Geotagging Named Entities in News and Online Documents

Jiangwei Yu
Department of Computing Science
University of Alberta
Edmonton, Alberta, Canada
jiangwei@ualberta.ca

Davood Rafiei
Department of Computing Science
University of Alberta
Edmonton, Alberta, Canada
drafiei@ualberta.ca

ABSTRACT

News sources generate constant streams of text with many references to real world entities; understanding the content from such sources often requires effectively detecting the geographic foci of the entities. We study the problem of associating geography to named entities in online documents. More specifically, given a named entity and a page (or a set of pages) where the entity is mentioned, the problem being studied is how the geographic focus of the name can be *resolved* at a location granularity (e.g. city or country), assuming that the name has a geographic focus. We further study dispersion, and show that the dispersion of a name can be estimated with a good accuracy, allowing a geo-centre to be detected at an exact dispersion level. Two key features of our approach are: (i) minimal assumption is made on the structure of the mentions hence the approach can be applied to a diverse and heterogeneous set of web pages, and (ii) the approach is unsupervised, leveraging shallow English linguistic features and the large volume of location data in public domain.

We evaluate our methods under different task settings and with different categories of named entities. Our evaluation reveals that the geo-centre of a name can be estimated with a good accuracy based on some simple statistics of the mentions, and that the accuracy of the estimation varies with the categories of the names.

1. INTRODUCTION

In the past few years, there has been a growing interest in both extracting entities from the web [9, 20] and associating geographical descriptors to web resources [1, 3]. The two lines of research point to an interesting but open issue of identifying the associations between named entities and their geographical boundaries in the web pages where the entities are mentioned.

Many named entities have a geographic centre or focus where the entity is better known or associated with; an organization may be tagged with a location where it is headquartered in; an artist may be associated with his/her hometown; a sport team may be identified by the location where it is based in; and a disease may be associated with a location where it is first reported. Sometimes a named entity may be associated with multiple locations, but its mentions in an article can often be traced to one location (usually the one that is more relevant to the readership of the article).

Exploring this relationship between a named entity and its geo-centre(s) has important implications. Knowing the geography of the named entities that appear in the matching pages of a query can help, for example, a search engine to better localize the search [18] or diversify the results. When "Shubert Theater" is mentioned in the matching pages of a query, knowing that the named entity refers to the theater in New Haven, Connecticut, rather than the one in New York City, may help an advertising system to supply more relevant ticket information. The geography information about entities, once collected, can be used to answer entity-based queries (e.g. "Where is entity X?") and location-based queries (e.g. "What is in location Y?"). The geography information may also be used to disambiguate entities in the same way it is used to disambiguate person names in official oath and declaration documents (e.g. "I <name> of <location> declare").

Existing approaches on relation extraction [2] can only extract a relation between a named entity and a location when they both appear in the same sentence or are linked using some surface text patterns. Toponym resolution methods (e.g. [15]) don't usually have this restriction but can only be applied when the named entity is a location.

Obtaining accurate and effective geographical information for general named entities on the Web can be challenging for several reasons. *First*, many different semantic relationships can exist between a named entity and a location, and it is not possible to bound the relationship to a sentence or some predefined templates (e.g. *headquarters* and *birth place* [16]). *Second*, there is a large variation in the scope and the geographical spread of named entities. While many entities are only known in certain regions with limited boundaries (e.g. municipal and provincial politicians, schools, small companies and organizations, etc.), some entities with a large spread may not be bounded to a city or a province (e.g. federal ministers, heads of states, etc.). It should be noted that entities that are better known globally may not have a fixed geo-boundary; examples are well-known celebrities (e.g. Lady Gaga), multi-national organizations and companies (e.g. ACM, McDonald), well-publicized products (e.g. Macbook Air), etc; detecting such names or possibly geo-

CIKM'16 , October 24-November 28, 2016, Indianapolis, IN, USA
© 2016 ACM. ISBN 978-1-4503-4073-1/16/10...$15.00
DOI: http://dx.doi.org/10.1145/2983323.2983795

tagging them is out of the scope of this paper, but we can imagine that one may compile a list of such names (given that they are well-known) and totally avoid the problem of geo-tagging. *Finally*, both entity and location mentions in web pages can be ambiguous. For example, the term "Springfield" by itself can refer to 30 different cities in the USA and names like Houston and Dallas can be both person names and place names.

Problem Formulation. In this paper, we study the problem of geo-tagging the mentions of named entities in web pages. A problem that often arises in working with locations is the overlap and inclusion relationships between locations (e.g. Phoenix, Arizona, and USA); in our case, a geo-centre can be at any or all of country, state or province, and city levels. We assume these relationships are described in a tree structure with an inclusion relationship between parents and children. For example, the root of the tree may represent the whole world, with the nodes in the second, the third and the fourth levels representing respectively countries, states and cities; such a tree is often referred to as a *gazetteer*. The problem can then be formulated as:

Given the mentions of a named entity n in a document or a set of documents P, and a tree structure of locations (gazetteer), resolve the geo-centre of n, by identifying a node or nodes in the tree that best represents the location of n in document(s) P.

This is a *document-centric* view of a geo-centre, where the orientation, and not the type, of a relationship is sought. It may be treated as more general than a *type-centric* view (e.g. [16, 5]) since the type of the relationship is not always known or may not be important. Also defining a geo-centre with respect to a document (or documents) can narrow down the scope of the name and its geo-centre, making it less ambiguous. For example, entities can have multiple and maybe non-overlapping geo-centres (e.g. places of birth and work), but the mentions of those entities in a given document(s) may well focus on one aspect resolving in a single location.

We develop a probabilistic model that assigns geo-centre(s) to a name based on the geo-centre of the pages that mention it and the distribution of location references in the proximity of the name. The model is unsupervised, and allows ambiguity at the location level with the probability mass distributed over all candidate locations. We evaluate our model and its variations on a gold standard set of names with known geo-centres and a set of pages that mention those names. Our contributions can be summarized as follows.

- We formalize the problem of geo-tagging named entities in documents and study the interactions between a geo-centre and the spread.
- We propose a framework for detecting the geo-centre of named entities cited in documents, based on a location-aware generative model of documents.
- We devise an algorithm that estimates the dispersion of a name, and show that a geo-centre can be detected at an exact dispersion level.
- We evaluate our algorithms and formulations on a real dataset composed of various classes of named entities and their mentions in web pages, and report the performance results, broken down to different algorithms and different classes of named entities.

