GEOTAGGING NAMED ENTITIES IN WEB PAGES

by

Jiangwei Yu

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Science

Department of Computing Science

University of Alberta

Abstract

We study the problem of geotagging named entities where the goal is to identify the most relevant location of a named entity based on the content of the Web pages where the entity is mentioned. We hypothesize the relationship between the mentions of an entity and its geo-center in web pages, and propose a framework that explores this hypothesis and provides a model that can give a ranked list of locations at different location granularities for an entity. We further study the problem of dispersion, and show that the dispersion of a name can be estimated and a geo-center can 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 large gazetteers.

We evaluate our methods under different settings and with different categories of named entities. Our evaluation reveals that the geo-center 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.

Acknowledgements

First of all, I would like to express my sincere thanks to my supervisor Dr. Davood Rafiei for his wise guidance and persistent support. I would also like to thank my colleagues Afsaneh Esteki and Muhammad Waqar for the discussions on the tools and resources involved in this work. Thanks also go to my friends for the great time we had together. Finally, I would like to thank my family for their love and support.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

During the past few years, there has been a growing interest in both extracting entities from the web [12, 27] and associating geographical descriptors to web resources [2, 5]. 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 that they are mentioned.

Many named entities have a geographic center or focus where the entity is better known or associated with; a person may be associated to a location where s/he was born, raised, worked, or known in general; an organization may be tagged with a location where it is headquartered in or has a large customer base, and a disease may be associated with a location where it is first reported, has a large spread, etc.

Exploring this relationship between a named entity and its geo-center(s) has important implications. Knowing the geography of named entities that appear in the matching pages of a query can help in disambiguating the entities, allowing, for example, a search engine to better localize the search [25] or diversify the results [23]. For example, when a user who frequently visits the website of Yale University posts a microblog containing “Shubert Theater”, it may be more accurate for an advertising system to supply the ticket information of the theater in New Haven, Connecticut, rather than the one in New York City, because it is very likely that a frequent visitor to the website of Yale University is a resident in New Haven. A real world example is that anyone who has filled one of the forms that read “I—of—declare”, can assert that location is an important and sometimes inseparable attribute of an entity and also the one that may uniquely identify an entity when combined with name.

Obtaining accurate and effective geographical information for named entities on the Web can be challenging for several reasons. *First*, 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, small companies and organizations, schools, sports teams, local theaters, etc.), some entities with a large spread or better known globally may not have a fixed geo-boundary; examples are well-known celebrities (e.g. Lady Gaga) and politicians (e.g. Barak Obama), multi-national organizations (e.g. ACM) and companies (e.g. McDonald, Google, Amazon), well-publicized products (e.g. Macbook Air), etc. Also entities can have multiple and sometimes non-overlapping geo-centers (e.g. places of birth and work). *Second*, both entity and location mentions in web pages can be ambiguous. For example, the term "Springfield" can refer to 63 places in 34 states in the USA [2] and names like Houston and Dallas can be both person names and place names. *Finally*, many different semantic relationships can exist between a named entity and a location; unlike some special cases where the relationship may be described using some templates (e.g. *headquarter* and *birth place* [24]), preparing an exhaustive set of patterns that can capture the semantics of all or even most of these relations is difficult (if not impossible).

Problem Formulation. In this thesis, we propose a framework for tackling 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-center can be at any or all of country, state or province, and city levels. We assume these relationships are described in a *gazetteer* (a geographical dictionary) and our experiments make use of one in the form of a tree with the root node representing the whole world, and the nodes in the second, the third and the fourth levels representing respectively countries, states and cities. The problem can be formulated as:

Given a named entity N and a set of its relevant pages $P_1, P_2, ..., P_n$ and a tree structure of locations (gazetteer), estimate N's geo-center, $G(N)$, where the named *entity has more significant visibility and familiarity than in any other location.*

Hence the problem of finding the geo-center is equivalent to finding a node in the location gazetteer.

We develop a probabilistic model that assigns geo-center(s) to a name based on the geo-center 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-centers and a set of pages that mention those names.

1.1 Thesis Statements

Our thesis statement is that the relationship between the mentions of an entity and its geo-center in web pages can be modelled by simple statistics. We assume that all mentions of the same name in a web page refer to the same entity hence would have the same geo-center.

1.2 Thesis Contributions

Our contributions can be summarized as follows.

- We hypothesize that when a name has some sort of regional orientation and is mentioned in a web page, it either inherits the geo-center of the page or is explicitly qualified with a different geo-center.

- We propose an unsupervised framework that estimates the likeliness of a location being the geo-center of a given named entity based on the mentions of the named entity and locations in web pages. Exploiting the hierarchical relationships between locations (provided by a large gazetteer), we analyze locations at the country, state and city levels.

- We devise an algorithm that estimates the dispersion of a name, and show that a geo-center can be detected at an exact dispersion level.

- We evaluated our work on real world datasets composed of various kinds of named entities and good performance has been shown by the results.

1.3 Thesis Organization

The remainder of this thesis is organized as follows. We review related work in Chapter 2 and give an overview of our framework in Chapter 3. Chapter 4 introduces the method for extracting location candidates from web pages and linking them with the geographic entities in the gazetteer. Probabilistic frameworks for estimating the geo-centers at specified dispersion levels are presented and evaluated in Chapter 5. In Chapter 6, we study the problem of estimating an exact dispersion level of the geo-center of a name. We discuss techniques for filtering names that do not have regional orientation in Chapter 7. We summarize the thesis and discuss future work in Chapter 8.

Chapter 2

Related Work

To our best knowledge, there is no prior work addressing the same problem studied in this paper. In this chapter, we review the geotagging of web documents and resources in social network systems. We also briefly discuss the connections between our problem and other related topics.

2.1 Geotagging Web Documents

Geotagging web documents is the most related work to our problem. Existing approaches to extracting the geographic foci (geo-centers) of web pages focus on building hand-made rules to disambiguate and aggregate the locations mentioned in the pages.

Web-a-where [2] is one of the most widely studied methods, which comes up with a geographic focus for a web page by assigning confidence scores to every location using some predefined rules and propagating scores between locations that are in a containment relationship. It generates a candidate list of geographic foci mixing locations regardless of their administrative levels, *i.e.*, there are locations at different levels in the rank-list of geographic foci for the web page. Our approach considers each administrative level (country, state and city level) separately at first and then determines the correct level. Our approach is based on a probabilistic model capturing the relation between the mentions of named entities and locations, rather than some hand-picked confidence scores that are used by the web-a-where method.

Ding et al. [11] introduce the idea of the *power* and *spread* of 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 of a geo-center to exist in a set of locations. Our work differs from Ding et al.'s in that our approach does not rely on the link structure of web pages, and can be applied to both individual pages and collections of pages.

Wang et al. [26] propose a system that computes three types of geographic attributes of a web page: 1) provider location (the physical location of the web resource owner); 2) content location (the geographic location that the content of the web resource is about); 3) serving location (the geographic scope that the web resource reaches). The provider location is determined by the extracted *address strings*. A binary classification model trained with features like URL, title, anchor text, page content, etc., is used to estimate whether extracted address strings are the provider location. To estimate the content and serving locations, the authors adopt methods similar to those proposed in [2, 11].

A comparison of different approaches for assigning geographic scopes to web documents is studied by Anastcio et al. [4]. More recently, methods inspired from the previous works have been applied to geotagging web pages in Russian [22] and Chinese [9].

2.2 Geotagging in Social Network Systems

With the growth of the mobile usage and social network systems, there have been studies on locating a user by analyzing the user generated content (UGC), based on 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 network, Facebook [6] and Twitter [10, 16, 19], where known user geotags are used to train models that can predict the users' geo-centers. In contrast, our approach is unsupervised for the reason that named entities with tagged locations are not as prominent as UGC in social network systems.

There are some attempts to develop unsupervised methods that address the sparseness of data that may be available for geotagged users, for example in a geographic region or in a social network system. Li et al. [18] propose an unsupervised approach, called GLITTER, to estimate the location of a microblog user in by leveraging clues of points-of-interest (POI) that are mentioned by the user, and the cities that contain those POIs, as provided by Factual[1]. Our approach is similar to GLITTER in that locations at the same level (country, state and city) are analyzed respectively and that an aggregation algorithm is used to find out the geo-center from locations at different levels. 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.

2.3 Other Related Topics

Related work also includes the literature on detecting the *locality* of a web page, *i.e.*, whether a web page has a geographical orientation (*e.g.*, topics related to a specific area) or contains general information for global readers (*e.g.*, electronics and health topics). Existing approaches [3, 15] train a binary classifier based on a set of pages in the Open Directory Project[2] and features extracted from pages, such as occurrences of place names and metrics based on them. The same problem arises in the context of named entities. Our work also tries to detect if a named entity has a regional geo-center or is of general interest, considering locations mentioned in pages. Our work differs in that it is an unsupervised method and in estimating the relevance of a name to a specific regional orientation based on the information entropy of the estimates for different locations to be the geo-center of the name.

In some of the existing works, *geotagging* is known as the process of interpreting words as geographic locations and providing the coordinate values of the locations. Lieberman et al. [20] proposes a spatial lexicon model that distinguishes between a *global lexicon* of locations known to a large scope of audiences, and a

[1] `http://factual.com/`

[2] `http://www.dmoz.org/`

local lexicon that is audience-specific. The local lexicons identified are then used for resolving ambiguity of mentions of locations in newswire articles, based on assumptions that an article author will develop a linguistic contextual framework for the audience (i.e., tell the audience where the story happens) and that the author is aware of the nature of the expected audience's spatial lexicon (i.e., know what an audience knows). Paradesi [21] has developed a system, TwitterTagger, that geotags tweets. Different from our definition, the geotagging process is not about associating tweets with their locations, but is to find out the references (mentions) to geographic locations in tweets and to annotate these references with their actual address.

