Identifying Short-Names for Place Entities from Social Networks

Faizan Wajid
Department of Computer Science
University of Maryland, College Park
College Park, MD 20740
fwajid@cs.umd.edu

Hong Wei
Department of Computer Science
University of Maryland, College Park
College Park, MD 20740
hyw@cs.umd.edu

Hanan Samet
Department of Computer Science
University of Maryland, College Park
College Park, MD 20740
hjs@umiacs.umd.edu

ABSTRACT

Organizations can be identified by a myriad of terms apart from their official names. While abbreviations remain a common "short-name" to reference organizations, the prevalence of other short-names has risen in conjunction with social networks. When a user enters a short-name as a locational search query, it remains a challenge to infer the relationship between the short-name and the organization it ostensibly represents. For a number of organizations around the Washington D.C., Maryland, and Virginia area, we first generate a list of possible short-names for each of them. We then search through their tweets to build a corpus of short-names associated with each organization. By measuring our list against the corpus, we can identify potential short-names, and return the location of the organization.

CCS CONCEPTS

•**Information systems** →*Information extraction; Geographic information systems;*

KEYWORDS

Short Names, Twitter, Toponym Recognition, GeoNames

ACM Reference format:
Faizan Wajid, Hong Wei, and Hanan Samet. 1997. Identifying Short-Names for Place Entities from Social Networks. In *Proceedings of ACM SIGSPATIAL Workshop on Recommendations for Location-based Services and Social Networks, Redondo Beach, CA, November 7 2017 (LocalRec '17),* 4 pages.
DOI: 10.475/123_4

1 INTRODUCTION

Organizations and institutions around the world are often referred to by names other than their official names. Alternative names are given to such entities for sake of brevity and/or convenience, and other times adopted by people due to marketing or advertisements. Despite this long-standing practice, only some of these alternative names are used officially, and increasingly social networks now provide another medium in which these alternative names (which we call *short-names*) flourish.

To start with a motivating example: as students of the University of Maryland, we colloquially refer to our institution as "UMD", and when talking about our department, we say "UMD CS". While these terms have become canonical in our daily jargon, mentions to terms such as "Maryland Comp Sci" or the more ambiguous "UMCS" are also encountered.

Many search engines are adept in recognizing a variety of an entity's short-names but fall short of recognizing many official *and* unofficial ones. Moreover, they lack the ability to keep up with newly-generated short-names as they appear on social networks. This naturally extends to the problem of recognizing the correct place entity in spatial queries, given a short-name. In the above example, the case of "UM" is an officially recognized short-name of the University of Maryland, however spatial queries do not return this reference regardless of a user's proximity to the university.

Solving this problem is beneficial to many fields such as named entity recognition [22] and furthermore one of subtask *toponym recognition* [10, 11, 15, 26], which is to recognize textual references (place names such as "UMD") to geographical locations in the text, as well as the subtask of *toponym resolution*[16], which is to disambiguate between multiple interpretations to place names like "London", both of which play an essential role in many map-based information aggregation systems such as news monitoring [2, 8, 12, 17, 24, 25], crime tracking [27], diseases tracking [13, 21] and etc. However, many proposed approaches [3, 5, 7, 9, 14, 18–20, 23] to toponym resolution are utilizing the GeoNames, an open-source geographical databases of millions of place names [1], as its source of place names. Although a table of alternative names is provided in GeoNames to increase its chance of recognizing a place which might have several candidate names, such a table usually falls short in practice [4]. For example, GeoNames does not have an alternative name entry to associate "UMD" with the "University of Maryland" and thereby will fail to return a correct geographical latitude and longitude values for "UMD".

This paper relates our efforts to first identify the many types of short-names that exist for any given place entity (the organization or institution in question), and then procedurally generate potential short-names for them. Since these short-names are entities that are associated with spatial data in the sense that they are instances of toponyms, the goal then is to associate user-provided short-names to place entities to make locational searches more convenient.

The remainder of this paper is organized as follows. First we discuss our data sources and text formatting techniques before we describe our short-name generation heuristics (Section 2). We then report the results of applying our methodology (Section 3). Following this, we briefly present our demonstration (Section 4). Finally, we outline plans for improvement and future considerations (Section 5).

LocalRec '17, Redondo Beach, CA

DOI: 10.475/123_4

2 METHODOLOGY

We first limit our scope only to Washington, D.C. along with it's two neighboring states (Maryland and Virginia), and further focus on organizations and institutions (referred to as *place entities*) that are founded or headquartered (or both) in these states. While the concentration of population is around the D.C. metropolitan area, expanding to the neighboring states give us more entities to work with. We group place entities per their website's top-level domain (TLD) yielding four categories: *Commercial* for business-related entities, *Education* for universities and other scholastic institutions, *Government* for Federal, state, and municipal institutions, and *Military* for military and defense-related research institutions. By collecting tweets from these various organizations, we identify some heuristics for the generation of short-names.

