Towards Geo-referencing Infrastructure for Local News

Guoray Cai*
College Of Information Sciences and Technology
Penn State University
University Park, PA, USA
cai@ist.psu.edu

Ye Tian
College Of Information Sciences and Technology
Penn State University
University Park, PA, USA
yxt157@ist.psu.edu

ABSTRACT

Local news articles are an important source of knowledge about local events, place-specific culture, and peoples' thoughts about their environment. Reliable geocoding of such articles is the first step towards unlocking such local knowledge for community engagement and development. However, existing geo-referencing methods and tools do not work well for local news because they do not reflect the ways local people encode and communicate geographical knowledge. This paper argues that local news requires a different method and infrastructure support for effective geo-referencing. To gain insights on the unique aspects of local gazetteers and the nature of ambiguities, we present an analysis of a collection of local new articles. We found that place references in local news have their special vocabulary, and that their ambiguities are handled differently by local people. We translated such insights into a gazetteer-based geocoding solution that combines progressive geocoding with a smart footprint recommender. Progressive geocoding service uses Nominatim (OpenStreetMap) as the initial gazetteer to jump-start the construction of local gazetteer for a community and by the community. LocusRecommender automatically suggests the best matches from gazetteer ranked by a set of heuristic rules. Preliminary evaluation shows that our smart footprint recommender predicts 80% of the answers by its top-three recommendations.

CCS Concepts

•**Applied computing** → **Cartography;** *Document metadata;* •**Information systems** → *Location based services; Information extraction;*

Keywords

Geographical information retrieval; geo-referencing; local news articles

*Author of correspondence.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

GIR16, October 31-November 03 2016, Burlingame, CA, USA

© 2016 ACM. ISBN 978-1-4503-4588-0/16/10…$15.00

DOI: http://dx.doi.org/10.1145/3003464.3003473

1. INTRODUCTION

Geographic descriptions in texts reflect human conceptualization and experiences of space and places [16, 11]. Different from other forms of geographical data, text-based spatial descriptions are subject to all sorts of ambiguities that prevent effective use [11, 15]. To unlock the huge potentials of the place-based knowledge in textual documents, significant efforts have been invested in geo-referencing texts to coordinate-based spatial "footprints" that capture the geographic content, or focus, of textual data. Geographic footprints can be encoded with varying levels of detail as points, lines, or polygons. Geospatial referencing textual documents refers to the task of discovering location phrases and creating unambiguous representation (or footprints) of the meaning of those textual references [11, 15, 8, 21].

This paper focuses on the geo-referencing problem in the context of geographical retrieval of local news and community forums. Local news is a medium for communication among residents, and they reflect the nature of local geographical knowledge. The residents' local knowledge concerning their living environment is often invisible, descriptive and vague, and thus difficult to collect [23]. Such knowledge is considered important in community planning and participatory decision-making. A locality covered by a local newspaper normally contains one or more geographical communities that share common concerns or passions on something they do. As a special type of community of practice [28], local communities have interests in doing something together and getting better over time, therefore they can be seen as a "social learning system" where actors of the community accumulate and share knowledge and practices through a repertoire of resources like experiences, stories, and tools. Local newspapers can be considered as one important tool to support the accumulation and sharing of local knowledge.

Because local news serves the important role of being repositories of political, social, and cultural knowledge in the community of practice, it is important to make such knowledge accessible to community members who can maintain the awareness of what the community have known or done in the past when facing a new or repeating problem. Towards this goal, geo-referencing local news can contribute to improved accessibility of community knowledge. Moy et al. [20] found that citizens' attention to local news promoted political participation. Despite the importance of geo-referencing local news, we found that existing methods and tools of geo-referencing text are not ideally tuned for processing local news.

Geo-referencing text data to proper geospatial footprints

remains a challenging problem. Two general approaches have been followed: gazetteer-based approach and statistical language model approach. The choice of method is often corpus-dependent. Existing research has been mostly concerned with developing a generic method for geocoding any textual documents, usually with a focus on global news. For newspapers targeted at local communities, geographical locations are likely to be in finer granularity and demand gazetteer-based approach.