In recent years, there has been a growing interest in relation extraction, which closely relates to the problem we are tackling in this thesis, under the context of relationships between a named entity and locations. Relation extraction studies the problem on extracting semantic relations between entities from textual content. Approaches that focus on predefined relations [1, 8, 13] (*e.g.*, acquisition between companies) require hand-crafted input seeds or patterns to find similar matches in a large collection of pages. More relation patterns are learned from a given set of matches, and similarly more matches can be found with more patterns. Our work differs from these systems in that no prior knowledge of textual patterns between a name and its geo-center is needed.

There has been more recent studies on open relation extraction [7] to address the problem of relation extraction when the number of relations is massive and thus requiring training data or hand-crafted patterns are unrealistic. These studies are similar to our work in their unsupervised nature and the ability to identify relations between named entities and locations (*e.g.*, place of birth). However, while relation extraction systems are more focused on relations between named entities that are mentioned closely, our work can detect the geo-center for a name even if the location is not mentioned nearby based on the hypothesis that a name may inherit the geo-center of the web page where it is mentioned.

Chapter 3

System Overview and Dataset Preparation

In this chapter, we give an overview of the proposed framework and its underlying assumptions. Our framework aims at geotagging various named entities in different categories, including person names, organization names and implicit physical locations. In this regard, we also present our gold standard dataset composed of names in different categories.

3.1 Hypothesis

Our framework is based on the hypothesis that *when a named entity is mentioned in a Web page, it inherits the geo-center of the page unless the name is explicitly qualified with a different geo-center.* We further assume that all mentions of the same name in a web page refer to the same entity hence would have the same geo-center.

To verify this hypothesis on a set of pages that have regional orientation, we randomly selected 30 headline articles from six newspaper websites in North America. We manually inspected each named entity tagged as PERSON or ORGANIZATION by the Stanford Named Entity Recognizer [14]. We excluded named entities that satisfy one of these two conditions: i) the named entity does not have a geo-center; ii) the geo-center of the named entity spans over more than one country. A named entity that satisfies one of these conditions does not have a regional orientation, and is not considered in our study. Table 3.1 lists the names excluded from our study

Table 3.1: Named entities without regional orientation.

Name	Reason
Lorna Morello	Character
Alex Vause	Character
NHL	North America
UN Security Council	Global Influence
United Nations	Global Influence
Red Cross	Global Influence
International Energy Agency	Global Influence
Jupiter	Scientific, Global Influence
Kepler	Scientific, Global Influence

Table 3.2: Different relations between named entities mentioned in Web pages and their geocenters.

Relation	Count	%
A only	51	31.5
B only	44	27.2
A and B	65	40.1
C	2	1.2

and the reasons for their exclusion. For the remaining 162 named entities that had geo-centers mentioned in the same pages, each name was assigned a geo-center. Each named entity, based on the relation between the named entity and its geo-center, is placed into one or more of the following three bins. Note that a named entity can be placed in both bins A and B.

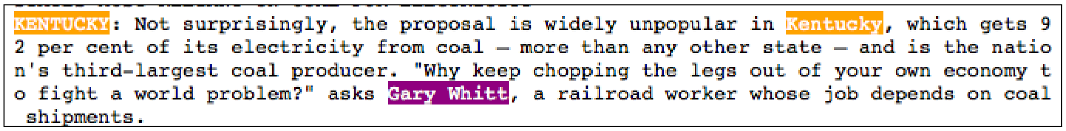
A. The named entity inherits the geo-center of the page.

B. Either the geo-center of the named entity is mentioned nearby, or there is a name with the same but known geo-center mentioned nearby. We interpret ”nearby” as appearing in the same sentence, *i.e.*, the clue for the geo-center co-occurs with the named entity in the same sentence in a web page.

C. The named entity has a geo-center, but neither A nor B holds.

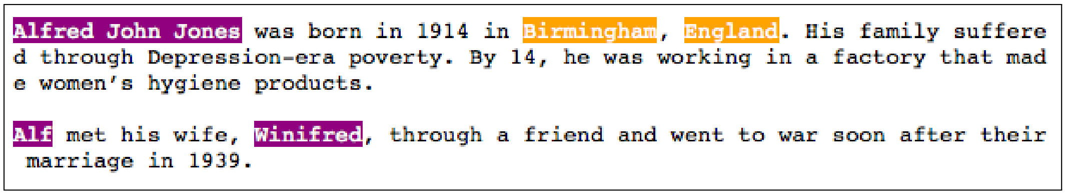
Table 3.2 shows the results of the study. We can see that in 71.6% (31.5% + 40.1%) of the cases the named entity inherits the geo-center of the page. We refer to this as our *inheritance hypothesis*. In 67.3% (27.2% + 40.1%) of the cases the geo-center or a name with the same but known geo-center is mentioned nearby. We refer to this as our *near-location hypothesis*.

Only 2 of 161 (1.2%) cases did not fall into one of the bins A and B. In one case, as shown in Figure 3.1, names in each paragraph have different geo-centers, which are mentioned at the beginning of the paragraph, but not mentioned in each sentence hence they are not placed in bin B. In another case, as shown in Figure 3.2, we know that *Winifred* is the wife of *Alfred John Jones (Alf)*, who was born in *Birmingham, England*. But we do not know the birthplace of Winifred. Neither A nor B applies to the named entity *Winifred*, and this is because: 1) the web page is telling a story of families from different places so we are not sure of the geo-center of the page; 2) Only by inferring the context of multiple sentences and with the knowledge that *Alf* is short for *Alfred*, we may conclude that the geo-center of *Winifred* is *Birmingham, England*. *Alf* itself is not a named entity with a known geo-center.



```
KENTUCKY: Not surprisingly, the proposal is widely unpopular in Kentucky, which gets 9
2 per cent of its electricity from coal – more than any other state – and is the natio
n's third-largest coal producer. "Why keep chopping the legs out of your own economy t
o fight a world problem?" asks Gary Whitt, a railroad worker whose job depends on coal
 shipments.
```

Figure 3.1: The hypothesis cannot cover the case for the named entity *Gary Whitt*, whose geo-center is *Kentucky, USA*.



```
Alfred John Jones was born in 1914 in Birmingham, England. His family suffere
d through Depression-era poverty. By 14, he was working in a factory that mad
e women's hygiene products.

Alf met his wife, Winifred, through a friend and went to war soon after their
 marriage in 1939.
```

Figure 3.2: The hypothesis cannot cover the case for the named entity *Winifred*.

The reported experiment confirms that our hypothesis holds for almost all the cases in this study. And with only about one percent of the cases falling outside A and B, we believe that our hypothesis is 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 that this hypothesis will hold to the same degree in general web pages, which is an area where more studies are needed. That said, in developing our framework, we assume the same or similar rules are followed by the editors or writers of non-newswire web pages.

3.2 The Framework

Our framework takes as input a named entity and a set of web pages and estimates the geo-center of the name based on its mentions in those pages.

If a page mentions both the named entity and a location, it is likely that there exists some relation between the name and the location. Therefore, our approach focuses on the extracted mentions of the target named entity and candidate locations and leverages these clues to build models for geo-center estimation.

Our framework consists of three stages.

1. In the first stage, 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[1] depending on its context. Such preprocessing addresses the problem of location ambiguity in web pages and is a common practice in the related works [2, 18].

2. In the second stage, the geotagging of the named entity is performed at the same level of cities, states and countries, based on the clues provided by the disambiguated locations mentioned in web pages.

3. In the final stage, the administrative levels of the geo-center are classified based on the results for each level in the second stage, and one of the city, state and country levels is selected as the exact administrative level for the geo-center of the named entity.

By tackling the problem in this way, we can improve the overall performance of the framework by optimizing the algorithm in each stage separately. We will illustrate the details of each stage in subsequent chapters.

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

3.3 Gold Standard Datasets

Since a full coverage of all kinds of named entities is not realistic, we focus on three sets of proper names: "persons", "locations" and "organizations", which are key types in common named entity recognizers.

The first set consists of *person names* with different granularities of geocenters. As politicians usually have a clear level of administration, we can obtain the geocenter of politicians with high confidence. We collected names of heads of states in the world[2] 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[3] and party leaders[4] of provinces and territories are categorized as names at the state/province level. Names of city mayors[5] and councilors[6] 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 many named entity recognition tools. This kind of names are very indicative about location information, especially with the common use of social networks on mobile phones, where user posts often contain these names. Understanding the city level geocenter of these names is important. To evaluate the performance of our geotagging framework on implicit physical locations, we collected names of museums, theaters and towers in the United States. It is logical to consider the located city as the geo-center of an implicit physical location, as these places may be well known in the same city but far less well known in other cities even in the same state or province. We collected names and longitude/latitude information from the Geonames database and the longitude/latitude pairs are converted into cities with the Google Map API[7].