The remainder of this paper is organized as follows. Section 2 presents both the model and an overview of our framework. Extracting location candidates from web pages and linking them with a gazetteer is addressed in Section 3. A probabilistic framework for estimating the geo-centre at a specified dispersion level is presented in Section 4 and evaluated in Section 5. We review related work in Section 6 and conclude the paper in Section 7.

2. THE MODEL

Consider a document generative model with a vocabulary V; let V is partitioned into a set of location terms V_L and a set of non-location terms $V - V_L$. Each document can be generated with respect to a location, referred to as the *geo-centre* of the document, to indicate the relevance of the document content to the location. For example, the geo-centre of a narrative story can be set in a place where the events in the story take place; a news article may be generated with respect to the location of its readership. Given a set D of documents and a named entity n, the probability that document $d \in D$ generated in the context of a location $l \in V_L$ mentions a named entity n can be written as

$$P(d|l,n) = \frac{P(l|d,n)P(d|n)}{P(l|n)}. \tag{1}$$

As a geo-centre of named entity n in document d, we want to estimate $P(l|d,n)$. The language model establishes a relationship between the probability of generating a document and the probability that l is a geo-centre of n in that document, but we still need to make some assumption to do our estimation. One hypothesis is that *each named entity mentioned in a document inherits the geo-centre of the document*, i.e. $P(l|d,n) = P(l|d)$. We refer to this as *the inheritance hypothesis*. Based on this hypothesis and using maximum likelihood estimation, the probability of generating l under a unigram model is

$$\hat{P}(l|d,n) = \frac{tf_{l,d}}{\sum_{l' \in V_L} tf_{l',d}} \tag{2}$$

where $tf_{l,d}$ is the frequency of location term l in d. Not all named entities mentioned in a document inherit the geo-centre of the document. Also long documents may have multiple geo-centres and only one of those may be relevant to a named entity. Our second hypothesis, referred to as *near-location hypothesis* states that *the geo-centre of a named entity must be mentioned in nearby text.* This has to be interpreted in the context of the first hypothesis in that when a named entity does not inherit the geo-centre of the page, it is expected to be qualified explicitly with a geo-centre; otherwise the named entity can be ambiguous [1]. For example, an article published in San Francisco Chronicle can cite names which are not related to the geo-centre of the article or the newspaper and those names can be ambiguous if not explicitly qualified. Section 4.1 gives an estimate based on this hypothesis and looks into ways of combining the two estimates.

2.1 Hypothesis testing

The two hypotheses need some verification. For this purpose, we randomly selected 30 headline articles from six different newspaper websites in North America; each newspaper had a regional focus which was also expected to be

[1] This rule may not hold for more global names with no fixed geo-centre, but tagging those names is out of the scope of this paper

the focus of the published articles. We manually inspected each named entity tagged as PERSON or ORGANIZATION by the Stanford Named Entity Recognizer [11], and excluded 9 named entities that either did not appear to have a regional orientation, or the geo-centre spanned over more than one country [2]. The remaining 162 named entities had their geo-centres mentioned in the same pages, and accordingly a geo-centre was assigned to each. We then placed each named entity into one or more of the following three bins, based on the relation between the named entity and its geo-centre: **A)** The named entity inherits the geo-centre of the page; **B)** In the same sentence where the named entity is mentioned, either the geo-centre of the named entity is mentioned, or there is a name with the same but known geo-centre mentioned; **C)** The named entity has a geo-centre, but neither A nor B holds. Note that a named entity can be placed in both bins A and B.

Among the 162 names analyzed, there were 51 (31.5%) placed only in bin A; 44 (27.2%) placed only in bin B; and 65 (40.1%) placed in both bins A and B. In other words, in 71.6% of the cases *the named entity inherits the geo-centre of the page*, supporting our **inheritance hypothesis**. In 67.3% of the cases *the geo-centre (or a name with the same but known geo-centre) is mentioned nearby*, which can be treated as an evidence supporting our **near-location hypothesis**. Only 2 of 162 (1.2%) cases did not fall into any of the bins A and B.

Based on this and other evidence, one can observe that the two hypotheses are capturing the intrinsic rules for introducing a named entity with certain regional orientation adopted by most of the newswire article writers. We are not expecting the same level of agreement in general web pages, and this is an area where more studies are perhaps needed. That said, we assume the writers and editors of non-newswire web pages more or less follow the same or similar guiding rules and principles; this is also consistent with the findings that document features may be beneficial in disambiguating location names [17].

Overview of the framework Our framework consists of three stages. First, locations mentioned in each page are extracted and translated into canonical forms. For example, a mention of *Edmonton* is translated into *Edmonton, Alberta, Canada* or other locations[3] depending on its context. Such preprocessing addresses the problem of location ambiguity in web pages and has been successfully used in the past [1, 14]. Second, a geotagging of the named entity is performed at one of city, state or country levels, using our models and based on the clues gathered in the previous stage and possible relationships to the named entity. Finally, one of the city, state and country levels is selected as the exact level for the geo-centre of the named entity, based on the results for each level in the second stage.

3. FINDING CANDIDATE LOCATIONS

As our geotagging is based on the mentions of a target named entity and candidate locations in web pages, extracting accurate location information plays a crucial role in the performance of the whole work. The source of a web page often contains HTML tags and potentially scripts that are not exploited by our framework. We use the *Keep Everything Extractor* in the boilerpipe library[4] [12] to extract the full text of the pages.

3.1 Mentions extraction

To extract mentions of locations, we use the Stanford Named Entity Recognizer to tag potential locations in text. A side effect of running the recognizer is that the text is tokenized into a sequence of terms. A mention is a subsequence of this sequence described by the indices of its first and last terms. In this work, we extract three kinds of mentions:

Mentions of a target named entity. Similar to text, the target named entity is tokenized, and its mentions in text are identified.

Mentions of locations. We consider consecutive terms tagged as LOCATION by the named entity recognizer as mentions of locations. We will explain the disambiguation of these mentions in Section 3.2.

Mentions of adjectival and demonym forms of locations. We collect the adjectival forms of countries [5] and states in the U.S. [6]. Mentions matching these forms are resolved to the corresponding locations (e.g., mentions of *Canadian* are considered the same as mentions of *Canada*). The matching rules for mentions of a target named entity apply here.

Note that we do not allow the mentions to overlap (e.g. *Edmonton Oilers* and *Edmonton*, both at the same offset). The extraction is done in the order described above, which implies that we ignore potential locations and demonyms embedded in the target named entity. It should be noted that such location mentions inside a named entity are useful clues that should be exploited. However, we decided against using them in our experiments for two reasons: (1) to report performance results that are less dependent on the selection of the entities, and (2) to possibly underestimate (but not overestimate) the performance of our system.