Following the gazetteer-based approach, this paper seeks to understand the unique requirement on a computational infrastructure for geo-referencing local news. Our assumption is that local spatial texts have their own vocabulary and they are ambiguous in different ways in contrast to geographic descriptions in global news. If this is the case, two research questions must be answered:

1. *What kind of gazetteers should be used for geo-referencing local news?* Due to uniqueness of each locality, there is a lack of local gazetteers that reflect the richness and details of local spatial language.

2. *What kind of location disambiguation strategies work the best for geo-referencing local news?* Location ambiguity in the local has its own unique nature that presents both challenges and opportunities to resolve them.

We seek answers of above questions by analyzing spatial language patterns in articles collected from the Centre Daily Times [1] newspaper, which publishes local news and events in centre county, Pennsylvania. Using Nominatim as the initial gazetteer and human coders for disambiguation, we find that a small percentage (16%) of the local references belong to case I where toponyms are mapped to footprint representation without ambiguity by directly consulting OpenStreetMap and Google Maps. 59% of the local references belong to case II where location references are matched with multiple entries in the gazetteer, but such ambiguity can be resolved by applying certain heuristics in the literature. The rest 12% of geographical references belong to case III where location references are not resolvable by gazetteer and have to rely on human annotation to create footprints. For those location references that belong to case II, human coders applied a set of disambiguation strategies that is different from generic ones.

Based on such insights, we propose a framework for developing geo-referencing infrastructure that addresses the unique requirements of local news. The approach is implemented by embedding a smart footprint recommender (LocuRecommender) within a progressive geo-coding environment. This configuration of geo-referencing system ensures that a local gazetteer is incrementally enriched through local community use and the demand for the effort of human geo-coding is kept to the minimum.

2. WHAT IS SPECIAL ABOUT LOCAL NEWS AND LOCAL PLACES?

Local newspapers are an important knowledge repository for understanding important local events, politics, and public issues [5]. Local news are textual artifacts of human experience on the places they live and interact with, and they play an important role in the making of place meaning [26]. They reflect what local people think and are written for the local people. Local newspapers contribute to making sense of place [3] in a locality by its "linguistic participation of place" [10] which gives meaning to a place by constructing the synopsis of people, events, locus, and time [29]. In 2011, the Pew Research Center found that nearly three-quarters of U.S. adults are "quite attached to following local news and information", and local newspapers are the most important source for news[2].

Local newspaper production and use facilitate the accumulation of local knowledge. Local knowledge is a collection of facts and relates to the entire system of concepts, beliefs and perceptions that people hold about the world around them. This includes the way people observe and measure their surroundings, how they solve problems and validate new information. It includes the processes whereby knowledge is generated, stored, applied and transmitted to others. Local knowledge is the knowledge that people in a given community have developed over time, and continue to develop. Six characteristics distinguish local knowledge from other knowledge of the world:

- It is based on experience
- It is developed over time by people living in a given community, and is continuously developing.
- It is adapted to the local culture and environment.
- It is embedded in community practices, institutions, relationships and rituals.
- It is held by individuals or communities.
- It is dynamic and changing.

Based on these characteristics, we may anticipate that local knowledge is unique from place to place. Therefore the gazetteer used for geo-referencing local newspaper articles should be place-specific.

In order to gain insights into the types of local spatial language phenomena and their nature of vagueness, we conducted an analysis on a set of local newspaper articles selected from Centre Daily Times[3], a daily newspaper located in State College, Pennsylvania in the United States (see Figure 1). It has about 40,000 readerships primarily from Centre County, PA, with some readerships from neighboring counties. The region is evenly divided among rural, primarily farm, communities and the urban, university town. Centre Daily Times has made all newspaper articles in the last 20 years available in digital form through a subscription service from NewsBank[4].