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

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

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

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

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

[7] https://developers.google.com/maps/documentation/geocoding/, vis-

The third set is the names of *organizations*, including universities, sports teams and technology companies, collected as follows:

1. Names of universities and their locating cities were extracted from the list of Top 100 U.S. Universities by U.S. News[8];

2. 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;

3. From CrunchBase[9] 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-center of that organization.

For each collection of names, we collected related pages from December 2013 to March 2014 by searching the names in the search engine Exalead[10] 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.

ited on Dec 30, 2013

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

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

[10] http://www.exalead.com/search/web/

Chapter 4

Location Candidates Extraction

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. In this chapter, we discuss the steps for extracting location candidates from raw HTML sources.

4.1 Full Text Extraction

The source of a web page contains HTML tags and potentially scripts that are not exploited by our framework. We are interested in the textual content of the pages. Hence given the source of a web page, the first step of the analysis is extracting the text content. We use the *Keep Everything Extractor* in the boilerpipe library[1] developed by Kohlschütter et al. [17] to extract the full text of the page.

4.2 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 n terms (English words and punctuations),

$$W = \{w_1, w_2, \cdots, w_n\}. \tag{4.1}$$

A mention m is a subsequence of W described by the indices of its first and last terms, denoted by $s(m)$ and $t(m)$. Hence a mention m may be denoted as

[1] `https://code.google.com/p/boilerpipe/`

$w_{s(m)}, w_{s(m)+1}, \cdots, w_{t(m)}$.

In this work, we extract three kinds of mentions:

1. **Mentions of a target named entity.** We tokenize the target named entity, resulting in a sequence of k tokens $\omega_1, \omega_2, \cdots, \omega_k$. A mention m of the target named entity which starts at $s(m)$ and ends at $t(m)$ of the sequence W (inclusive) must satisfy

$$\omega_i = w_{s(m)+i-1} \tag{4.2}$$

for $1 \leq i \leq k$.

2. **Mentions of locations.** We consider consecutive terms in W tagged as LOCATION by the named entity recognizer as mentions of locations. We will explain the disambiguation of these mentions in Section 4.3.

3. **Mentions of adjectivals and demonyms of locations.** We also collect the adjectival forms of countries [2] and states in the U.S. [3]. 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 each adjectival or demonym is the same as the target named entity described above.

Note that we do not allow the mentions to overlap with each other. 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 to possibly underestimate (but not overestimate) the performance of our system.

4.3 Location Disambiguation

With location mentions extracted, an important question is: *Which geographic entity does each mention of a location refer to?*

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

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

Ambiguities often arise in resolving locations, and such ambiguities can be categorized into two types. The first is the GEO/non-GEO ambiguity, which characterizes the situation where a named entity can possibly be both a location or another kind of entity. The second kind of ambiguity is the GEO/GEO ambiguity. Mentions identified in the text may refer to more than one locations. For example, the term "`London`" without considering any context information can refer to "`London, England, United Kingdom`" or "`London, Ontario, Canada`". Without properly addressing this issue, we cannot provide valuable information for grounding the named entity.

4.3.1 Methods for Location Disambiguation

We tackle the problem of disambiguating location mentions as follows. 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 with canonical information in a tree structure to obtain a list of possible matches [4]. If no results are returned, we simply treat the mention as non-location word. Otherwise, we choose a geographic entity from the list and associate it with all the mentions in the page with the same surface text according to the one-sense-per-discourse principle.

We have experimented two disambiguation strategies. The first strategy is to select the most populated location from the list. For example, by searching the term "London" we may obtain two canonical locations: "London, England, United Kingdom" and "London, Ontario, Canada". Since the first one is more populated and thus is known by more people, we resolve it as the geographic entity the term "London" refers to.

The second strategy is based on the first one but also considers the mentions of other locations in the same page. The intuition is that people (authors and readers of web pages as well as other text media) can tell the canonical location of a mention by referring to mentions of locations of higher level of granularity in the

[4] To achieve that, we have modified an open source project built upon the Lucene index of the entries in the GeoNames database.

context. In the previous example, if the page also mentions “Canada” but not “England”, “British”, “UK”, “United Kingdom” or places such like, it is very likely that “London” here is referring to the city in the province of Ontario, Canada.

Not all location mentions may help with resolving a disambiguation. Instead of deciding which other mention or mentions should be considered for a resolution, we use the *term distance* between a mention to be resolved and the mention that helps the disambiguation. That is, a mention nearby should have more contribution on a resolution. A supporting example is the way English canonical addresses are mentioned, with place names of different levels written one followed by another. At the same time, there may be multiple mentions with the same terms in the same page. Given an unresolved location with text surface s (e.g., Edmonton) whose mentions are

$$Mentions(s) = \{m_1(s), m_2(s), \cdots, m_k(s)\}, \tag{4.3}$$

and a set of location candidates

$$L = \{l_1, l_2, \cdots, l_{|L|}\}, \tag{4.4}$$

a confidence score for that s actually refers to l_i is defined as

$$DS(l_i|s) = \sum_{c \in l_i} \sum_{\substack{m_j(c) \in Mentions(c) \\ m_j(c) \notin Mentions(s)}} \frac{1}{\min_{m(s) \in Mentions(s)} D_M(m(s), m_j(c))} \tag{4.5}$$

where c is a constituent term of l_i at city, state or country level. 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 $m_j(c)$ in the page. If $m_j(c)$ is not a mention of s, we treat $m_j(c)$ as a clue to resolve s. For example, *Alberta* is a clue to resolve *Edmonton*.

As the closer mentions are expected to contribute more to the confidence score, the contribution of $m_j(c)$ is calculated as the reciprocal of the minimum term distance between $m_j(c)$ and the mentions of s. Such term distance is formally defined as the number of terms between two mentions,

$$D_M(m_1, m_2) = \min\{|t(m_2) - s(m_1)|, |t(m_1) - s(m_2)|\}. \tag{4.6}$$

By the one-sense-per-discourse principle, we resolve all mentions of s ($m_i(s) \in Mentions(s)$) to the canonical location l with the largest $DS(l|s)$ for $l \in L$.

The problem of location disambiguation has been studied in many prior works. One of the most widely used strategies is proposed by Amitay et al. [2], 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. There are two commonalities between their method and strategies proposed here.

1. When it comes to ambiguity, all methods rely heavily on the default meaning of a term - the most populated location.

2. The clues that are used for resolving an ambiguous term both in Amitay et al.'s work and in our second method are the *co-mentions* of locations in the same page. Amitay et al. assign high confidence scores (only ranges are described) to a term that can be resolved by looking at its neighbouring text. For example, the term *Alberta* is resolved to *Alberta, Canada* with high confidence if *Canada* occurs right after some mention of *Alberta*. With an overall effect similar to Amitay et al.'s work, in our second method constituents of a location appeared nearby contribute high scores to the location. For example, an occurrence of *Alberta, Canada* contributes 1 for that *Alberta* resolves to *Alberta, Canada*.

We implemented our own methods because we are not sure how the weights could be assigned to different locations in Amitay et al.'s work.

4.3.2 Experiments and Evaluation

In this section, we evaluate the performance of our two location disambiguation strategies using the dataset introduced in Section 3.3.

We check if the strategies can correctly find the ground truth in web pages that are related to the given named entity. Specifically, for each page of a named entity,

Category	Total Pages	Pages with geo-centers	
		Population	Mention
Companies	1079	570	575
Theaters	974	755	728
Museums	843	666	670
Towers	497	314	314
Politicians	603	487	491
SportsTeams	1486	1280	1286
Universities	1002	825	848
Overall	6484	4897	4912

Table 4.1: Numbers of pages with geo-centers retrieved from web pages using location disambiguation strategies based on populations and mentions.

we apply both strategies and obtain a list of candidate locations. If the geo-center of the named entity is included in the list, we count it a hit for the corresponding page.

Table 4.1 shows the results of this experiment. We have listed results for each category and for all names. We can see that the strategy that considers all mentions of locations in a web page (MENTION) has a slight advantage over the strategy that picks the most populous location (POPULATION). For all categories except Theaters, the method MENTION has identified more geo-centers for the corresponding name than the method POPULATION. On the other hand, for the category of Theaters, the number of pages with geo-centers resolved by POPULATION is 2.8% greater than that by the strategy MENTION.

We can see that the strategies have close performance with a slight advantage for the strategy MENTION. Given that the method POPULATION would not be applicable for locations with less populations, we opt for the strategy MENTION in our later experiments.

Chapter 5

Geotagging at Specific Levels

Given a named entity, a collection of relevant pages with mentions of the named entity, a set of disambiguated locations in each page, and a desired level of dispersion (e.g., country, state, city), our task is to determine the most relevant location (geo-center) of the target named entity at the specified level of dispersion.

In this chapter, we propose models to rank locations at a given level for a target named entity. In the next chapter, we address the same problem but with no level of dispersion given.

The structure of this chapter is as follows. In Section 5.1 we introduce the models for finding the geo-center at the city level. Upon these models, an aggregation algorithm is proposed in Section 5.2 for locating the geo-center at the state and the country level. Leveraging the techniques for analyzing single pages, we combine the results from multiple pages for the same named entity in Section 5.3.

5.1 Geotagging at the City Level

Instead of geotagging at higher levels, we start with the city level for the reason that when considering geotagging at the state or the country levels, one should not ignore the city level locations because the mentions of such locations have strong indication of the higher locations they belong to. Moreover, locations at the city level are the most fine-grained locations in our gazetteer. That means the modelling process should begin at the city level anyway.

Our hypothesis, as given in Section 3.1, specifies two sources where the geo-center of a named entity can be identified.