3.2 Location disambiguation

An extracted mention of a location can be ambiguous, meaning that it may not refer to a unique location, for example, in a gazetteer.

Location disambiguation (and similarly *toponym resolution*) is out of the scope of this work, and any relevant method from the literature may be used. A well-cited work is that of Amitay et al. [1], which applies a set of rules (*e.g.*, checking if a location is qualified by another location) to assign a confidence score to each resolved location. We decided to implement our own method since we were not sure how the weights could be assigned to different locations in Amitay et al.'s work. In particular, for each mention of a possible location (tagged by the NER tool), we search the text of the mention (this can be a multi-word term) in a database of geographic entities [7]; the database has the canonical description of each location and is structured as a tree to describe the containment relationships between loca-

[2] The excluded names are Lorna Morello, Alex Vause, NHL, UN Security Council, United Nations, Red Cross, International Energy Agency, Jupiter, and Kepler.

[3] http://en.wikipedia.org/wiki/Edmonton_(disambiguation)

[4] https://code.google.com/p/boilerpipe/

[5] http://en.wikipedia.org/wiki/List_of_adjectival_and_demonymic_forms_for_countries_and_nations

[6] http://en.wikipedia.org/wiki/List_of_demonyms_for_U.S._states

[7] This was the Geoname database in our setting.

tions. Our search returns a list of possible matches, ordered based on population, with the top k selected as candidates [8].

Given an unresolved location with surface text s (e.g., Edmonton), its set of mentions $M(s)$, and a set of location candidates L, a confidence score for that location $l_i \in L$ resolves s is defined as

$$DS(l_i, s) = \sum_{c \in C(l_i)} \sum_{mc \in M(c) - M(s)} \max_{ms \in M(s)} \frac{1}{D_M(ms, mc)},$$

where $C(l_i)$ is the set of constituent terms of l_i. For example, the constituent terms of the city of Edmonton in Alberta, Canada are *Edmonton*, *Alberta*, and *Canada*. We enumerate over all constituent terms of a location and for each constituent c, we look for each of its mentions mc in the page. If mc is not a mention of s, we treat mc as a clue to resolve s. For example, *Alberta* is a clue to resolve *Edmonton*.

To account for the distance between mentions and that near mentions are more important, the contribution of mc is calculated as the reciprocal of the minimum term distance between mc and the mentions of s. The term distance is formally defined as the number of terms that fall between two mentions, i.e.

$$D_M(m_1, m_2) = \min\{|t(m_2) - s(m_1)|, |t(m_1) - s(m_2)|\}. \tag{3}$$

where $s(m)$ and $t(m)$ respectively denote the start and the end indexes of mention m. By the one-sense-per-discourse principle, we resolve all mentions of s to the canonical location l that has the maximum confidence score $DS(l, s)$ for $l \in L$. In case of a tie, the most populated candidate location is chosen, as done in some early work [1].

4. DETECTING GEOTAGS

Given a named entity, a collection of relevant pages with mentions of the named entity, a set of disambiguated locations in each page as candidates, we are set to determine the most relevant location of the target named entity. A caveat is that not all locations are at the same level of dispersion and because of the containment and overlap relationships between locations, a comparative ranking may not be possible. In this section, we first assume a level of dispersion is known or given, thus all candidate locations are at the same level of dispersion (e.g. city). Having developed a solution under this setting, we then address the problem of detecting a right level of dispersion for a named entity.

4.1 Geotagging at the city level

We start geotagging at the lowest level of dispersion, i.e. the “city” level in our gazetteer. This is due to the fact that for geotagging at higher levels such as country or state, one should not ignore the lower-level locations such as city and that the mentions of such locations support the higher-level locations they belong to.

Our hypotheses, as given in Section 2.1, specify two sources where the geo-centre of a named entity can be identified. Accordingly, we propose two models for geotagging a named entity, each tapping into one of these sources, before combining them into a single model.

Inheritance based model of geo-centre Our inheritance hypothesis suggests a method for estimating the geo-centre of a named entity, provided that the geo-centre of the page that mentions the named entity is known. Any page geotagging method can be used here (as some reviewed in Section 6), and our maximum likelihood model in Equation 2 also provides a way to estimate a geo-centre. In a simple evaluation of this estimate, we used the 30 headline articles reported in Section 2.1 and assigned a geo-centre to each based on a manual inspection. The assignment took into account the relevance of the events reported in the page and the expected readership of the page. We found that in 21 of the 30 cases the geo-centre of the page was the location where the newspaper was published. And among these 21 pages where we were sure of their geo-centres, we found that in 15 pages (71%), the location identified by the model was either the geo-centre or part of the geo-centre of the page. Based on this observation and the hypothesis that named entities inherit the geo-centre of the page that mentions them, a geo-centre for named entities can be identified.

Near location based model of geo-centre Our study, as reported in Section 2.1, showed that in 67.3% of the cases the geo-centre is mentioned near the named entity in the same sentence for purposes such as introduction and disambiguating the named entity, following our near-location hypothesis. A document cannot always be broken down into sentences especially if it is not a well-written piece of text; a better way of describing this relationship is in terms of the distance between the mentions of a candidate location and a named entity.

Given a document d, a named entity n mentioned in d, and a set L of locations that are also mentioned in d, $P(l|d, n)$ can be estimated as

$$\hat{P}(l|d, n) = \frac{\frac{1}{D_E(l,n)}}{\sum_{l' \in L} \frac{1}{D_E(l',n)}}, \tag{4}$$

where $D_E(l, n)$ is the minimum term distance between mentions of l and n, referred to as *Entity Distance*. The relevance scores are normalized to a number between 0 and 1, so they can be interpreted as probabilities.

The Entity Distance between two named entities n_1 and n_2 can be defined as

$$D_E(n_1, n_2) = \min_{\substack{m_1 \in M(n_1), \\ m_2 \in M(n_2)}} D_M(m_1, m_2), \tag{5}$$

where $D_M(m_1, m_2)$ is the term distance between mentions m_1 and m_2, as defined in Eq. 3, and $M(n)$ is the set of mentions of an entity n.

A mixture model A problem in using any single model is that we are not certain if the premise of the model holds. For example in our case, we don’t know if a named entity inherits the geo-centre of the page or not. One way to address this is to use some sort of a mixture model. A question to be addressed here is what should be the mixture model and how the two models should be weighted. Let $\hat{P}_{inh}(l|d, n)$ be the estimate obtained using the inheritance model (as defined in Eq 2) and $\hat{P}_{near}(l|d, n)$ be the estimate obtained using the near-location model (as in Eq 4).