2.1 Data Gathering

To build this list of organizations, we leverage Wikipedia primarily to obtain lists and references for place names. This means gave us an exhaustive list for universities and other educational institutions, including their enrollment numbers per year. We also used Google Maps to locate high-profile companies in certain areas (e.g. the Dulles Technology Corridor) which we cross-referenced from their Forbes Fortune 500 rating. We opted to use high-profile companies based on their influence on the general public, and therefore generate visibility and traffic on social media. As the goal is to provide locations, we store the latitude and longitude for each organization's headquarters or administrative office.

The next step was to select good candidates from this list of organizations and to make sure we only work with unique entities. To this end, we use mutual reference to compare whether the link to the entity's Twitter profile on the entity's website is *the same as* the link to the homepage on the Twitter profile. We discard any entities that do not meet this requirement. In fact, we can further build trust by verifying the same with the entity's other social media accounts, as we will describe later.

From a list of 100 entities aggregated, we verify if they are active on Twitter by measuring how many tweets they publish or are mentioned in per day. We then select the latest 3,200 tweets published. For our benefit, the Twitter API provides APIs for collecting tweets[1] and performing mutual reference[2]. We subject the tweets to simple normalization whereby we remove stopwords, as well as the # symbol to treat hashtags as regular text. We also keep the order of words in tweets to maintain context.

To develop a ground-truth, we must manually identify the various short-names associated with each entity. The Wikipedia article lists common short-names and abbreviations for each entity; we also collected high-frequency hashtags (prior to normalization) and determined whether or not they were used as short-names (as opposed to promotional events). While this list of short-names per entity is not collectively exhaustive, it does provide a representative sample of the various names an entity can have, for both official and unofficial purposes. Given the breadth of manual work required, we uniformly sampled 30 entities from the aggregate list, and arrived at a condensed collection of 11 Education entities, 7 Government, 4 Military, and 8 Commercial.

[1] https://dev.twitter.com/rest/reference/get/statuses/user_timeline
[2] https://dev.twitter.com/rest/reference/get/users/lookup

2.2 Short-name Generation

We extracted a number of patterns from our list of short-names and compiled heuristics which we programmatically execute on every place entity name to build a list of potential short-names. The list below enumerates each list-item with a "B" as each list-item effectively functions as a bucket for collecting short-names that fall into it.

- B1. *Initializations:* The simplest short-name we observed was the *letter abbreviation* where only the first letter of each word is kept (e.g. "United States of America" is "U.S.A."). Note this technique only applies to place-entities with two or more words. We also build the abbreviations with and without prepositional keywords such as "of", and "and" as their inclusion in short-names varies by entity.
- B2. *State abbreviations:* Some entities contain the name of the state they're located in (particularly universities) and many utilize the official state abbreviation in the short-name (e.g. "Maryland" is "MD"). Other cases involve using the first letter of the state's name, falling in the above category.
- B3. *Word swapping:* Another short-naming practice entails rearranging the words in the place entity's name, e.g. representing "Department of State" as "State Department". While uncommon, we notice this technique was used more frequently with Federal agencies and a limited number of universities.
- B4. *Common abbreviations:* Many words have well recognized abbreviations, such as "U" or "Univ" for "University" and "Dept" for "Department". We also include condensing repeat letters to the number of times they occur (e.g. "Community College" as "C2").
- B5. *Syllables:* In some cases, place entity short-names also include abbreviations based on syllables, but more noticeably the first syllabic element in a word. For example, the "University of Michigan" is recognized as "U.Mich", where "Mich" is the first syllabic element in Michigan. We leverage the Moby Hyphenated Word List [28] to identify the first syllabic element for our place entities.

3 RESULTS

To evaluate our methodology, we assess three points: 1. correctly generate all possible short-names for any given entity as populated in our corpus; 2. correctly identifying the entity, and therefore location, when provided a short-name; and 3. significant short-names encountered in tweets but *not* populated in our corpus.

With respect to the first point, our short-name generation methodology identified all official and unofficial short-names associated with each entity that exist in our corpus. We report only one officially recognized short-name was missed: "CommerceGov" for the "Department of Commerce", simply because we did not consider a case such as this (the inclusion of the entity's TLD (Top-Level Domain) with part of it's name). While it's true that greedy generation would certainly arrive at results like this, we note the necessity of this approach due to the flexibility of language and the common

Table 1: Average occurrence of bucket-items per entity in each category.