- **Inheritance Hypothesis**. The named entity inherits the geo-center of the page.
- **Near-location Hypothesis**. Either the geo-center of the named entity is mentioned nearby, or there is a name with the same but known geo-center mentioned nearby. We interpret ”nearby” as appearing in the same sentence, *i.e.*, the clue for the geo-center co-occurs with the named entity in the same sentence in a web page.

Accordingly, we propose two models for geotagging a named entity, each tapping into one of these sources, before we combine the two.

5.1.1 Near Location based Model of Geo-Center

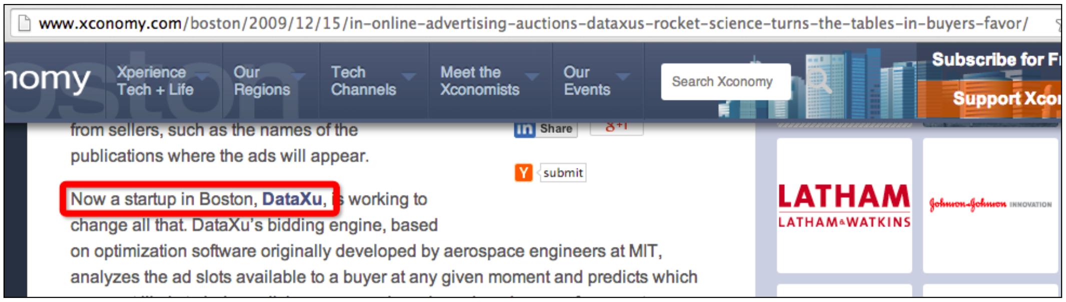
Our study, as reported in Section 3.1, showed that in 67.3% of the cases the geo-center is mentioned near the named entity in the same sentence for purposes such as introduction and disambiguating the named entity, following our Near-location Hypothesis.

We can turn our hypothesis to a model by rewriting the observation as *the less the term distance between mentions of the candidate location and the target named entity is, the more likely the candidate location will be the geo-center.*

Supporting examples span from newswire articles to information profile pages. In Figure 5.1, a technology company called *DataXu* is mentioned next to a mention of its headquartered city (Boston, MA) in a newswire article. In Figure 5.2, in a profile page of the *Krikorian Redlands Cinema* the address of the cinema is mentioned right below the name.

Based on this observation, we propose a model to estimate the geo-center at the city level. Given a set of locations at the city level, $L = \{l_1, l_2, \cdots, l_{|L|}\}$, the relevance of a location as a geo-center for a named entity N mentioned in page P is defined as

$$R_d(l|P,N) = R_d(l|L,N) = \frac{\frac{1}{D_E(l,N)}}{\sum_{l' \in L} \frac{1}{D_E(l',N)}}, \tag{5.1}$$

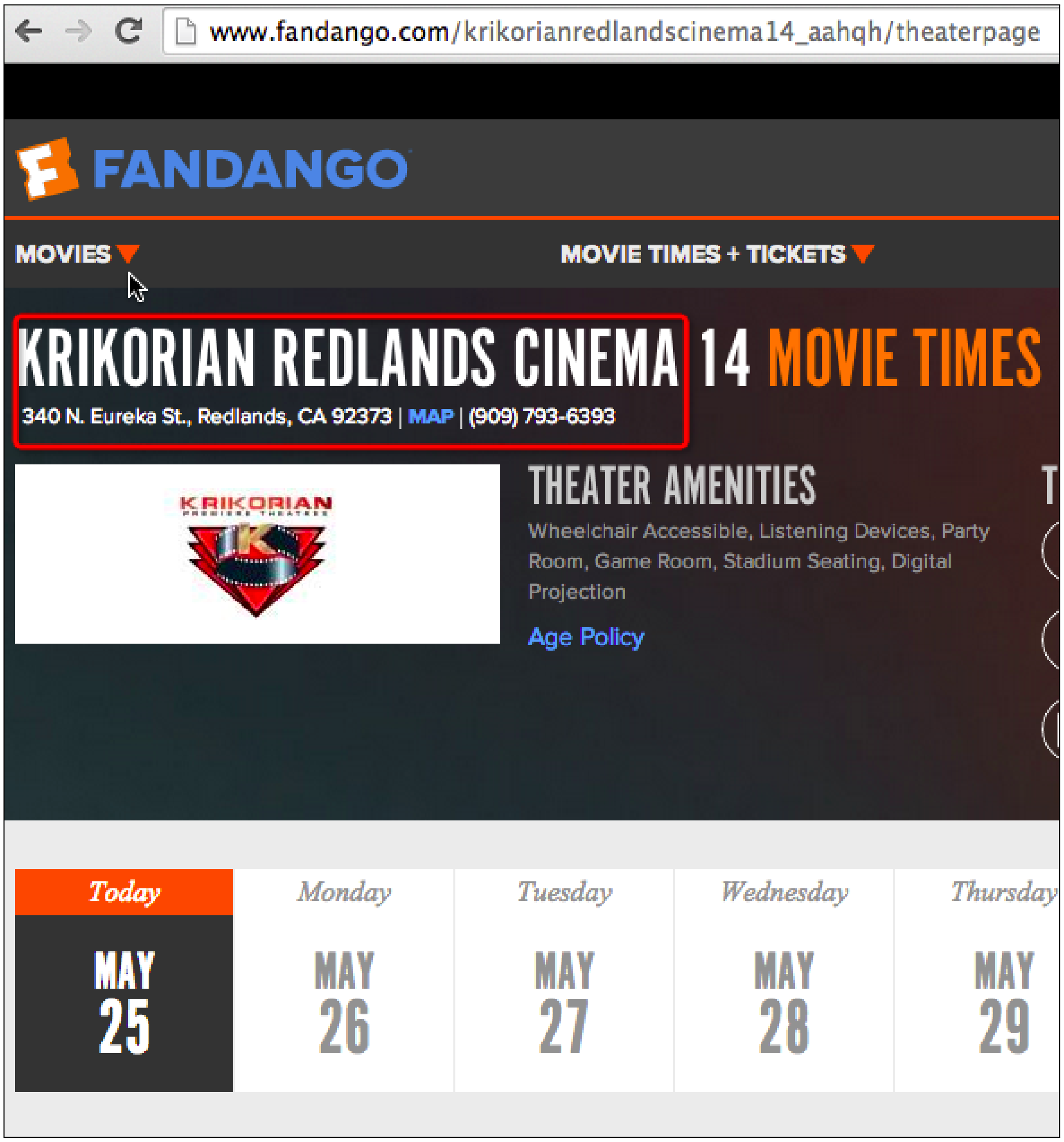


www.xconomy.com/boston/2009/12/15/in-online-advertising-auctions-dataxus-rocket-science-turns-the-tables-in-buyers-favor/

from sellers, such as the names of the publications where the ads will appear.

Now a startup in Boston, DataXu, is working to change all that. DataXu's bidding engine, based on optimization software originally developed by aerospace engineers at MIT, analyzes the ad slots available to a buyer at any given moment and predicts which

Figure 5.1: The geo-center is mentioned near the target named entity in a news article.



www.fandango.com/krikorianredlandscinema14_aahqh/theaterpage

KRIKORIAN REDLANDS CINEMA 14 MOVIE TIMES

340 N. Eureka St., Redlands, CA 92373 | MAP | (909) 793-6393

THEATER AMENITIES

Wheelchair Accessible, Listening Devices, Party Room, Game Room, Stadium Seating, Digital Projection

Figure 5.2: The geo-center is mentioned near the target named entity in a profile page.

where L is a set of locations mentioned in page P, $D_E(l, N)$ is the minimum term distance between mentions of l and N, referred to as *Entity Distance*, defined as

$$D_E(e_1, e_2) = \min_{\substack{m_1 \in Mentions(e_1),\\ m_2 \in Mentions(e_2)}} D_M(m_1, m_2), \tag{5.2}$$

where e_1 and e_2 are two named entities and $D_M(m_1, m_2)$ is the term distance between mentions m_1 and m_2, defined in Eq. 4.6.

The reciprocal of $D_E(l, N)$ correlates with our observation where locations mentioned nearby the named entity are more likely to be the geo-center. The scores are normalized to a number between 0 and 1, so they can be interpreted as probabilities.

5.1.2 Inheritance based Model of Geo-Center

According to the study in Section 3.1, in 71.6% of the cases a named entity inherits the geo-center of the web page where it is mentioned; we referred to this as our Inheritance Hypothesis. This observation suggests a method for estimating the geo-center of a named entity mentioned in a web page. A related problem now is estimating the geo-center of a page.

The problem of geotagging a web page has been studied in the past (*e.g.*, [2]) as reviewed in Chapter 2. Any of the methods reviewed in that chapter can be used here; however, we adopt a simpler model based on the following small study on the dataset reported in Section 3.1:

In a manual inspection of the 30 web pages in the dataset, we assigned a geo-center to each page based on 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-center of the page was the location where the newspaper was published. Among these 21 pages that we were sure of their geo-centers, we found that in 15 pages, the most frequently mentioned location was either the geo-center or part of the geo-center of the page.

This observation suggests that the more frequent a location is mentioned, the more likely it is the geo-center of the page. Based on this observation and our

Inheritance Hypothesis, we can estimate the geo-center of a named entity N mentioned in a web page P as

$$R_f(l|P,N) = R_f(l|L,N) = \frac{|Mentions(l)|}{\sum_{l' \in L} |Mentions(l')|}. \tag{5.3}$$

where l is a location from the set of locations L mentioned in P and $|Mentions(l)|$ is the number of mentions of l in the page.