Consider the extreme case where the values of $\hat{P}_{near}(l|d, n)$ are evenly distributed for every location l mentioned in d; this suggests that either the name does not have a unique geo-centre or a unique geo-centre cannot be identified using $\hat{P}_{near}()$. Under this setting, two questions that arise are: (1) how much should we rely on these values when the

[8] In our experiments, k is set to 5.

name actually has a clear geo-centre? (2) Can we opt for another measure which can provide a larger margin between the probabilities?

As a measure of the non-uniformity of the ranks and to indicate that one or more locations stand out from the rest in document d, we introduce $J(d, n)$ which is defined in Eq. 6 in terms of the Shannon Entropy of the vector induced by $\hat{P}_{near}(l|d, n)$ for different values of l. The non-uniformity here is based on our near-location model $\hat{P}_{near}$. $H(d, n)$, as defined in Eq. 7, is the distance-based entropy of the probabilities distributed over the cities mentioned in d, and $H_{max}(p)$ is the maximum entropy over locations in d, which is $\log |L|$, achieved when all locations have the same probabilities.

$$J(d, n) = 1 - \frac{H(d, n)}{H_{max}(d)} = 1 - \frac{H(d, n)}{\log |L|} \tag{6}$$

$$H(d, n) = -\sum_{l \in L} \hat{P}_{near}(l|d, n) \log \hat{P}_{near}(l|d, n) \tag{7}$$

When the gap between the maximum probability and the second largest probability is large, $H(d, n)$ is small and $J(d, n)$ is close to 1, indicating a strong tendency toward the model that is based on term distance for correctly capturing the geo-centre. Conversely, $J(d, n)$ is close to zero when the entropy value approaches its maximum, meaning that the top values of $\hat{P}_{near}(l|d, n)$ are close and the model may not be effective in detecting a geo-centre.

Now a combined model can be written as the mixture:

$$\hat{P}(l|d, n) = J(d, n) \cdot \hat{P}_{near}(l|d, n) + (1 - J(d, n)) \cdot \hat{P}_{inh}(l|d, n). \tag{8}$$

The first term of the mixture characterizes the joint probability of two events: 1) the near location based model provides a correct estimate with probability $J(d, n)$; 2) l is the geo-centre, estimated by the near location based model with probability $\hat{P}_{near}(l|d, n)$. The second term of Eq. 8 characterizes the situation where the near location based model cannot capture a unique geo-centre, and the target named entity inherits the geo-centre of the document.

It is noteworthy that we have experimented with other models as well, such as replacing $J(d, n)$ with an inheritance-based entropy; some of these variations are reviewed and evaluated in Section 5.3.

4.2 Geotagging at higher levels

The problem of geotagging arises not only at the city level, but also at higher levels for names that are more widely known. To find out the geo-centre of a named entity n at, say the state level, we consider both the mentions of cities and states (provinces) in the page.

An aggregation of the scores at city and state levels can be done in two steps. *First*, the mentions of both state-level and city-level locations are considered and the relevance of each location is estimated based on the proposed models in Section 4.1. Specifically, the relevance is computed over the union of city and state level locations as the domain of discourse. *Second*, for each state level location ls, the relevance score of the cities that belong to ls are added to the state-level score; this give a relevance score that ls is the geo-centre of n at the state level. If a state level location is not mentioned in d, but its cities are mentioned, its score will be the sum of its cities' score. The relevance score can be computed using any of the models presented in Section 4.1. One can verify that the relevance scores for all state-level locations sum up to 1, forming a probability distribution.

We can easily generalize this method to even higher levels, such as country, by propagating the probability mass of states and cities to that of the countries they are located in.

4.3 Corpus aggregation

Given a named entity n and a set of relevant documents D, one may treat each document as an independent evidence with equal weight. Under this setting, the probability that location l is the geo-centre of n, namely $\hat{P}(l|n)$, can then be measured as the expectation of $\hat{P}(l|d_i, n)$ for $d_i \in D$, i.e.

$$\hat{P}(l|n) = \frac{1}{|D|} \sum_{d_i \in D} \hat{P}(l|d_i, n). \tag{9}$$

Assuming that the geo-centre is unique, the location with the maximum value of $\hat{P}(l|n)$ can be declared as the geo-centre of n.

4.4 Location refinement

When the probability distribution of a named entity is known at the country level, we can use it as a priori for the estimation of the state-level geo-centre. The underlying assumption is that the probability of a location to be the geo-centre at a fixed level of dispersion can be affected by the mentions of its parent locations in the tree structure of locations. Let ct, st and cn be locations at city, state and country levels respectively, and st be the parent of ct and cn be the parent of st. Taking the prior information into account, the geotags at state and city levels may be refined as $\hat{P}(cn|n) \cdot \hat{P}(st|n)$ and $\hat{P}(st|n) \cdot \hat{P}(ct|n)$ respectively.

4.5 Detecting the dispersion level of a name

We have so far assumed that a level of dispersion is either known or given, and now we want to relax this condition and determine most suitable level of dispersion for a given named entity among the three levels: city, state, and country.

Inspired by the idea that the geo-centre is more likely to be unique at the desired level of dispersion, we aim to find $v \in \{city, state, country\}$ that maximizes the expectation of $J^v()$ over the set of documents D, i.e.

$$\arg\max_v \frac{1}{|D|} \sum_{d_i \in D} J^v(d_i, n). \tag{10}$$

$J^v()$, as defined in Eq. 6, measures the probability that a unique geo-centre can be detected at level v and is assessed in terms of $\hat{P}()$ with L set to the locations at level v; but it can as well be measured in terms of $\hat{P}_{inh}()$ or $\hat{P}_{near}()$ (and we evaluate them in our experiments).

5. EXPERIMENTS AND EVALUATION

We evaluate the performance of our algorithms in terms of the accuracy of the results on various datasets. Our evaluation includes geotagging at city, state and country levels and on the basis of the mentions of a name in a single page or a set of pages. In each case, we present a comparison of our results to a few baselines.

5.1 Gold standard datasets

A full coverage of all kinds of named entities is not realistic, hence we focus on three sets of proper names that are common in web pages: names of "persons", "locations" and "organizations", according to MUC-6[9] types.

The first set consists of *person names* with a geographic boundary at the granularity of country, state, or city. As politicians usually have a clear level of administration, we can obtain the geo-centre of politicians with high confidence. We collected names of heads of states in the world[10] into the set of country level politicians, where the country of the politician is the ground truth. We also collected names of politicians in Canada. Names of governors[11] and party leaders[12] of provinces and territories are categorized as names at the state/province level. Names of city mayors[13] and councillors[14] are categorized as names at the city level.