The goal of this content analysis on local newspaper articles is to understand two questions:

Q1 *What kinds of spatial language are used in the locality*? Answers to this question will make clear the nature of gazetteers needed for geoparsing local newspaper.

[1] www.centredailytimes.com

[2] www.pewinternet.org/2012/04/12/72-of-americans-follow-local-news-closely/

[3] www.centredailytimes.com

[4] infoweb.newsbank.com

Figure 1: The Geographic area served by the Centre Daily Times newspaper

Q2 *What are the common strategies human use to resolve ambiguity in local spatial references*? Answers to this question will reveal proper decision rules in disambiguation algorithms.

Answers to these two questions are reported in section 3.1 and section 3.2. But, first, we describe the experimental design and data analysis methods.

2.1 Data Collection and Experimental Protocol

For this experiment, we downloaded 600 articles (1/2/1997–2/16/1997) of Centre Daily Times newspaper, all featuring local news/events, written by Centre Daily Times editors. The reason that we select earliest available CDT news articles is that we want to leave more recent data studying how local spatial language changes over time.

These articles are parsed to extract their header information and body of text and are stored into a relational database. Each article came with a unique identifier. We hired four researchers who are either native or have been living in the local area for six or more years. This ensures that the researchers have local knowledge similar to those residents in the area.

For the collection of newspaper articles, We conducted content analysis by the four researchers. Each researcher was assigned 150 articles. They were instructed to follow four-step protocol as follows:

[Step 1] Open an article and read through the whole article to understand the story;

[Step 2] Geoparsing: identify all place names (or prepositional phrases) and highlight them in the article;

[Step 3] Geomatching and disambiguation: for each identified location reference, search the best gazetteers to find all matches, compare candidates, and finally make a choice from the candidates. In case no candidate match is found or all candidates are rejected by the human analyst, he or she can create a new footprint and add it to the local gazetteer.

[Step 4] Coding: depending on the gazetteer matching outcomes and the disambiguation strategies used by the analyst, each location reference is coded by a case number (see section 2.4 for details on coding scheme).

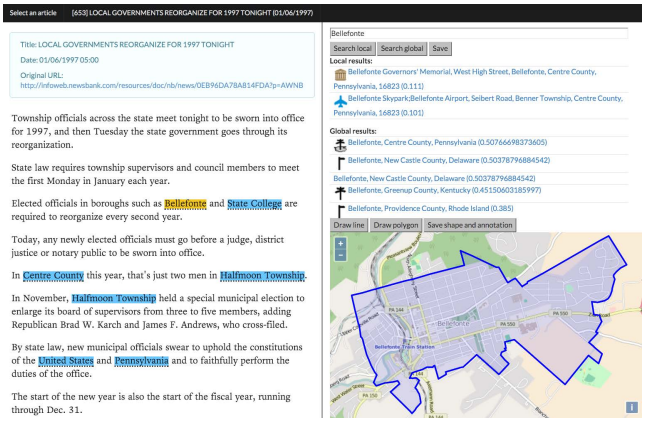

Figure 2: GeoAnnotator workbench

The protocol of the above analysis generally follows the three-phase process of geo-referencing prescribed by Larson [15]. The process is conducted by human analysts with the assistance of the **GeoAnnotator** workbench.

2.2 GeoAnnotator: An Analyst Workbench for Geo-referencing

We designed GeoAnnotator for use in this experiment. It is an interactive geospatial referencing workbench written as a web-based application[5]. Figure 2 shows the GeoAnnotator interface.

The interface is laid out in three view areas:

DocumentView (left side) manages newspaper articles as working document that can be viewed and annotated interactively. A user starts in *DocumentView* to select an article from the collection. During geoparsing, the user reads through the article and recognizes phrases that are location references. When a new phrase is recognized, the user selects that phrase and the system automatically tags it as a location reference (using color code in the view). The annotated text is also automatically copied to the search box in the *CandidatesListView*.