5.1.3 The Combined Model

As discussed above, each model is only capturing one aspect of a geo-center. According to the study in Section 3.1, we should capture both aspects in order to cover most of the cases where the geo-center is identified.

For a page P, consider the case when the values of $R_d(l|P,N)$ are evenly distributed for every location l mentioned in P, which suggests locations have close probabilities to be the geo-center. In other words, it is very likely that the name does not have a unique geo-center. In this case, some of the questions that arise are: how much should we rely on these values when the name actually has a clear geo-center? Can we opt for another measure which can provide a larger margin between the probabilities?

To answer these questions, we use the Shannon Entropy of the vector induced by $R_d(l|L,N)$ for $l \in L$ to measure the probability that the name has exactly one geo-center. This probability is defined as $J(L|R_d,N)$ in Eq. 5.4, where the conditional notation indicates that the probability is based on our near-location model R_d. $H(L|R_d,N)$, as defined in Eq. 5.5, is the distance-based entropy of the probabilities distributed over the cities mentioned in the page, and $H_{max}(L)$ is the maximum entropy value for all locations, which is $\log |L|$, achieved when all the probabilities are equal to $\frac{1}{|L|}$.

$$J(L|R_d,N) = 1 - \frac{H(L|R_d,N)}{H_{max}(L)} = 1 - \frac{H(L|R_d,N)}{\log |L|} \tag{5.4}$$

$$H(L|R_d,N) = -\sum_{l \in L} R_d(l|L,N) \log R_d(l|L,N) \tag{5.5}$$

When the gap between the maximum probability and the second largest probability is large, $H(L|R_d, N)$ is small and $J(L|R_d, N)$ is large, indicating a strong tendency toward the model that is based on term distance for correctly capturing the geo-center. Conversely, $J(L|R_d, N)$ is close to zero when the entropy value approaches to its maximum, meaning that the top values of $R_d(l|L, N)$ are close and the model may not be effective in detecting a geo-center.

Given page P, the probability of l to be the geo-center at city level estimated by the combined model is defined as

$$R(l|P, N) = R(l|L, N) = J(L|R_d, N) \cdot R_d(l|L, N) + (1 - J(L|R_d, N)) \cdot R_f(l|L, N). \tag{5.6}$$

Eq. 5.6 is a mixture of two terms. The first term characterizes the joint probability of two events: 1) the near location based model provides a correct estimate with probability $J(L|R_d, N)$; 2) l is the geo-center, estimated by the near location based model with probability $R_d(l|L, N)$.

The second term of Eq. 5.6 characterizes the situation where the near location based model cannot capture a unique geo-center. According to our hypothesis in Section 3.1 we assume that the target named entity inherits the geo-center of the page, and the probability of l being the geo-center is estimated by the Inheritance based model.

It is noteworthy that we have experimented with other models as well, such as replacing $J(L|R_d, N)$ with a frequency based entropy, and we compare some of these variations in Section 5.4.

5.2 Geotagging at State and Country Levels

As we tackle the geotagging problem not only at the city level, but also at other levels for names that are more widely known, we try to generalize the above probabilistic models to the state and country levels.

To find out the geo-center of a named entity N at the state level, we consider both the mentions of cities and states (provinces) in the page. Assume that the set of cities mentioned in page P is L_{city}. Let $L_{state} = \{ls_1, ls_2, \cdots, ls_{|L_{state}|}\}$ be the state level locations mentioned in page P.

The aggregation of the scores at city and state levels are done in two steps. *First*, the mentions of both state-level and city-level locations are considered and the probability of each location is estimated based on the proposed models in Section 5.1. Specifically, the probability is computed over the union of city and state level locations as our domain of discourse. *Second*, for each state level location ls in page P, the probabilities of the cities that belong to ls are added to the original probability; this give a new probability of ls being the geo-center of N at the state level. If a state level location is not mentioned in the page, but its cities are mentioned, its new probability will be the sum of its cities' probabilities.

$$R(ls|P,N) = R(ls|L_{city} \cup L_{state}, N) + \sum_{l \text{ is a city in } ls} R(l|L_{city} \cup L_{state}, N) \tag{5.7}$$

The formal definition is given in Eq. 5.7. The aggregation is applicable to any of the probabilistic models presented in Section 5.1. One can verify that the new probabilities for all state-level locations sum up to 1.

We can easily generalize this method to country level locations by propagating the probability mass of states and cities to that of the countries they are located in. Let $L_{country} = \{lc_1, lc_2, \cdots, lc_{|L_{country}|}\}$ be the country level locations mentioned in page P. The probability of lc being the geo-center at the country level in page P is given by

$$\begin{aligned} R(lc|P,N) = {} & R(lc|L_{city} \cup L_{state} \cup L_{country}, N) \\ & + \sum_{ls \text{ is a state in } lc} R(ls|L_{city} \cup L_{state} \cup L_{country}, N) \\ & + \sum_{l \text{ is a city in } ls} R(l|L_{city} \cup L_{state} \cup L_{country}, N). \end{aligned} \tag{5.8}$$

5.3 Corpus Aggregation

With a collection of relevant pages of a named entity N, one may locate the geo-center of N by aggregating the results of geotagging at each page. Previously, we have devised models and algorithms for estimating the geo-center at a given level of granularity in single pages. Now we want to estimate the geo-center of N at a given level of granularity (city, state, or country) by aggregating the results from pages in a collection of relevant pages of N.

5.3.1 Model Intuition and Definition

We treat each page P_i ($1 \leq i \leq n$) relevant to named entity N as a test for the event that a candidate location is the geo-center of N. A location that is selected as the geo-center is expected to pass all the tests. The probability for a location l to pass the test of P_i is the probability for l to be the geo-center given P_i, namely $R(l|P_i, N)$ (See Eq. 5.1, Eq. 5.3, and Eq. 5.6). Thus the probability for a location l to pass the tests of all the pages is the product of $R(l|P_i, N)$ for all pages P_i in the corpus of N.

However, it is possible that some pages do not mention the location being considered. To avoid making the probability zero, let λ be the minimum value of the original probabilities for any location mentioned in any pages of N. Furthermore, to avoid a decreasing value of the probability when the number of pages increases, we compute the n-th root of the product mentioned above. Thus, the corpus-wise probability of l to be the geo-center of N is defined as

$$R(l|N) = \left(\prod_{i=1}^{n} \max\{R(l|P_i, N), \lambda\} \right)^{1/n}, \tag{5.9}$$

where n is the number of pages in the corpus of N, and $\lambda = \min\{R(l'|P_j, N)\}$ for any l' whose level is the same as that of l and any page P_j ($1 \leq j \leq n$) such that $R(l'|P_j, N) > 0$.

For each level, we sort the locations at that level according to $R(l|N)$. The geo-center at the corresponding level is the location with the maximum value of $R(l|N)$.

5.3.2 Implementation Glitch

Technically, the production of float numbers in Eq. 5.9 is error-prone. To reduce the error introduced by the multiplication of float numbers, we calculate Eq. 5.9 with the transformation in Eq. 5.10.

$$R(l|N) = \exp\left[\frac{1}{n}\sum_{i=1}^{n}\log(\max\{R(l|P_i, N), \lambda\})\right] = \exp\left[\frac{1}{n}S(l|N)\right] \tag{5.10}$$

Instead of the production, Eq. 5.10 is computed with addition. In this way, we can actually use the result of the summation of logarithm values, denoted by $S(l|N)$, to sort different locations. The location with the maximum value of $S(l|N)$ is selected to be the geo-center at the given level.

5.3.3 Location Refinement

We refine the model with the following heuristics. 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-center. The underlying assumption of this refinement is that the probability of a location to be the geo-center at its level of dispersion can be affected by mentions of its parent locations in the tree structure.

We assume that a city ct is included in the state level location st, which is in the country level location cn. The refined model computes the probabilities for st and similarly ct being the geo-centers at the state and city levels as shown in Eq. 5.11 and Eq. 5.12 respectively.

$$Refine(st|N) = R(cn|N) \cdot R(st|N) \tag{5.11}$$

$$Refine(ct|N) = Refine(st|N) \cdot R(ct|N) \tag{5.12}$$

Note that we have $Refine(cn|N) = R(cn|N)$ as the country level locations cannot be refined from the root of the tree structure.

When implementing this extension, we can simply update the summation part in Eq. 5.10 to include the terms for the priori of the enclosing locations so that the multiplication of float numbers is avoided.

5.4 Experiments and Evaluation

We run experiments in two stages to evaluate the performance of the proposed approach to finding the geotag of named entities at a specific level. In the first stage, each time a named entity and a relevant page are given, we find the geo-centers of the named entity at city and state levels based on the content of the page. In the second stage, we evaluate the proposed approach on a corpus where page results are aggregated.

5.4.1 Evaluation Settings

From the data we collected in Section 3.3, we build evaluation datasets according to different tasks.

For geotagging at the city level, we collect single pages or document corpora that mention the city name of any ground truth location. For geotagging at the state or country level, we choose pages or corpora that contain mentions of the ground truth location at that level or its lower levels. Specifically, we say a corpus contains the ground truth location if and only if there exists a page in the corpus that contains the ground truth location.

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-center estimated by our algorithm matches one of the ground truth locations of the target named entity, we treat the answer correct. In this way we can compute the accuracy by dividing the number T_c of correct answers by the total number T of data points in the dataset:

$$Accuracy = \frac{T_c}{T}. \tag{5.13}$$

Geotagging with single pages. The task of geotagging single pages is for evaluating the performance of the proposed models to estimate probabilities of a location being the geo-center at the given level. Hence if the geo-center is dropped in the pre-processing steps (see Chapter 4) or there is only one candidate, the page is excluded because the result cannot be changed regardless of which model is used.