The second set is the names of *implicit physical locations*, which are usually tagged as location named entities by named entity recognition tools. These kinds of names are very indicative of the location information, especially with the common use of social networks on mobile phones, where user posts often contain these names. Understanding the city level geo-centre of these names is important. To evaluate the performance of our geotagging framework on implicit physical locations, we collected the names of museums, theatres and towers in the United States. We considered the city where each entity was located as the geo-centre of the entity. The location of an entity was determined based on its longitude/latitude information from the Geonames database and through mapping the longitude/latitude pairs into cities with the Google Map API[15].

The third set is the names of *organizations*, including universities, sports teams and technology companies. Names of universities and their locating cities were extracted from the list of Top 100 U.S. Universities by U.S. News[16]. Names and home cities of sports teams of four major sport leagues in the North America (NHL, NBA, MLB and NFL) were extracted from the official website of each league. From CrunchBase[17] we collected names of technology companies founded after 2008 with Series C funding (which means they are likely to be known by the public) and their headquarters at the city level. Similar to the implicit physical locations, we consider the city where an organization is located as the geo-centre of that organization.

Our dataset consisted of 101 theaters, 100 museums, 100 towers, 100 universities, 204 companies, 123 sport teams and 125 politicians. For each collection of names, we gathered related pages from December 2013 to March 2014 by searching the names in the search engine Exalead[18] with the names as queries. For each name, (up to) the top 30 pages returned by the search engine are used to build our dataset. The dataset used in our experiments is all available online [19].

5.2 Evaluation settings

Based on the data reported in Section 5.1, we collect pages that mention both a target named entity and its ground truth location in one of the city, state or country levels. Unlike a city-level geotagging where the ground truth is expected to be mentioned at the city level, our state- and country-level geotagging may include pages that mention the ground truth at lower levels only. Furthermore, in our geotagging of page corpora, where we want to evaluate the ability of our algorithms in aggregating the results of different pages, we drop a corpus if it contains less than 5 pages simply because a comparison between different strategies is less meaningful.

For the task of geotagging with single pages, each page is considered as a data point and for geotagging at the corpora level, each corpus of a named entity is considered as a data point. For each data point, if the geo-centre estimated by an algorithm matches one of the ground truth locations of the target named entity, we treat the answer correct. With this setting, we can compute the accuracy by dividing the number of correct answers, denoted by T_c, by the total number of data points in the dataset, denoted by T:

$$Accuracy = \frac{T_c}{T}. \tag{11}$$

5.3 City level geotagging with single pages

Our city-level experiments evaluate the performance of the models proposed earlier as well as a few baselines, as reported here. The uppercase letters in the parentheses are the short names used in Table 1a to refer to the models.

- **Random (RAND):** The random model assigns equal probabilities to all candidate locations in a page. This model is used for sanity check and to assess the performance of other models against chance.
- **Frequency (FREQ):** This is the model defined in Eq. 2; it is based on the hypothesis that named entities inherit the geo-centre of the page where they are mentioned.
- **Term Distance (TD):** This is the model defined in Eq. 4, which is based on the principle that nearby names tend to share a common orientation. This model also serves as a baseline disambiguation model in the style of Lesk's word sense disambiguation [13].
- **Mixing by Multiplication (MM):** This model is based on the following two events: (1) the geo-centre appears near the target named entity and (2) the named entity inherits the geo-centre of the page. Assuming independence, the model takes the probability of the conjunction, i.e. the product of the values of Eq. 2 and Eq. 4.

[9] http://cs.nyu.edu/faculty/grishman/muc6.html

[10] http://en.wikipedia.org/wiki/List_of_current_heads_of_state_and_government, visited on Mar 11, 2014

[11] http://en.wikipedia.org/wiki/Provinces_and_territories_of_Canada, visited on Sep 10, 2013

[12] http://www.parl.gc.ca/Parlinfo/compilations/ProvinceTerritory/PartyStandingsAndLeaders.aspx, visited on Sep 10, 2013

[13] http://www.fcm.ca/home/about-us/big-city-mayors-caucus.htm, visited on Sep 10, 2013

[14] http://www.edmonton.ca/city_government/city_organization/city-councillors.aspx, visited on Sep 10, 2013

[15] https://developers.google.com/maps/documentation/geocoding/#ReverseGeocoding, visited on Dec 30, 2013

[16] http://colleges.usnews.rankingsandreviews.com/best-colleges/rankings/national-universities/, visited on Jan 11, 2014

[17] http://crunchbase.com/search/advanced/companies/2144281, visited on Dec 30, 2013

[18] The engine (http://www.exalead.com/search/web/) was selected since it did not block our queries unlike more well-known engines and also because of the simple "keyword-in-document" model of our queries.

[19] http://www.cs.ualberta.ca/~drafiei/datasets/geotagging

- **Mixing by Addition (MA):** This model is also based on the two events described for MM, except that the mixture model is defined as the mean of the two probabilities.
- **Mixing based on the frequency entropy (MFE):** It is similar to the definition in Eq. 8, with the difference that the non-uniformity $J(d, n)$ is computed based on $\hat{P}_{inh}(l|d, n)$ instead of $\hat{P}_{near}(l|d, n)$.
- **Mixing based on distance entropy (MDE):** This is our proposed model in Eq. 8.

The results are shown in Table 1a. The scores in bold indicate the best accuracy achieved among all the models for the corresponding category. The model based on frequency alone (`FREQ`) achieves the best performance on Sports Teams with an accuracy of 0.585. It is noteworthy that the surface text of named entities in the sports category often include the home city or state of the team. We ignore this information to evaluate the ability of our approach for capturing other clues of the geo-centres mentioned in the page. Because of this setting, when estimating the geo-centre of a team, it may be more promising to look for clues from the page geo-centre than finding locations nearby, as the home location is less likely to be mentioned again nearby.

However, `FREQ` does not perform well on other categories especially on the category of technology company names, with an accuracy of only 0.370. An examination of the pages revealed that many of the pages related to a company are focusing on the business end of the company. As the companies in our dataset are about technology, their business might not be limited to the areas of their home offices, and this makes it less likely for the main topic of the page to have a strong location indication. Instead, the geo-centre is often mentioned near the mention of the company name for readers to gain knowledge about the company's location, which is indicated by the results of the model `TD`, whose accuracy (0.630) is the best among those investigated.

As we can see, models that are only based on the term distance or frequency may perform well in one category but bad in another. In contrast, the mixed models are more balanced. For categories of landmark names (theatres, museums, towers, and universities [20]) and person names (politicians), `MA`, `MM`, `MFE` and `MDE` are superior to the other models.