Bucket	edu	com	gov	mil
B1	68.3%	88.8%	75.0%	66.7%
B2	44.4%	22.2%	12.5%	33.3%
B3	5.00%	22.2%	12.5%	16.6%
B4	70.3%	38.5%	31.3%	33.3%
B5	2.67%	11.1%	6.25%	16.6%

usage of colloquialisms in social media. We provide Table 1 to delineate the commonplace usage of each type of abbreviation method per entity in each organization category and, for brevity, only show the average occurrences. We also draw attention to the fact that all bucket-item values are greater than zero, and can clearly see that certain categories exhibit short-name preferences.

To our second point, we simply perform a reverse look-up of the short-name on our corpus and return the salient place name and location. We improve the recognition of each short-name to entity place name by assigning it a score by virtue of tf-idf. In the phase of scanning through tweets, short-names that are more frequently mentioned alongside a particular entity will appear in more volume. While it's certainly possible for multiple entities to share the same short-name, the entity that is more renowned will be more recognized. Given our spatial domain being restricted to local organizations, we did not encounter conflicting entries, however in cases of equal scores, a simple tie-break would suffice.

With regards to the third point, we encountered a handful of instances where unique short-names not in our corpus were used to reference place entities. For example, with respect to the Federal Reserve, numerous mentions were made to "TheFeddy", and for George Washington University, we saw "GDubs". In many of these cases, we only spotted these short-names because they contained a substring of a corpus entry. We ignored low frequency mentions and report that an average of three non-corpus short-names appeared per category. As we increase our data-set and define more heuristics, we will certainly be able to capture such cases more effectively.

4 DEMONSTRATION

We present a simple short-name geographical database as seen in Figure 1. A rudimentary website (that takes visual cues from GeoNames) allows users to input short-names into a search bar and retrieve the place-name, along with a small list of other high-frequency short-names. A button titled "Show on map" displays the entity on an interactive map, and clicking on the place name redirects the user to the entity's Wikipedia article.

5 CONCLUSION AND FUTURE WORK

In this paper, we describe short-name generation heuristics for spatial querying of organizations. Organizations can be identified with many names, and while some of these are used officially, social networks are another venue where short-names of both varieties flourish. The goal here is to associate user-provided short-names to place entities (the organization or institution in question) to make locational searches more convenient.

In our study, we utilized a list of 30 organizations to build our ground-truth and collect data. The small dataset is mainly due to difficulties in hand-labeling the various short-names an entity can have, and though we take advantage of hashtags, they alone are not sufficient in building a list of potential short-names. As such, we provide generated short-names that were not in our list, but appeared very frequently in the entity's tweets. This outcome alone leads us to wonder the questions of how new short-names arise and the propensity of their adoption, and if these factors play in how long the short-name will live.

To add, our list of companies also did not include conflicting organizations, that is, multiple organizations that share the same short-name. This will invariably be an issue as we increase spatial resolution from local organizations to global and locational assistance will not be sufficient to alleviate this problem. Providing score is a stable first-step, however better approaches will need to be considered.

Another shortcoming was that we limited ourselves to tweets, however many short-name mentions appear across Facebook, Yelp, and other popular social media domains. Incorporating their data would be beneficial in allowing us to improve our generation method, including some level of sentiment analysis. Leverage Amazon Mechanical Turk to crowd-source hand-labeling is also very practical, as our belief is that people will be better able to identify short-names, making our ground-truth more resilient.

As we mentioned earlier, we manually associated latitude/longitude for each place name. By incorporating Facebook into the mutual reference process, we can obtain the location directly from the entity's profile. The idea here is to build a fully automated pipeline that can not only yield a unique list of organizations, but to return location from short-name inference.

Our efforts to build a short-name generator provided us with sufficient training data to employ machine learning techniques for more effective recognition of short-names to their place name equivalent, supplemented with location data. A simple approach would be to utilize the StanfordNER [6] which performs name-entity recognition on both Location and Organization. More advanced techniques could be used for querying, as well as sentiment analysis of the short-name (and context) provided.

In terms of features, many extensibility options exist. One such example is embedding short-name detection into NewsStand [24] to provide late-breaking news relative to the search query. The major benefit of this integration is to provide larger spatial context for both big and small place entities, which could include social events happening nearby. Another interesting venue would be to enable crowd-sourcing methods and allow users to submit potential short-names for place names.

While our results have been positive, there are still many interesting questions we seek to answer from our findings. We are encouraged to develop and enhance our methods further for better spatial querying of short-name. With users by and large referring to social media to find places to visit, improved (and timely) short-name recognition will become a critical feature in spatial querying solutions.