Geotagging with page corpora. In the task of geotagging with the entire corpus of a named entity, we focus on evaluating the ability of our algorithm to aggregate results of different pages. Thus we drop a corpus if it contains less than 5 pages simply because we may not get enough information from the corpus.

5.4.2 City level geotagging with single pages

In this experiment, we evaluate the performance of the models proposed earlier as well as other baselines as explained below. The uppercase letters in the parentheses are the short names used in Table 5.1 to refer to the models themselves.

- **Random (RAND):** The random model assigns equal probabilities to all the candidate locations of being the geo-center.

- **Frequency (FREQ):** This is the model defined in Eq. 5.3.

- **Term Distance (TD):** This is the model defined in Eq. 5.1.

- **Mixing by Multiplication (MM):** It is defined as the product of the values of Eq. 5.3 and Eq. 5.1. This model assumes the term distance and the frequency as the two factors that determine the probability to be the geo-center.

- **Mixing by Addition (MA):** This model considers the two events (1) the geo-center appears near the target named entity and (2) the named entity inherits the geo-center of the page, and it assumes the two probabilities are roughly the same. The mixture model is defined as the mean of the two probabilities, i.e. the sum of the values of Eq. 5.3 and Eq. 5.1 divided by two.

- **Mixing based on the frequency entropy (MFE):** It is similar to the definition in Eq. 5.6, with the difference that the disorderedness of probabilities is computed based on frequency instead of term distance. That is

$$R(l|L,N) = J(L|R_f,N) \cdot R_f(l|L,N) + (1 - J(L|R_f,N)) \cdot R_d(l|L,N). \tag{5.14}$$

  where

$$J(L|R_f,N) = 1 - \frac{H(L|R_f,N)}{\log|L|}, \tag{5.15}$$

$$H(L|R_f,N) = \sum_{l \in L} R_f(l|L,N) \log R_f(l|L,N), \tag{5.16}$$

  and L is the city level locations in page P.

- **Mixing based on distance entropy (MDE):** This is our proposed model in Eq. 5.6.

For each category of names, we compare the accuracy of different models. The results are shown in Table 5.1. The scores in bold in each column indicate the best accuracy achieved among all the models for the corresponding category.

	Theater	Museum	Tower	Uni-versity	Company	Sports Team	Politician
RAND	0.220	0.228	0.233	0.235	0.342	0.166	0.273
FREQ	0.575	0.661	0.608	0.606	0.356	**0.591**	0.833
TD	0.756	0.691	0.608	0.678	**0.641**	0.376	0.840
MA	**0.773**	0.748	0.670	0.716	0.613	0.528	0.887
MM	**0.773**	0.751	0.688	**0.733**	0.605	0.536	0.882
MFE	0.756	0.721	0.636	0.695	0.639	0.438	0.863
MDE	0.758	**0.759**	**0.727**	0.712	0.605	0.560	**0.891**

Table 5.1: Accuracy of city-level geo-center estimation in single pages by different models.

The results show that the model only based on frequency (`FREQ`) achieves the best performance on Sports Teams with an accuracy of 0.591. 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-centers mentioned in the page. Because of this setting, when estimating the geo-center of a team, it may be more promising to look for clues from the page geo-center than finding locations nearby, as the home location is less likely to be mentioned again closely.

However, the model `FREQ` does not perform well on other categories especially on the category of technology company names, with an accuracy of only 0.356. We analyzed the pages and found 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, which makes it less possible for the main topic of the page has strong location indication. Instead, the geo-center 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 (0.641) is the best among those investigated.

As we can see, models only based on the feature of 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 (theaters, museums, towers, and universities [1]) 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 accuracy in the categories of museums, towers and politicians. In the 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 0.639 in company names but falls short in sports teams (0.438) and towers (0.636) when compared with the other three mixed models.

In summary, we find that the relation between mentions of named entities and geo-centers should be modelled with term distance and frequency simultaneously so that a measure may play a more important role when the other measure cannot estimate the geo-center with a good confidence. The experimental results show that the proposed model MDE defined by Eq. 5.6 is more reliable than the others. Hence we will use it in later stages of our experiments.

5.4.3 State level geotagging with single pages

We experimented with the proposed algorithm as well as a series of baselines for state level geotagging with single pages. All algorithms leverage the probabilistic model MDE selected in Section 5.4.2 to assign initial probabilities to the candidate locations, with different aggregation approaches explained below.

- **States only (S):** This algorithm considers the state names mentioned in a page as candidate locations. The probabilities assigned by the model MDE are used for ranking the candidate locations.

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

- **Cities only (C):** This algorithm first computes the probabilities of cities in a page with the model MDE. Then the probabilities of cities in the same states are added up. 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 described in Section 5.2.

	Theater	Museum	Tower	Uni- versity	Company	Sports Team	Politician
S	0.617	0.848	0.801	0.664	0.405	0.361	0.641
C	0.740	0.688	0.667	0.643	0.619	0.588	0.725
MSC	0.809	0.877	0.801	0.707	0.618	0.472	0.813
ASC	0.817	**0.886**	0.813	0.730	0.616	0.493	0.821
AMS	**0.858**	0.853	**0.838**	**0.759**	**0.690**	**0.600**	**0.861**

Table 5.2: Accuracy of state-level geo-center estimation in single pages by different aggregation methods.

Table 5.2 reports the accuracy of applying each algorithm to different categories. We can see that the algorithm only considering either states (S) or cities (C) does not achieve the best result in any category when compared to the other three (MSC, ASC and AMS). It is noteworthy that the algorithm that only considers mentions of cities performs better than the one that only analyzes mentions of states in categories of theaters, companies, sports teams and politicians, which suggests that an algorithm for geotagging at the state level should not ignore the mentions of their cities.

The above results show that our proposed algorithm has a robust performance in all categories. It can find out the state level geo-center in a given page with accuracy over 0.83 in the categories of landmarks and politicians and achieves at least 60%

accuracy in the organizations category. The baselines may perform well on some categories but fall much behind in others.

We can also see that our proposed algorithm (AMS) has a higher accuracy than MSC and ASC, which combine the results given by methods S and C in all categories only except the category of museums. This advantage of ASC is due to the fact that when the number of mentions at different levels are not balanced, the algorithms MSC and ASC may overestimate the probability for a location whose level has few candidates, while the proposed algorithm AMS keeps such difference by distributing original probabilities to both levels simultaneously.

As suggested by the experimental results, we will use AMS as our algorithm of choice in the rest of our experiments.

5.4.4 Aggregating results of different pages

For each named entity and its set of relevant pages, we run experiments using different algorithms to aggregate the scores from individual pages. In addition to the algorithm proposed in Section 5.3, we implement a few baselines for comparison. The methods are explained below.

- **Maximum probability of the location in all pages (MP):** For each candidate location, the algorithm takes the maximum probability in all pages and the candidates are ranked by these values.

- **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 descending order by these frequencies.

- **Product of MP and NP (MP·NP):** This algorithm ranks the candidate locations by the product of the values in MP and NP. The idea here is that the geocenter should be mentioned in many pages and we expect in some pages it is tagged with a high probability by our algorithm for page level geotagging.

- **Average of probabilities of the location in all pages (AP):** This algorithm adds up the probability of a location to be the geo-center at each page, and divides the sum by the total number of pages. The intuition behind this algorithm is that the geo-center should have a high average probability in the corpus.

- **Product of probabilities of the location in all pages (PP):** This algorithm ranks the locations by the probability given by Eq. 5.9.

	Theater	Museum	Tower	University	Company	Sports Team	Politician
MP	0.556	0.463	0.412	0.431	0.250	0.539	0.542
NP	0.733	0.707	0.618	0.824	**0.672**	0.831	0.958
MP·NP	**0.756**	**0.756**	0.618	0.824	0.656	0.865	0.958
SP	0.711	0.732	0.618	**0.843**	0.656	0.854	0.958
PP	0.733	**0.756**	**0.676**	**0.843**	**0.672**	**0.910**	**0.958**

Table 5.3: Accuracy for city-level geo-center estimation in document corpora by different aggregation methods.

	Theater	Museum	Tower	University	Company	Sports Team	Politician
MP	0.716	0.662	0.706	0.610	0.174	0.532	0.538
NP	**0.896**	**0.818**	**0.824**	0.829	0.779	0.856	0.862
MP·NP	0.866	0.792	0.809	0.817	0.767	0.793	0.846
SP	0.881	0.805	0.809	0.829	0.721	0.793	**0.908**
PP	0.881	**0.818**	0.794	**0.841**	**0.814**	**0.874**	0.892

Table 5.4: Accuracy for state-level geo-center estimation in document corpora by different aggregation methods.

The results for the city-level and the state-level geo-center estimation in document corpora are shown respectively in Table 5.3 and Table 5.4.

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. If in this page the probability of a location is overestimated it will affect the final decision. But in other algorithms this problem is alleviated by using metrics that capture characteristics in all pages.

From Table 5.3 and Table 5.4 we can also see that our proposed algorithm PP has competitive results in all categories for both state and city levels. In addition, PP achieves the best recall in the category of sports teams for both levels. From the previous experiments, we have found that the categories of company names and sports team names are the most difficult ones. Since PP performs the best in these two categories and competitive in others, PP is selected as the algorithm for aggregating results of different pages in our framework.