Among these mixed models, `MDE` is a robust one. It has the highest overall accuracy and performs the best in the categories of museums and towers. In our other categories, its performance is also comparable to the best ones. Conversely, the remaining three mixed models all fall behind `MDE` in the categories of towers and sports teams. The model `MFE` achieves an accuracy of 0.628 (the second best) in company names but falls short in sports teams (0.447) and towers (0.609) when compared with the other three mixed models. The model `MM` has a close performance compared to the model `MDE`, but falls 2.3% behind in the category of sports teams, which amounts to 21 pages in 899 pages.

Overall we find that the relation between the mentions of named entities and geo-centres should be modelled using both the term distance and the frequency, and that one measure may play a more important role when the other measure cannot estimate the geo-centre with a good confidence. The experimental results show that the proposed model `MDE` defined by Eq. 8 is more reliable than the others. Hence we will use it in the rest of our reported experiments.

5.4 State level geotagging with single pages

We experimented with our proposed algorithm as well as a series of baselines for state level geotagging with single pages. All algorithms leverage the probabilistic model `MDE` to assign initial probabilities to the candidate locations, with different aggregation approaches explained below.

- **States only (S):** As the name suggests, only the state names are plugged into the model `MDE` for ranking. In other words, this model treats state names as abstract terms and does not take into account possible containment relationships between state names and city names mentioned in the same page.
- **Cities only (C):** Each state is defined in terms of the cities that it contains, hence the probability of each city in a page is assessed using the model `MDE`, and then the probabilities of cities in the same states are added up to form a probability distribution for states. The ranking of states is determined by these new probabilities.
- **Maximum of S and C (MSC):** For each state-level location and its probabilities given by algorithms `S` and `C` described above, the algorithm `MSC` takes the maximum and divides it by two to maintain a valid probability distribution over the candidate locations.
- **Average of S and C (ASC):** This algorithm is similar to `MSC`. The difference is that we take the mean of the probabilities given by `S` and `C`.
- **Analyzing mentions of states and cities simultaneously (AMS):** This is the proposed algorithm in Section 4.2.

Table 1b reports the accuracy of applying each algorithm to different categories. We can see that the algorithm that only considers mentions of cities (`C`) performs better than the one that only analyzes mentions of states (`S`) in categories of theaters, universities, companies, sports teams and politicians, which suggests that city level locations play a significant role even in geotagging at the state level.

The above results show that our proposed algorithm (`AMS`) outperforms the baselines in all categories. The reason for this difference in performance can be probably explained as follows: 1) Either `C` or `S` only considers one level, which is less thorough compared to `AMS`; 2) When the numbers of mentions at different levels are not balanced, the baselines `MSC` and `ASC` may overestimate the relevance for a location whose level has few candidates, while the proposed algorithm `AMS` keeps such difference by distributing original probabilities to both levels simultaneously.

We can also see that there is quite some gap between the accuracy of our proposed algorithm (`AMS`) and that of `MSC` and `ASC` in all categories except the category of museums. This is because of the fact that when the numbers of mentions at different levels are not balanced, the algorithms `MSC` and `ASC` can overestimate the probability for a location whose level has few candidates and this can lower their accuracy. On the other hand, `AMS` keeps such difference by distributing original probabilities to both levels simultaneously. We will use `AMS` as our algorithm of choice in the rest of our experiments.

[20] A university can be considered both as a landmark or an organization.

Table 1: Accuracy of geotagging with single pages using different models.

(a) City level

	RAND	FREQ	TD	MA	MM	MFE	MDE
Theater	0.230	0.597	0.752	0.755	**0.757**	0.750	0.745
Museum	0.190	0.601	0.645	0.691	0.689	0.673	**0.694**
Tower	0.228	0.620	0.549	0.641	0.663	0.609	**0.696**
University	0.238	0.572	0.664	0.694	**0.713**	0.677	0.692
Company	0.306	0.370	**0.630**	0.626	0.626	0.628	0.623
Sports Team	0.187	**0.585**	0.370	0.532	0.539	0.447	0.562
Politician	0.239	0.836	0.838	**0.890**	0.881	0.860	0.883
Overall	0.225	0.594	0.606	0.672	0.677	0.638	**0.681**

(b) State level

	S	C	MSC	ASC	AMS
Theater	0.575	0.765	0.773	0.775	**0.844**
Museum	0.806	0.819	0.873	0.878	**0.904**
Tower	0.778	0.772	0.820	0.827	**0.877**
University	0.621	0.746	0.730	0.743	**0.807**
Company	0.366	0.652	0.566	0.571	**0.673**
Sports Team	0.345	0.589	0.472	0.483	**0.598**
Politician	0.618	0.751	0.805	0.803	**0.869**
Overall	0.564	0.718	0.702	0.708	**0.782**

5.5 Aggregating results of different pages

For each named entity and its set of relevant pages, we experiment with a few algorithms to aggregate the scores from individual pages. In particular, we evaluate the following baselines in addition to the algorithm proposed in Section 4.3.

- **Maximum probability of the location (MP):** This model is based on the assumptions that (1) at least one page can correctly resolve the geo-centre of a name and (2) a correct geo-centre has the highest rank among candidates. Thus, candidate locations are ranked based on their maximum probabilities in the set of relevant pages.
- **Number of pages that mention the location (NP):** This algorithm counts the number of pages that a candidate location is mentioned and ranks the locations in a descending order by these frequencies.
- **Product of MP and NP (MP·NP):** Candidate locations are ranked based on the product of the values in MP and NP. The intuition is that the geo-centre should be mentioned in many pages and in some pages it should be ranked high by our page-level geotagging.
- **Product of probabilities (PP):** This algorithm considers each $d_i \in D$ as a test for the event that a candidate location l is the geo-centre, whose probability is defined as $\hat{P}(l|d_i, n)$. The product of $\hat{P}(l|d_i, n)$ for $d_i \in D$ measures the joint probability of a location passing all the tests, assuming independence between documents in D. For smoothing, let λ be be the minimum positive value of $\hat{P}(l|d_i, n)$ for all $d_i \in D$ and l. When a location is not mentioned in a document, we use λ as the probability instead. Locations are then ranked according to the following quantity:

$$\hat{P}_{PP}(l|n) = \left(\prod_{i=1}^{|D|} \max\{\hat{P}(l|d_i, n), \lambda\} \right)^{\frac{1}{|D|}}. \qquad (12)$$

- **Average of probabilities (AP):** Locations are ranked according to Eq. 9.

