio
MIT Center for Civic Media
77 Massachusetts Avenue
Cambridge, MA 02139, USA
dignazio@mit.edu

Rahul Bhargava
MIT Center for Civic Media
77 Massachusetts Avenue
Cambridge, MA 02139, USA
rahulb@media.mit.edu

Ethan Zuckerman
MIT Center for Civic Media
77 Massachusetts Avenue
Cambridge, MA 02139, USA
ethanz@media.mit.edu

Luisa Beck
Independent
Emmi.Beck@gmail.com

ABSTRACT

The growing diversity of news sources available online has led to a significant methodological change in field of global news coverage. Studies of media attention and framing require sophisticated analytic tools to permit analysis of a large volume of content consumed by a broad readership. Geographic focus continues to be a topic of interest to media organizations, media analysts, and media consumers. Detecting and recognizing geographic locations (toponyms) in news media is a well-established field with many commercial and open source tools available. An evaluation is performed of various existing tools to compare their accuracy and appropriateness for use within media organizations and for media analysis. The concept of *focus*, indicating the location an article is primarily about, is extended into the news realm and added to an existing tool to increase relevance for the aforementioned applications. Potential applications as well as initial experiments using geoparsing for news organizations are discussed, in addition to ideas for future work building on these tools.

1. INTRODUCTION

While newspaper publishing might be in decline, online news publishing and reading is an increasingly important facet of contemporary life. The widespread adoption of the Internet as a networked platform has expanded our definition of news to include blogs, citizen journalism, social media and other participatory forms of information creation. We live in the attention economy [20] where increased attention means increased access to resources. Likewise, little or no attention means that the world will ignore your problems. Many research labs, including the MIT Center for Civic Media, develop tools for large-scale quantitative analysis of the news as a space of global public discourse and agenda-setting.

Utilizing the MediaCloud platform [4], we analyze stories published in tens of thousands of news sources and make results available in public-facing tool. As news stories are about people, places and things, one of the essential dimensions of news analysis is geography. Newspapers are geographically oriented (e.g. The New York Times, The Boston Globe) and online news organizations also have a geographic scope of attention (e.g. The Huffington Post's offices, content and readers are mostly located in the domestic US). The technical challenge of understanding geography from news stories or any form of unstructured text is called geoparsing [10], geotagging [12] or geocoding [8]. We argue that in addition to sophisticated geoparsing software, a concept of geographic focus is necessary for any thorough media analysis effort and point towards potential applications for news organizations.

2. FULL TEXT GEOPARSING

Geoparsing software is increasingly utilized across a variety of online domains including web pages [2, 13, 23], blogs [16], Wikipedia [15], Twitter [19, 14], web APIs [8], libraries [7, 6] and news stories [3, 9, 12, 11, 22, 24]. There are a number of commercial products available including Thomson Reuters's OpenCalais[1], Yahoo!'s Placespotter[2], and MetaCarta's GeoTag [3]. There are open source products such as CLAVIN[4].

Geoparsing systems typically are made up of two stages. First, text is processed to identify possible place names, and geolocated to create a list of possible physical locations. Some systems use part-of-speech tagging to accomplish this toponym-recognition phase [13]. The technical challenge here is separating place names with names of other entities like people. For example, is "Jordan" a reference to a country, the basketball player, or a religious reference?

NewsKDD 2014, New York, New York

[1] http://www.opencalais.com/

[2] https://developer.yahoo.com/boss/geo/docs/key-concepts.html

[3] http://www.metacarta.com/products-platform-geotag.htm

[4] http://clavin.bericotechnologies.com

The second function of geoparsing systems is to associate a single latitude and longitude with each location mentioned, called toponym resolution [13, 1] or geographic name disambiguation [5, 21]. These typically employ gazetteer-based approaches [8, 7], combined with heuristics [9], to select among candidate locations. The technical challenges of this stage have to do with locating the place mention to the right place. For example, there are over 200 places in the world named ”Springfield” - which one is correct in a given context?

2.1 News Story Focus

Although a variety of geoparsing systems are available, they seldom focus on news content. There are some notable exceptions. GDELT [9] extracts events from news reports. NewsStand [22] geoparses and places news stories on a map. The European Media Monitor [3] does near real-time news analysis, including geoparsing, for multilingual documents. Web-a-Where [2] geolocates online mentions of places on webpages, and adds in a crucial element of focus to suggest the specific location a webpage is really about. We argue this intuitive concept of focus is critically important for newsroom applications.

Our work builds on the concept of geographic focus developed for geoparsing news stories. The CLIFF tool is our own implementation of the pipeline outlined by Amitay et. al.[2]: 1. Spotting (toponym recognition), 2. Disambiguation (toponym resolution) and 3. Focus determination. Focus can be defined as the geographic place/s that a news article is about. For example, an article like ”G-8 Meeting Ends with Cordial Stalemate on Syria” [5] mentions numerous places but its focus is Syria.