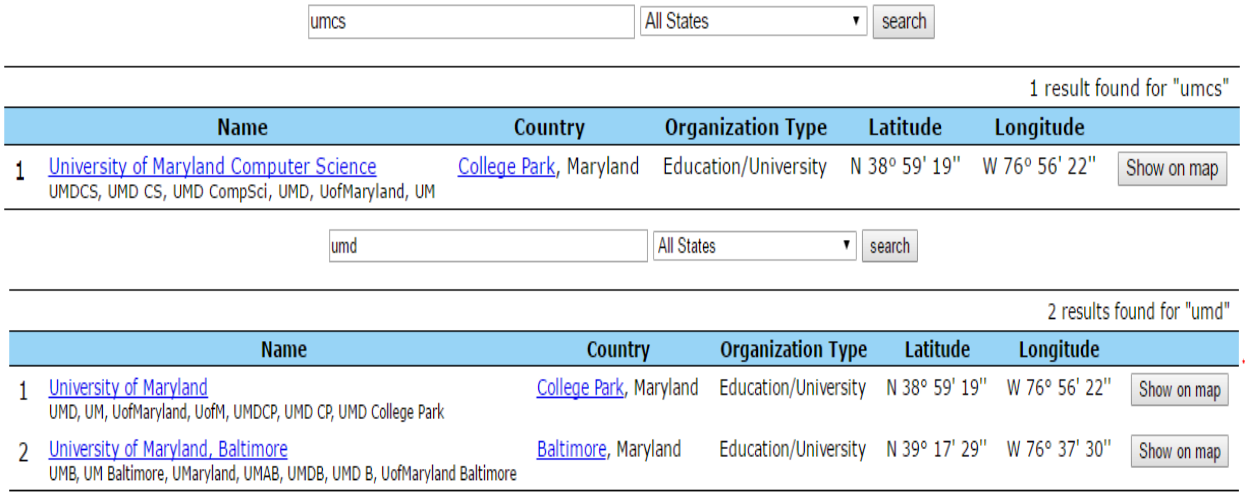

Figure 1: Two sample screenshots of the short-name spatial query web interface.

REFERENCES

[1] GeoNames. (2017). http://www.geonames.org/

[2] Ahmed Abdelkader, Emily Hand, and Hanan Samet. Brands in NewsStand: Spatio-temporal Browsing of Business News. In *Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems (SIGSPATIAL '15)*. ACM, New York, NY, USA, Article 97, 4 pages.

[3] Marco D. Adelfio and Hanan Samet. Structured Toponym Resolution Using Combined Hierarchical Place Categories. In *Proceedings of the 7th Workshop on Geographic Information Retrieval (GIR '13)*. ACM, New York, NY, USA, 49–56.

[4] Dirk Ahlers. Assessment of the Accuracy of GeoNames Gazetteer Data. In *Proceedings of the 7th Workshop on Geographic Information Retrieval (GIR '13)*. ACM, New York, NY, USA, 74–81.

[5] Carlo Bernaschina, Ilio Catallo, Eleonora Ciceri, Roman Fedorov, and Piero Fraternali. Towards an Unbiased Approach for the Evaluation of Social Data Geolocation. In *Proceedings of the 9th Workshop on Geographic Information Retrieval (GIR '15)*. ACM, New York, NY, USA, Article 10, 2 pages.

[6] Jenny Rose Finkel, Trond Grenager, and Christopher Manning. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. In *ACL*. Ann Arbor, MI, 363–370.

[7] Rodolfo Gonzalez, Gerardo Figueroa, and Yi-Shin Chen. TweoLocator: A Non-intrusive Geographical Locator System for Twitter. In *Proceedings of the 5th ACM SIGSPATIAL International Workshop on Location-Based Social Networks (LBSN '12)*. ACM, New York, NY, USA, 24–31.

[8] Nick Gramsky and Hanan Samet. Seeder Finder: Identifying Additional Needles in the Twitter Haystack. In *Proceedings of the 6th ACM SIGSPATIAL International Workshop on Location-Based Social Networks (LBSN '13)*. ACM, New York, NY, USA, 44–53.

[9] Bo Han, Paul Cook, and Timothy Baldwin. Text-based Twitter User Geolocation Prediction. *J. Artif. Int. Res.* 49, 1 (Jan. 2014), 451–500.

[10] Diana Inkpen, Ji Liu, Atefeh Farzindar, Farzaneh Kazemi, and Diman Ghazi. *Detecting and Disambiguating Locations Mentioned in Twitter Messages.* Springer International Publishing, Cham, 321–332.