The results for the city and the state levels are respectively shown in Table 2a and Table 2b. We can see that the algorithm `MP`, which takes the maximum of page-level probabilities, has the worst performance. This is because by taking the maximum the results only reflect the characteristics of one page. But in other algorithms this problem is alleviated by using metrics that capture characteristics of all pages.

From Tables 2a and 2b we can also see that the algorithms `AP` and `PP` achieve the best overall results. They outperform the other methods in all categories, which suggests that the relevance of a location measured in single pages can be effectively combined under the assumption of these two models (i.e. the independence assumption in `PP` and the equal importance assumption in `AP`).

Since `AP` is comparable to `PP` in terms of the city level geotagging and performs slightly better at the state level, we select `AP` as our algorithm for aggregating the results of different pages.

5.6 Location refinement

In Section 4.4, we proposed a refinement for adjusting the probability of a location by taking the probability of its parent location as a priori. Table 3 compares the accuracy with and without this refinement step in different categories, using `AP` for aggregating results from different pages. We can see that there is no clear winner between the two methods. The refined method has a relatively higher accuracy at the city level but performs worse at the state level. An explanation is that we have been already exploiting the containment relationships between locations in our location disambiguation (Section 3.2), and using the same relationship again does not provide additional evidence to further refine the locations.

5.7 Selecting the most relevant level

This experiment is conducted on politicians (see Sec. 5.1) for the reason that a politician has a clear dispersion in terms of their serving regions. Similar to our previous experiments on a corpus, we removed names with less than 5 relevant pages, resulting in 44 names at the city level, 30 at the state level and 32 at the country level.

The performance is measured in terms of accuracy, as defined in Eq. 11, where T is the total number of names and T_c is the number of names for which our algorithm correctly identifies the geo-centre at the exact level of dispersion.

Methods We experiment with the algorithm given for determining the level of dispersion in Section 4.5 as well as a few baselines. The details are illustrated below. Unless it is stated otherwise, $v \in \{city, state, country\}$.

- **Total Locations (TL):** This model assumes the most relevant level is dominant. For a page P_i, let the number of distinct locations at level v be $TL(i, v)$. This method adds up the values of TL for all pages at each level. Then the level with the most distinct locations is chosen:

Table 2: Accuracy for geotagging with document corpora using different aggregation methods.

(a) City level

	MP	NP	MP·NP	PP	AP
Theater	0.426	0.660	0.723	**0.745**	**0.745**
Museum	0.349	0.619	**0.651**	**0.651**	0.603
Tower	0.341	0.488	0.488	**0.537**	0.512
University	0.444	**0.819**	0.806	**0.833**	0.819
Company	0.286	0.610	0.638	0.619	**0.648**
Sports Team	0.480	0.730	0.810	0.810	**0.830**
Politician	0.567	0.900	0.900	**0.933**	**0.933**
Overall	0.400	0.683	0.716	**0.725**	**0.725**

(b) State level

	MP	NP	MP·NP	PP	AP
Theater	0.632	0.816	**0.842**	0.803	**0.842**
Museum	0.725	0.900	0.900	**0.912**	**0.912**
Tower	0.743	0.843	0.843	0.843	**0.857**
University	0.744	0.919	0.930	**0.942**	**0.942**
Company	0.145	0.803	0.821	0.829	**0.855**
Sports Team	0.553	0.860	0.825	0.877	**0.904**
Politician	0.580	0.884	0.884	**0.913**	0.899
Overall	0.559	0.858	0.859	0.873	**0.887**

Table 3: Accuracy for geo-centre estimation in document corpora with and without the location refinement.

	City level		State level	
Category	AP	AP Refined	AP	AP Refined
Theater	0.745	0.745	0.842	0.842
Museum	0.603	**0.651**	0.912	0.912
Tower	0.512	**0.537**	0.857	0.857
Company	0.819	**0.833**	0.942	0.942
Sports Team	0.648	**0.667**	0.855	**0.872**
University	**0.830**	0.820	**0.904**	0.895
Politician	**0.933**	0.900	**0.899**	0.855
Overall	0.725	**0.736**	**0.887**	0.884

Table 4: Level classification results by different algorithms.

Method	Correct Geo-centers	Accuracy
TL	35	0.333
TM	69	0.657
FN	69	0.657
DN	71	0.676
PN	**75**	**0.708**

$$v_{TL} = \arg\max_v \sum_{i=1}^{n} TL(i, v). \quad (13)$$

- **Total Mentions (TM):** This algorithm is similar to TL, with the difference that all (and not just unique) mentions of a location in a page, denoted as $TM(i, v)$, are counted.

$$v_{TM} = \arg\max_v \sum_{i=1}^{n} TM(i, v). \quad (14)$$

- **Frequency Non-uniformity (FN):** This model prefers the level that has the largest non-uniformity distribution of the ranks over locations. For each level v and each document d_i, the non-uniformity of the inheritance-based probability distribution, denoted as $J^v(p_i, n)$, is computed, with L set to the set of locations at level v. The algorithm then finds the mean over documents for each level and chooses the level with the maximum mean.

$$v_{FN} = \arg\max_v \frac{1}{|P|} \sum_{i=1}^{|D|} J^v(d_i, n). \quad (15)$$

- **Distance Non-uniformity (DN):** This algorithm is similar to FN, with the difference that the non-uniformity function is $J^v(d_i, n)$, as given in Eq 6.
- **Probability Non-uniformity (PN):** This is the algorithm described in Section 4.5. The non-uniformity is computed as defined in Eq. 10. It differs from the methods FN and DN in that the score is based on the probability mass given by the combined model $\hat{P}$ instead of $\hat{P}_{inh}$ or $\hat{P}_{near}$.

Results and discussion Table 4 shows for each algorithm both the accuracy and the number of names whose centres are correctly identified. PN outperforms other baselines with an accuracy of 70.8%. Also similar models that measure the uniqueness of a geo-centre (namely FN and DN) are competitive, suggesting that methods that are based on the non-uniformity of the scores or ranks are effective in detecting a level of dispersion.

6. RELATED WORK

Our work relates to the lines of research on geotagging web resources and toponym resolution; we are not aware of any work on geotagging the more general class of named entities covered in this paper.

Existing approaches to extracting the geographic foci (geo-centres) of web pages [1, 19, 10] generally use hand-made rules to disambiguate location mentions or to aggregate the scores of multiple mentions. Web-a-where [1] comes up with a geographic focus for a web page by assigning a confidence score to every location mentioned in the page (using some predefined rules) and propagating the scores between locations that are in a containment relationship. Our work also exploits the containment relationships between locations, but unlike Web-a-where, we do not use hand-picked confidence scores.