5.4.5 Location Refinement

In Section 5.3.3, we have proposed a refinement method for adjusting the probability of a location by multiplying the probability of its parent location in the tree structure. Here we compare the accuracy with and without such refinement. As the refinement is done to the probabilities analyzed from the whole corpus of pages, we choose the corpus aggregation method PP selected in the previous experiment to be the baseline for comparison. Table 5.5 compares the accuracy with and without the refinement in different categories.

	City level		State level	
Category	PP	PP Refined	PP	PP Refined
Theaters	0.733	0.700	0.881	0.833
Museums	0.756	0.780	0.818	0.844
Towers	0.676	0.735	0.794	0.824
Companies	0.672	0.766	0.814	0.872
Sports Teams	0.910	0.865	0.874	0.883
Universities	0.843	0.843	0.841	0.890
Politicians	0.958	0.917	0.892	0.877
Overall	0.796	0.802	0.845	0.862

Table 5.5: Accuracy for geo-center estimation in document corpora by with and without the location refinement.

We can see that the refinement method slightly improved the accuracy for both levels. At the city level the overall accuracy improved from 0.796 to 0.802 with the refinement implemented, and at the state level similar improvement is recorded from 0.845 to 0.862.

Another observation is that the improvement for the state level is more promising than that at the city level. This is somewhat expected as the refinement from the country level to the state level is more reliable than that from the state level to the city level, given that it is more accurate to estimate the geo-center at the country level than that at the state level.

Chapter 6

Geotagging at Arbitrary Levels

6.1 Classification on Levels of Dispersion

With the probabilities of locations as the geo-centers at different levels of the location hierarchy, an interesting question is what is the most suitable level of dispersion for the geo-center of a named entity N among the three levels: city, state, and country.

Inspired by the idea that the desired level should be the one where the geo-center is more likely to be unique, we exploit the entropy-based models proposed in Section 5.1 for estimating the probability (confidence) that a set of locations contain a unique geo-center. We hypothesize that the exact level of dispersion should be the level that is most likely to have a unique geo-center. Hence, our approach is to find the level of dispersion with the maximum confidence defined as

$$C(v|N) = \frac{1}{n}\sum_{i=1}^{n} J(L_v(P_i)|R, N), \tag{6.1}$$

where v can be one of the three levels: city, state, and country, n is the total number of pages related to N and $L_v(P_i)$ are the locations at level v mentioned in page P_i.

As captured in Eq. 6.1, the probability for a level to be the exact dispersion is the mean of the confidence score of the set of locations at the corresponding level. The confidence score $J(\cdot|R, N)$, as defined in Eq. 5.4, measures the probability that a unique geo-center can be detected from the given set of locations based on the probability mass given by the combined model R defined in Eq. 5.6.

We choose the level with the maximum value to be the estimated level of dispersion of the geo-center. Finally, we can assign the name with a geo-center at the selected level based on the results of geotagging at the specific level.

6.2 Experiments and Evaluation

6.2.1 Evaluation Settings

In this task, we use only politician names for the evaluation for the reason that a politician has a clear dispersion in terms of their serving regions. On the other hand, organization names may have plausible reasons to be categorized into different levels according to their popularity (e.g., Edmonton Oilers is an NHL team based in Edmonton, Alberta, Canada but it is known by NHL fans in North America.) and landmarks are mostly known at the city level.

We built a dataset with names with different dispersions. Besides the names of politicians at the city level and the province level (equivalent to the state level) of Canada as described in Section 3.3, we added a list of names of *heads of states* in the most populated countries. In this way, we have names with their geo-centers at each level of locations. Similar to the previous experiments for aggregating results of pages in a corpus, we removed names with less than 5 relevant pages. Table 6.1 shows the statistics of our dataset for this experiment.

	Names
City Level	44
State Level	30
Country Level	32
Total	106

Table 6.1: The number of names with geo-centers at each level of dispersion.

Let the total number of names be T and the number of names for which our algorithm correctly identifies the geo-center at the exact level of dispersion be T_c. We use the metric accuracy as defined in Eq. 5.13 for this evaluation. The accuracy is the ratio of T_c to T.

6.2.2 Methods

We experiment with the algorithm given for determining the level of dispersion in Section 6.1 as well as a few baselines. The details are illustrated below. Note that we have level $v \in \{city, state, country\}$ and locations at level v and page P_i denoted as $L_v(P_i)$ unless it is stated otherwise.

- **Total Locations (TL):** For a page P_i, let the number of distinct locations at level v be $TL(i, v)$. This algorithm adds up the values of TL for all pages at each level and the level with the most locations is chosen. It is formulated as

$$v_{TL} = \arg\max_v \sum_{i=1}^{n} TL(i, v). \tag{6.2}$$

- **Total Mentions (TM):** This algorithm is similar to TL, with the difference that for each page the total number of mentions of locations at the given level, denoted as $TM(i, v)$, is used for counting. The level with the most mentions of locations is chosen.

$$v_{TM} = \arg\max_v \sum_{i=1}^{n} TM(i, v). \tag{6.3}$$

- **Frequency Disorderedness (FD):** In this algorithm, for each level v and each page P_i, the disorderedness of the frequency-based probability distribution over the locations at level v and page P_i, denoted as $J(L_v(P_i)|R_f, N)$, is computed as per the definition in Eq. 5.15. The algorithm then adds up $J(L_v(P_i)|R_f, N)$ for each page and chooses the level with the maximum sum.

$$v_{FD} = \arg\max_v \frac{1}{n} \sum_{i=1}^{n} J(L_v(P_i)|R_f, N). \tag{6.4}$$

- **Distance Disorderedness (DD):** This algorithm is similar to `FD`, with the difference that disorderedness score is replaced by $J(L_v(P_i)|R_d, N)$ as computed in Eq 5.4. It is formulated as

$$v_{DD} = \arg\max_v \frac{1}{n} \sum_{i=1}^{n} J(L_v(P_i)|R_d, N). \tag{6.5}$$

- **Probability Disorderedness (PD):** This is the algorithm described in Section 6.1. The disorderedness is computed as defined in Eq. 6.1. It differs from the methods FD and DD in that the score is based on the probability mass given by the combined model R instead of R_f or R_d.

6.2.3 Results and Discussion

Method	Correct Geo-centers	Accuracy
TL	34	0.320
TM	70	0.660
FD	71	0.670
DD	73	0.689
PD	**75**	**0.708**

Table 6.2: Level classification results by different algorithms.

The results are shown in Table 6.2. The column “Correct Goe-centers” shows the number of names whose geo-centers are correctly identified by each algorithm. The column “Accuracy” lists the accuracy of each algorithm.

We can see that the proposed algorithm PD performs better than any baseline with an accuracy of 70.8%. Another observation is that similar models that measure the probability to have unique geo-centers (FD and DD) have competitive performance, suggesting that choosing the level with a unique geo-center is an effective way for classifying levels dispersion.

Chapter 7

Regional Orientation

Recall that the proposed models assume that the named entities and the web pages used for geotagging should have certain regional orientation. In this chapter we suggest a way to filter named entities that do not have regional orientation.

7.1 Classification on Regional Orientation

Given a named entity and a set of pages, we try to answer the following question:

Based on the set of pages, can we determine the geo-center of the named entity with high confidence?

If the answer is positive, it implies that there is certain regional orientation with the name and we should be able to identify a geo-center for the name using the geotagging algorithm proposed. If the answer is negative, we can reject the name and not return a geo-center as the name may not have any regional orientation.

Again, we use the entropy-based confidence score to measure the probability that a unique geo-center can be identified from a set of locations. Here we compute the value of $J(L|R, N)$ where $L = \{l_1, l_2, \cdots, l_{|L|}\}$ is a set of country-level locations for a named entity N.

We come up with a threshold α for the value of $J(L|R, N)$. If and only if $J(L|R, N) \geq \alpha$, we consider the name N to have regional orientation.

7.2 Experiments

We collected names of endangered animal species on land as well as their living areas around the world[1]. As in Section 3.3, we query each name in the search engine Exalead to get its relevant web pages.

Based on the number of countries of their living areas, we divided the names of animals into two groups. The first group contains names that have exactly one country as their only geo-center. Examples are endangered animals like Giant Panda in China and Bumblebee Bat in Thailand. The other group consists of names that have more than one country as their living areas listed in the collected data. Chimpanzee living in Africa is an example of the second group, as Africa is a continent and contains more than one country.

We obtained 39 names for the first group and 51 names for the second group. After running the geotagging algorithm against each name at the country level, for each name N we obtained a list of countries $L = \{l_1, l_2, \cdots, l_{|L|}\}$. We calculate $J(L|R,N)$ as defined in Section 5.4. Names that cannot be tagged due to the lack of enough Web pages (at least 5) or names whose results did not contain any locations are removed. Thus, we obtained 33 names in the first group (noted as UNIQUE below) and 30 names in the second (noted as MULTIPLE below).

To determine a threshold α for predicting the group a name is in, we also collected a few names of animals that are more common than the aforementioned ones. For each name N, we calculate $J(L|R,N)$. The results are shown in Table 7.1.

N	$J(L\|R,N)$
Deer	0.127
Bat	0.066
Otter	0.064
Monkey	0.029
Cat	0.009

Table 7.1: Probabilities to have a unique country level geo-center for common animal species.

We can see that there are four names with $J(L|R,N)$ less than 0.1 and one name

[1] `http://www.earthsendangered.com/`, visited on March 11, 2014

with a probability slightly greater than 0.1. So we empirically set the threshold $\alpha = 0.1$.