While Web-a-Where and subsequent research on focus [13, 23] has oriented around geotagging web pages, we are interested in applying the concept of focus specifically to news stories. Applying the concept of Focus Determination in the news context helps make geoparsing more relevant to the way that news organizations and readers think about news stories. For example, the Boston Globe requires journalists to tag their stories with a piece of metadata indicating the geographic focus. The New York Times Annotated Corpus [18] has hand-assigned locations associated with places in the article. We see a pattern emerging in which focus is an important piece of meta-data for content analysis in the newsroom, and for media studies at large.

3. EVALUATING PLACE MENTIONS AND FOCUS

To create a geoparser with a concept of focus targeted for media research and newsroom applications, we began by evaluating three existing technologies.

3.1 Methodology for Evaluating Existing Geoparsers

In order to establish a baseline for measuring the performance of geoparsing technologies in relation to both place mentions and focus, we created a manually coded data set of news stories. While there are geographic data sets for assessing place mentions, there are no existing data sets that explicitly address the concept of focus. As media sources vary in style guides, article lengths, and reading levels it was important to include multiple sources in our data set.

Table 1: Human Agreement Averages

Metric	Average Inter-Coder Agreement
Place Mentions	92.23%
Country of Focus	96.15%

Our goal was to measure how often humans agreed with each other on 1) place mentions in news articles and 2) overall focus determination at the country level for each article so that we had a strong intercoder reliability measurement.

Our data set is comprised of 75 articles —25 from the New York Times, 25 from the Huffington Post and 25 from the BBC. They were randomly sampled from the previously described MediaCloud archive for the month of February 2013. To account for variations in human coding, we developed strong coding guidelines including rules for edge cases.

Two of the authors separately hand-coded the set of 75 articles using these rules. See Table 1 for the results of this process. This gave us a good baseline for the peak possible performance of any given geoparser. Note that we evaluated systems against only those articles that had 100% inter-coder agreement.

3.2 Performance of Existing GeoParsers

From an initial survey of the field of available geoparsing technologies[6], we chose to evaluate two of the widely used products in the space that do both toponym recognition and toponym resolution: OpenCalais and Placespotter. We included the open-source CLAVIN tool as well. To assess performance against the human-coded ”Country of Focus” metric, we calculated a single country of focus based on the number of times it was mentioned in the text (this included mentions of cities or other locations determined to be within the country). This simple ”frequency of mention” strategy for determining focus served as a placeholder for evaluating accuracy of focus from each tool.

We ran our tests with the 75 articles from three different news sources and averaged the results, as presented in Table 2. Here are some of the relevant outcomes:

- Placespotter consistently outperforms OpenCalais and CLAVIN on recall and geographic disambiguation. Overall it picks up more place references than the other two technologies.

- Placespotter has a higher rate of false positives - incorrectly identifying text as a place —whereas CLAVIN and OpenCalais have high precision scores.

- Because of that, the F1 scores for CLAVIN and Placespotter are comparable (0.76 and 0.78 respectively) whereas OpenCalais scores lower (0.69)

- All of the technologies did relatively well with toponym resolution with Placespotter leading them at 96%.

- Using a simple strategy of frequency of mention as a measure of focus of an article allowed us to correctly locate an article at the country level 75.3% of the time using CLAVIN.

[5] http://www.nytimes.com/2013/06/19/world/europe/g-8-meeting-ends-with-cordial-stalemate-on-syria.html

[6] The list we compiled is here: http://bit.ly/UH4vRJ

Table 2: Performance of Existing Tools

Tool	Recall	Precision	F1	Disambiguation Accuracy	Focus Accuracy
Yahoo Placespotter	69.50%	87.70%	0.78	96.27%	69.02%
OpenCalais	53.03%	96.78%	0.69	90.28%	59.57%
CLAVIN	63.78%	94.25%	0.76	89.91%	74.30%

4. GEOGRAPHIC FOCUS

Based on the data presented above, and weighing in the open source nature of CLAVIN, we chose to build on top of CLAVIN and tune it further for detecting news story focus from large data sets of news articles.

4.1 Modifications to CLAVIN's Disambiguation Stage

As we built on top of the CLAVIN architecture, we extended and adapted it to our needs. Our first focus was on the disambiguation stage of the pipeline, as CLAVIN relied on a simple recursive approach with a preference towards picking candidate locations within the same country. Studies have found simple techniques like this can perform well [9], but in testing against our larger corpora of news articles we found an unsatisfactory level of disambiguation errors. To remedy this, we introduced a multi-stage heuristic disambiguation pipeline into the CLAVIN architecture and named our tool CLIFF[7][8].