Ding et al. [8] introduce the concepts of *power* and *spread* for a web page based on the link structure and the geographical scope of the page. The spread introduced in this work is similar to the entropy-based score used in our work for measuring the likeliness that a geo-centre exists. Our work differs from Ding et al.'s in that our approach does not rely on the link structure of pages, and can be applied to both individual pages and collections of pages.

When named entities are geographical locations, toponym resolution (e.g. [15, 7]) may be applied for disambiguation and to get the location coordinates. The problem can be considered as a binary classification where each possible interpretation of a toponym is considered correct or incorrect. The techniques developed here are mostly supervised, and may use features from a window of text where a toponym is mentioned. Our work also uses the context features but is not limited to geographical names.

Finally, there has been studies on locating a user based on his/her generated content, with the hypothesis that a user's location correlates with the content he/she generates in social networks. Most efforts fall in two of the world's most widely used social networks, Facebook [4] and Twitter [6], where known user geotags are used to train models that can predict the users' geo-centres. These approaches are supervised whereas named entities with tagged locations are not as prominent in web pages as they are in social network systems. An exception is GLITTER of Li et al. [14] which estimates the location of a microblog user by leveraging clues of points-of-interest (POI) that are mentioned by the user, and the cities that contain those POIs.Our approach differs from GLITTER in that the scores assigned to a location in our work is highly dependent on the positions of its mentions in each document, while GLITTER considers all the mentions of a location equally important; this may be a reasonable assumption for tweets with lengths limited to 140 characters but not for general web pages and documents.

7. CONCLUSIONS AND FUTURE WORK

In this paper we conduct a study on estimating the geo-centres of named entities based on their mentions in relevant web pages. We hypothesize that a name with regional orientation mentioned in a web page inherits the geo-centre of the page unless it is qualified with another geo-centre mentioned nearby. We propose an unsupervised framework to identify both a geo-centre and a level of dispersion. We devise a variety of models that estimate the probabilities of a unique geo-centre among a set of candidates and empirically evaluate our models and show that a good accuracy for all categories of names studied in this paper can be achieved.

Our study leads to a few potential research directions. First, while there are many entities with a unique geo-centre, there are some that may take multiple geo-centres over time. This may be less of an issue if the set of pages for which a geo-centre is detected focuses on one aspect that is likely to lead to a unique geo-centre. That said, studying the interactions between those centres and modelling their manifestations in documents or different categories of named entities is an interesting direction. Second, more clues may be leveraged when extracting candidate locations from web pages. In particular, the future work may consider extracting named entities with known geo-centres in web pages and using them as clues of locations. Also better capturing the structure of a web page can be helpful, and may lead to a more accurate model of measuring the distance between mentions of entities. Last but not least, our framework may be improved by targeting more specific classes of names and/or applying different models to different classes of names and pages. A deeper analysis of the characteristics of names, pages and evaluation models is needed toward such improvements.

Acknowledgments

This research is supported by the Natural Sciences and Engineering Research Council of Canada.

8. REFERENCES

[1] E. Amitay, N. Har'El, R. Sivan, and A. Soffer. Web-a-where: geotagging web content. In *Proc. of the SIGIR Conf.*, pages 273–280, 2004.

[2] N. Bach and S. Badaskar. A review of relation extraction. Technical report, Language Technologies Institute, Carnegie Mellon University, 2007.

[3] L. Backstrom, J. Kleinberg, R. Kumar, and J. Novak. Spatial variation in search engine queries. In *Proc. of the WWW Conf.*, pages 357–366, 2008.

[4] L. Backstrom, E. Sun, and C. Marlow. Find me if you can: improving geographical prediction with social and spatial proximity. In *Proc. of the WWW Conf.*, pages 61–70, 2010.

[5] M. Banko, M. Cafarella, S. Soderland, M. Broadhead, and O. Etzioni. Open information extraction from the web. In *Proc. of the IJCAI Conf.*, pages 2670–2676, 2007.

[6] Z. Cheng, J. Caverlee, and K. Lee. You are where you tweet: a content-based approach to geo-locating twitter users. In *Proc. of the CIKM Conf.*, pages 759–768, 2010.

[7] G. DeLozier, J. Baldridge, and L. London. Gazetteer-independent toponym resolution using geographic word profiles. In *Proc. of the AAAI Conf.*, pages 2382–2388, 2015.

[8] J. Ding, L. Gravano, and N. Shivakumar. Computing geographical scopes of web resources. In *Proc. of the VLDB Conf.*, pages 545–556, 2000.

[9] D. Downey, M. Broadhead, and O. Etzioni. Locating complex named entities in web text. In *Proc. of the IJCAI Conf.*, pages 2733–2739, 2007.

[10] R. P. et al. The design and implementation of SPIRIT: a spatially aware search engine for information retrieval on the internet. *Intl. Journal of GIS*, 21(7):717–745, 2007.

[11] J. Finkely, T. Grenager, and C. Manning. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proc. of the ACL Conf.*, pages 363–370, 2005.

[12] C. Kohlschütter, P. Fankhauser, and W. Nejdl. Boilerplate detection using shallow text features. In *Proc. of the WSDM Conf.*, pages 441–450, 2010.

[13] M. Lesk. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In *Proc. of the SIGDOC Conf.*, pages 24–26, 1986.

[14] G. Li, J. Hu, J. Feng, and K. Tan. Effective location identification from microblogs. In *Proc. of the ICDE Conf.*, pages 880–891, 2014.

[15] M. Lieberman and H. Samet. Adaptive context features for toponym resolution in streaming news. In *Proc. of the SIGIR Conf.*, pages 731–740, 2012.

[16] D. Ravichandran and E. Hovy. Learning surface text patterns for a question answering system. In *Proc. of the ACL Conf.*, pages 41–47, 2002.

[17] D. Smith and G. Crane. Disambiguating geographic names in a historical digital library. In *Proc. of the ECDL Conf.*, pages 127–136, 2001.

[18] T. Tezuka, T. Kurashima, and K. Tanaka. Toward tighter integration of web search with geographic information system. In *Proc. of the WWW Conf.*, pages 277–286, 2006.

[19] C. Wang, X. Xie, L. Wang, Y. Lu, and W. Ma. Detecting geographic locations from web resources. In *Proc. of the GIR Workshop*, pages 17–24, 2005.

[20] C. Whitelaw, A. Kehlenbeck, N. Petrovic, and L. Ungar. Web-scale named entity recognition. In *Proc. of the CIKM Conf.*, pages 123–132, 2008.