We apply this threshold to classify the two groups UNIQUE and MULTIPLE. The results are shown in Table 7.2.

Table 7.2: Classification results for endangered animals to have a unique country level geo- center.

	Prediction Correct	Prediction Incorrect
UNIQUE	28	5
MULTIPLE	17	13

We calculate the metrics commonly used for classification problems: Precision, Recall and F1-Score. Precision ($prec$) is the number of correct predictions divided by the total number of predications. Recall ($recl$) is the number of correct prediction of UNIQUE. F1-Score is the Harmonic mean of $prec$ and $recl$, as shown in Eq. 7.1.

$$f_1 = \frac{2 \cdot prec \cdot recl}{prec + recl} \tag{7.1}$$

From the results in Table 7.2, we have

$$\begin{aligned} prec &= (28 + 17)/(28 + 5 + 17 + 13) = 0.714 \\ recl &= 28/(28 + 5) = 0.848 \\ f_1 &= 0.776 \end{aligned}$$

We can see that our algorithm can detect whether a named entity has a unique geo-center or it has more than one country-level geo-center and does not carry much geographic orientation. We give a brief error analysis for the results below.

From Table 7.2, we can see that our algorithm made 5 incorrect predictions for names with a unique geo-center, classifying them as MULTIPLE. We found that, despite being listed under one country name in our ground truth, three species among them actually appear in more than one country: American Bison, Short-tailed Chinchilla and Mongolian Beaver. For the other two names, our algorithm made mistakes because there are lists of different animals and locations in the pages where these names are mentioned. Our measure suffers from the cases where two named entities and their geo-centers are closely mentioned.

Our algorithm incorrectly predicted 13 names with multiple geo-centers as UNIQUE. For three of them, Chimpanzee, Leopard and Jaguarundi, the errors are due to that we collected many pages whose geo-center is the United States, resulting in a bias towards the United States. For each of another 7 names, our algorithm correctly detected at least one of its geo-center but due to the sparseness of the data, there were not enough clues for other geo-centers to get a competitive score, resulting in a bias towards only one location. For the remaining 3 names, our algorithm detected incorrect geo-centers, where one of the failures was due to the error of location disambiguation.

Chapter 8

Conclusions and Future Work

In this thesis we conduct a study on estimating the geo-centers 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-center of the page unless it is qualified with another geo-center mentioned nearby. We propose an unsupervised framework based on our hypothesis to identify the geo-center when the target level of dispersion is either known or unknown. We devise a variety of models that estimate probabilities of the existence of a unique geo-center among a set of candidates. Experiments showed that the proposed probabilistic model and the corpus aggregation algorithm can achieve a good accuracy for all categories of names studied in this thesis. We also showed that we can detect the exact level of the geo-center for names with different dispersion.

There are a few potential directions for improving this framework in future.

First, more named entities of different categories can be studied. For example, in this work we only studied politicians for person names while names of more professions (e.g., lawyers, professors, etc) can be studied.

Second, more clues can be leveraged when extracting candidate locations from web pages. We have been using a free gazetteer GeoNames which mainly consists of geographic entities. The future work may consider extracting named entities with known geo-centers in web pages and using them as clues of locations. Names with known geo-centers can possibly be obtained from free databases like Factual[1] and

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

Wikipedia [2].

Third, capturing the structure of a web page may be helpful. Currently we extract text from HTML sources by simply keeping everything and aligning them sequentially regardless of the different sections of a page. In fact, correctly separating different sections may lead to a more accurate model of measuring the distance between mentions of entities.

Last but not least, our framework can be improved by applying different models to different classes of names and/or pages given that different models and algorithms excel in different categories in many tasks of our experiments, though deeper analysis on the characteristics of names and pages is required towards such improvement. Such improvement may address the problem of geotagging named entities in lists and tables mentioned in Section 7.2.

[2] http://www.wikipedia.org/

Bibliography

[1] Eugene Agichtein and Luis Gravano. Snowball: extracting relations from large plain-text collections. In *Proceedings of the fifth ACM conference on Digital libraries*, DL '00, pages 85–94, 2000.

[2] Einat Amitay, Nadav Har'El, Ron Sivan, and Aya Soffer. Web-a-where: geotagging web content. In *SIGIR*, pages 273–280, 2004.

[3] Ivo Anastácio, Bruno Martins, and Pável Calado. Classifying documents according to locational relevance. In *Proceedings of the 14th Portuguese Conference on Artificial Intelligence: Progress in Artificial Intelligence*, EPIA '09, pages 598–609, 2009.

[4] Ivo Anastcio, Bruno Martins, and Pvel Calado. A comparison of different approaches for assigning geographic scopes to documents. In *In Proceedings of the 1st INForum - Simpsio de Informtica*, 2009.

[5] Lars Backstrom, Jon Kleinberg, Ravi Kumar, and Jasmine Novak. Spatial variation in search engine queries. In *Proceedings of the 17th international conference on World Wide Web*, WWW '08, pages 357–366, 2008.

[6] Lars Backstrom, Eric Sun, and Cameron Marlow. Find me if you can: improving geographical prediction with social and spatial proximity. In *Proceedings of the 19th international conference on World wide web*, WWW '10, pages 61–70, New York, NY, USA, 2010.

[7] Michele Banko, Michael J Cafarella, Stephen Soderland, Matt Broadhead, and Oren Etzioni. Open information extraction from the web. In *IJCAI*, pages 2670–2676, 2007.

[8] Sergey Brin. Extracting patterns and relations from the world wide web. Technical Report 1999-65, Stanford InfoLab, November 1999.

[9] Min Chen, Xing Lin, Yi Zhang, Xingguang Wang, and Hao Yu. Assigning geographical focus to documents. In *Geoinformatics*, pages 1–6, 2010.

[10] Zhiyuan Cheng, James Caverlee, and Kyumin Lee. You are where you tweet: a content-based approach to geo-locating twitter users. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, CIKM '10, pages 759–768, 2010.

[11] Junyan Ding, Luis Gravano, and Narayanan Shivakumar. Computing geographical scopes of web resources. In *Proceedings of the 26th International Conference on Very Large Data Bases*, VLDB '00, pages 545–556, 2000.

[12] Doug Downey, Matthew Broadhead, and Oren Etzioni. Locating complex named entities in web text. In *Proc. of IJCAI*, 2007.

[13] Oren Etzioni, Michael Cafarella, Doug Downey, Ana-Maria Popescu, Tal Shaked, Stephen Soderland, Daniel S. Weld, and Alexander Yates. Unsupervised named-entity extraction from the web: an experimental study. *ARTIFICIAL INTELLIGENCE*, 165:91–134, 2005.

[14] Jenny Rose Finkel, Trond Grenager, and Christopher Manning. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 363–370, 2005.

[15] Luis Gravano, Vasileios Hatzivassiloglou, and Richard Lichtenstein. Categorizing web queries according to geographical locality. In *Proceedings of the twelfth international conference on Information and knowledge management*, CIKM '03, pages 325–333, 2003.

[16] Matthew Hurst, Matthew Siegler, and Natalie Glance. On estimating the geographic distribution of social media. In *International Conference on Weblogs and Social Media*, 2007.

[17] Christian Kohlschütter, Peter Fankhauser, and Wolfgang Nejdl. Boilerplate detection using shallow text features. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, WSDM '10, pages 441–450, 2010.

[18] Guoliang Li, Jun Hu, Jianhua Feng, and Kian-lee Tan. Effective location identification from microblogs. In *Data Engineering (ICDE), 2014 IEEE 30th International Conference on*, pages 880–891, 2014.

[19] Rui Li, Shengjie Wang, Hongbo Deng, Rui Wang, and Kevin Chen-Chuan Chang. Towards social user profiling: Unified and discriminative influence model for inferring home locations. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '12, pages 1023–1031, 2012.

[20] Michael D. Lieberman, Hanan Samet, and Jagan Sankaranarayanan. Geotagging with local lexicons to build indexes for textually-specified spatial data. In *ICDE*, pages 201–212, 2010.

[21] Sharon Myrtle Paradesi. Geotagging tweets using their content. In *FLAIRS Conference*, 2011.

[22] Alexei Pyalling, Michael Maslov, and Pavel Braslavski. Automatic geotagging of russian web sites. In *WWW*, pages 965–966, 2006.

[23] Davood Rafiei, Krishna Bharat, and Ananad Shukla. Diversifying web search results. In *Proceedings of the WWW Conference*, pages 781–790, New York, NY, USA, 2010.

[24] Deepak Ravichandran and Eduard Hovy. Learning surface text patterns for a question answering system. In *Proceedings of the ACL Conference*, pages 41–47, 2002.

[25] Taro Tezuka, Takeshi Kurashima, and Katsumi Tanaka. Toward tighter integration of web search with geographic information system. In *Proceedings of the WWW Conference*, pages 277–286, 2006.

[26] Chuang Wang, Xing Xie, Lee Wang, Yansheng Lu, and Wei-Ying Ma. Detecting geographic locations from web resources. In *Proceedings of the 2005 workshop on Geographic information retrieval*, GIR '05, pages 17–24, 2005.

[27] Casey Whitelaw, Alex Kehlenbeck, Nemanja Petrovic, and Lyle Ungar. Web-scale named entity recognition. In *Proceedings of the 17th ACM conference on Information and knowledge management*, CIKM '08, pages 123–132, 2008.