4.2 Strategies for Determining Focus

We tested several algorithms for detecting focus. Each was evaluated to determine the best-performing strategy. Our most naive heuristic selects the location a document is about by simply sorting the locations by frequency of mention. Our hypothesis was that if an article mentioned one particular country most often (or places in that country most often) then the article was most likely about that country. The country mentioned most often is selected as the geographic focus of the document. The same logic is applied to states and cities. In the case of two locations occurring the same number of times, we select both.

However, we knew our documents are news articles, and could infer some natural document structure based on that. For instance, news articles have a history of strong summarial titles. Similarly, they often provide valuable context in the first paragraph. These observations led us to try a heuristic that scored locations based on where they were located in the text of the document. Locations mentioned earlier in the text earned more points than those mentioned later. We tried a variety of weighting and scoring strategies, but none performed as well as the naive frequency-of-mention heuristic. A detailed analysis of those results falls out of the scope of this extended abstract.

4.3 Evaluating Focus

The previously described hand-coded corpus created a good starting point for comparison. However, it lacks the scale needed to make strong statements about accuracy. To address this, two other corpora were utilized. First, we used the New York Times Annotated Corpus, which includes a locations field on each article with tags ”hand-assigned by The New York Times Indexing Service” [18]. We also included the Reuters RCV-1 corpus, which includes a countries tag on each article created by ”a combination of auto-categorization, manual editing, and manual correction” [17]. Neither corpus claims to have annotated articles for locations they are ”about” i.e. they have not coded their stories for country-level focus. That said, we argue the locations our heuristic focus algorithms select should be a strict subset of the locations the corpora have annotated it with.

Table 3: Country-Level Focus Detection

Corpus	Sample Size	Accuracy
Hand-Coded Data	75	95%
New York Times Corpus	10,000	90%
Reuters RCV-1	10,000	91%

We compared our various implementations of focus detection on this metric. Table 3 shows the results of this testing.

A cursory manual review of errors suggests that most can be attributed to problems in precision, recall, or disambiguation that happen earlier in the pipeline. This high level of accuracy has given us the confidence to incorporate CLIFF/CLAVIN into the core Media Cloud content analysis system. CLIFF exposes an HTTP-based API to produce JSON results, allowing for easy integration.

5. APPLICATIONS FOR THE NEWS

While we have developed CLIFF for integrating into MediaCloud for research purposes, we see potential for integrating geographic analytics into the newsroom for editors, journalists and managers. For example, analytics about the geographic distribution of an organization's news may highlight over- or under-covered geographies which need editorial consideration. Executives and marketing may be interested in looking at how the locations and interests of subscribers and readers intersect with the news coverage. And journalists may want to geolocate a set of third party news sources to create a news analysis piece about how a particular story is unfolding from a geographic standpoint. Finally, newsrooms might look to repackage their archives into novel news discovery products organized around geography such as ”Terra Incognita”[9], a geographic news recommendation system we are piloting based on CLIFF.

Geographic information can have particular resonance for news applications in combination with other data sets. We have begun initial forays down this path with partners at the Boston Globe, resulting in the ”Mapping the Globe” project[10]. This website maps the geographic coverage of stories in the Boston Globe, and highlights unique keywords used in the coverage of each place. When geographic information, demographic data and news content are combined, they create a picture of how that particular geography is

[7] Cliff Clavin is a mailman on the long-running television show Cheers! set in Boston, MA, where the authors live.

[8] http://cliff.mediameter.org

[9] https://www.terra-incognita.org

[10] http://globe.mediameter.org

represented in a longitudinal way across many news stories. This can provide powerful insight into patterns, trends and biases of news coverage that are hard to see otherwise.

6. CONCLUSION

Our key contribution to this space is the application of the concept of geographic focus to news articles and the evaluation of how well three different technologies perform at assessing Amitay et al's concept of focus. We develop a simple method for deriving focus based on frequency of place mention and evaluate it using three well-recognized geoparsing technologies including Yahoo Placespotter, OpenCalais and CLAVIN. We also compute the standard measures of geoparsing performance (Precision, Recall, F1 and Accuracy of Toponym Resolution) for each of those technologies against several geographic data sets, including a hand-coded set of news articles from multiple media sources. Our data shows that the CLAVIN system is comparable in performance with Yahoo Placespotter and has the advantage of being free, open source and thus tunable to a news context. We demonstrate improvements in CLAVIN's toponym resolution strategy for news articles and show that we can accurately determine the geographic focus of news articles at the country level 95% of the time. Finally, we point out potential applications of geoparsing technologies for news organizations. Our future work includes evaluating each stage of our recognition pipeline more closely to assess its contribution to success and failure modes.

7. REFERENCES

[1] M. D. Adelfio and H. Samet. GeoWhiz: toponym resolution using common categories. In *Proceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, SIGSPATIAL'13, pages 532–535. ACM.