[11] Neil Ireson and Fabio Ciravegna. Toponym Resolution in Social Media. In *Proceedings of the 9th International Semantic Web Conference on The Semantic Web - Volume Part I (ISWC'10)*. 370–385.

[12] Alan Jackoway, Hanan Samet, and Jagan Sankaranarayanan. Identification of Live News Events Using Twitter. In *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Location-Based Social Networks (LBSN '11)*. ACM, New York, NY, USA, 25–32.

[13] Rongjian Lan, Marco D. Adelfio, and Hanan Samet. Spatio-temporal Disease Tracking Using News Articles. In *Proceedings of the Third ACM SIGSPATIAL International Workshop on the Use of GIS in Public Health (HealthGIS '14)*. ACM, New York, NY, USA, 31–38.

[14] Sunshin Lee, Mohamed Farag, Tarek Kanan, and Edward A. Fox. Read Between the Lines: A Machine Learning Approach for Disambiguating the Geo-location of Tweets. In *Proceedings of the 15th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL '15)*. ACM, New York, NY, USA, 273–274.

[15] Michael D. Lieberman and Hanan Samet. Multifaceted Toponym Recognition for Streaming News. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '11)*. ACM, New York, NY, USA, 843–852.

[16] Michael D. Lieberman and Hanan 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)*. ACM, New York, NY, USA, 731–740.

[17] Michael D. Lieberman and Hanan Samet. Supporting Rapid Processing and Interactive Map-based Exploration of Streaming News. In *Proceedings of the 20th International Conference on Advances in Geographic Information Systems (SIGSPATIAL '12)*. ACM, New York, NY, USA, 179–188.

[18] Michael D. Lieberman, Hanan Samet, Jagan Sankaranarayanan, and Jon Sperling. Spatio-textual Spreadsheets: Geotagging via Spatial Coherence. In *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (GIS '09)*. ACM, New York, NY, USA, 524–527.

[19] Michael D. Lieberman, Hanan Samet, and Jagan Sankaranayananan. Geotagging: Using Proximity, Sibling, and Prominence Clues to Understand Comma Groups. In *Proceedings of the 6th Workshop on Geographic Information Retrieval (GIR '10)*. ACM, New York, NY, USA, Article 6, 8 pages.

[20] Michael D. Lieberman, Hanan Samet, and Jagan Sankaranayananan. Geotagging with local lexicons to build indexes for textually-specified spatial data. In *2010 IEEE 26th International Conference on Data Engineering (ICDE 2010)*. 201–212.

[21] Michael D. Lieberman, Jagan Sankaranarayanan, Hanan Samet, and Jon Sperling. Augmenting spatio-textual search with an infectious disease ontology. In *2008 IEEE 24th International Conference on Data Engineering Workshop*. 266–269.

[22] Alan Ritter, Sam Clark, Mausam, and Oren Etzioni. Named Entity Recognition in Tweets: An Experimental Study. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP '11)*. 1524–1534.

[23] Hanan Samet. Using Minimaps to Enable Toponym Resolution with an Effective 100% Rate of Recall. In *Proceedings of the 8th Workshop on Geographic Information Retrieval (GIR '14)*. ACM, New York, NY, USA, Article 9, 8 pages.

[24] Hanan Samet, Jagan Sankaranarayanan, Michael D. Lieberman, Marco D. Adelfio, Brendan C. Fruin, Jack M. Lotkowski, Daniele Panozzo, Jon Sperling, and Benjamin E. Teitler. Reading News with Maps by Exploiting Spatial Synonyms. *Commun. ACM* 57, 10 (Sept. 2014), 64–77.

[25] Jagan Sankaranarayanan, Hanan Samet, Benjamin E. Teitler, Michael D. Lieberman, and Jon Sperling. TwitterStand: News in Tweets. In *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (SIGSPATIAL '09)*. ACM, New York, NY, USA, 42–51.

[26] Axel Schulz, Aristotelis Hadjakos, Heiko Paulheim, Johannes Nachtwey, and Max Mhlhuser. A Multi-Indicator Approach for Geolocalization of Tweets. (2013). https://www.aaai.org/ocs/index.php/ICWSM/ICWSM13/paper/view/6063

[27] Faizan Wajid and Hanan Samet. CrimeStand: Spatial Tracking of Criminal Activity. In *Proceedings of the 24th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (GIS '16)*. ACM, New York, NY, USA, Article 81, 4 pages.

[28] Grady Ward. Moby Hyphenation List. *AG: General Works: Dictionaries and other general reference books* (May 2002).