[2] E. Amitay, N. Har'El, R. Sivan, and A. Soffer. Web-a-where: Geotagging web content. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '04, pages 273–280. ACM.

[3] M. Atkinson and E. Van der Goot. Near real time information mining in multilingual news. In *Proceedings of the 18th International Conference on World Wide Web*, WWW '09, pages 1153–1154. ACM.

[4] Y. Benkler, H. Roberts, R. Faris, A. Solow-Niederman, and B. Etling. Social mobilization and the networked public sphere: Mapping the SOPA-PIPA debate. (2013).

[5] D. Buscaldi. Approaches to disambiguating toponyms. 3(2):16–19.

[6] J. S. Creel and K. Weimer. Toponym extraction and resolution in a digital library. In *Proceedings of the 12th ACM/IEEE-CS Joint Conference on Digital Libraries*, JCDL '12, pages 415–416. ACM.

[7] L. L. Hill, G. Hodge, and D. Smith. Digital gazetteers: Integration into distributed digital library services. In *Proceedings of the 2Nd ACM/IEEE-CS Joint Conference on Digital Libraries*, JCDL '02, pages 427–427. ACM.

[8] M. Karimzadeh, W. Huang, S. Banerjee, J. O. Wallgrun, F. Hardisty, S. Pezanowski, P. Mitra, and A. M. MacEachren. GeoTxt: a web API to leverage place references in text. In *Proceedings of the 7th Workshop on Geographic Information Retrieval*, GIR '13, pages 72–73. ACM.

[9] K. H. Leetaru. Fulltext geocoding versus spatial metadata for large text archives: Towards a geographically enriched wikipedia. 18(9).

[10] J. L. Leidner and M. D. Lieberman. Detecting geographical references in the form of place names and associated spatial natural language. 3(2):5–11.

[11] M. D. Lieberman and H. Samet. Adaptive context features for toponym resolution in streaming news. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '12, pages 731–740. ACM.

[12] M. D. Lieberman and H. Samet. Multifaceted toponym recognition for streaming news. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, pages 843–852. ACM.

[13] M. D. Lieberman, H. Samet, J. Sankaranarayanan, and J. Sperling. STEWARD: architecture of a spatio-textual search engine. In *Proceedings of the 15th Annual ACM International Symposium on Advances in Geographic Information Systems*, GIS '07, pages 25:1–25:8. ACM.

[14] J. Lingad, S. Karimi, and J. Yin. Location extraction from disaster-related microblogs. In *Proceedings of the 22Nd International Conference on World Wide Web Companion*, WWW '13 Companion, pages 1017–1020. International World Wide Web Conferences Steering Committee.

[15] R. Odon de Alencar, C. A. Davis, Jr., and M. A. Goncalves. Geographical classification of documents using evidence from wikipedia. In *Proceedings of the 6th Workshop on Geographic Information Retrieval*, GIR '10, pages 12:1–12:8. ACM.

[16] T. Qin, R. Xiao, L. Fang, X. Xie, and L. Zhang. An efficient location extraction algorithm by leveraging web contextual information. In *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems*, GIS '10, pages 53–60. ACM.

[17] T. Rose, M. Stevenson, and M. Whitehead. The reuters corpus volume 1-from yesterday's news to tomorrow's language resources.

[18] E. Sandhaus. The new york times annotated corpus LDC2008T19.

[19] J. Sankaranarayanan, H. Samet, B. E. Teitler, M. D. Lieberman, and J. Sperling. TwitterStand: news in tweets. In *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, GIS '09, page 42âĂŞ51. ACM.

[20] H. A. Simon. Designing organizations for an information-rich world. In D. M. Lamberton, editor, *The economics of communication and information*, pages 187–202. Elgar Reference Collection. International Library of Critical Writings in Economics, vol. 70.

[21] D. A. Smith and G. S. Mann. Bootstrapping toponym classifiers. In *Proceedings of the HLT-NAACL 2003*

Workshop on Analysis of Geographic References - Volume 1, HLT-NAACL-GEOREF '03, pages 45–49. Association for Computational Linguistics.

[22] B. E. Teitler, M. D. Lieberman, D. Panozzo, J. Sankaranarayanan, H. Samet, and J. Sperling. NewsStand: a new view on news. In *Proceedings of the 16th ACM SIGSPATIAL international conference on Advances in geographic information systems*, page 18. ACM.

[23] A. Zubizarreta, P. de la Fuente, J. M. Cantera, M. Arias, J. Cabrero, G. Garcia, C. Llamas, and J. Vegas. A georeferencing multistage method for locating geographic context in web search. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, CIKM '08, pages 1485–1486. ACM.

[24] E. Zuckerman. Global attention profiles - a working paper: First steps towards a quantitative approach to the study of media attention. (2003).