CandidatesListView (upper right) shows the list of geomatching results. When the analyst runs a search of a location reference against both local and global gazetteers for geomatching, all candidate matches are listed in this view, with global search results shown beneath local ones.

FootprintView (bottom right) shows the footprint of the currently selected item in the *CandidatesListView* or *DocumentView*.

The matched gazetteer entries from Nominatim are returned and listed in the *CandidatesListView*. User may click on any of the items in the list to see its footprint in the *FootprintView* below it. To choose the currently viewed footprint as the right answer, simply click on the "SAVE" button. The system will record the geocoding result into an annotation database. Each GeoAnnotation generated is captured in the system as a tuple of five values:

[5]http://gir.ist.psu.edu/annotation

Table 1: Geocoding annotation table.

DocID	LocPhrase	GazName	EntryID
575	"Bellefonte"	"Nominatim"	144829311
Footprint			
"POLYGON((-77.7905379 ..."			

Where *DocID* is a unique identifier to the document that contains the location reference, *LocPhrase*.

The human annotator can also manually create a footprint using the *Draw Line* or *Draw polygon* tool immediately above the *MapView* in the interface. This manual creation of footprint is necessary only in one of the following cases:

1. No match was found from the available gazetteer;
2. None of the matched entries are judged to be appropriate or good enough in the given context.

2.3 Selecting Gazetteers

One goal of this data analysis is to find out how much of the local spatial language is not covered by the available gazetteers. Therefore it is important for GeoAnnotator to support geomatching step with the best available gazetteers.

When making choice of the gazetteers, we considered both commercial and non-commercial sources. Examples of geographical information sources are gazetteers (such as GeoNames.org and Alexandria Digital Library (ADL) gazetteer [12]), GIS databases, or volunteered geographic information (VGI) [9] (such as OpenStreetMap). Each of these data sources have their advantage and limitation. Gazetteers are well-structured and maintained, but they have limited coverage on those less known places at local levels. GIS databases have geographical features in greater geometric precision and details, but they tend to have limited textual metadata for matching to place names. Open Source geographical information now comprises an ever growing part of geographical knowledge, and they frequently include vernacular descriptions of locations, as well as references to imprecise areas. The problem with using VGI is that their local coverage is uneven from locality to locality, and they tend to have various degree of quality and trust. Ideal VGI is supposed to be produced entirely by individual citizens that reside in their locality, and therefore potentially represent the best knowledge about their local.

Generic gazetteers, like GeoNames, follow a strict top-down approach, i.e., the gazetteer data is administered by the organization running the gazetteer. Only this toponymic authority can add places or place types to the gazetteer and correct erroneous entries, which slows down updates and hampers the inclusion of local and often tacit knowledge. Moreover, gazetteers tend to be less sophisticated in representing geographic footprints. Majority of gazetteer entries use points (coordinates) to represent areas (such as a city), which is considered inadequate for local features. One promising use of VGI is the enrichment of gazetteers with vernacular names and vague places [7]. Keßler and colleagues [13, 14] argued that next-generation gazetteers are more likely to be generated from volunteered geographical information. They showed how a gazetteer can be built using a bottom-up approach based on geotagged photos harvested from the web. VGI-based gazetteers are likely to have polygonal footprints for these gazetteer entries.

Based on our analysis of available gazetteers, we have carefully chosen to use three gazetteers:

- **A local gazetteer** – we use a subset of Nominatim covering Centre County (Pennsylvania) and vicinity.
- **Nominatim global** – the whole Nominatim database with global coverage.
- **Google Maps** – global coverage

Nominatim is so far the largest and most detailed gazetteer derived from volunteered geographical information (OpenStreetMap). Google Maps is among the best commercial services for gazetteer look-up purposes. We believe that the combination of the two represents the best available generic gazetteers.

2.4 Coding Scheme

Researchers analyzed each of the recognized place references and assigned a code using the following rules:

A: place names that can be matched to a gazetteer without any ambiguity. These place references may be (1) the full address of a place, or (2) a single and definitive match worldwide.

B: place names that can be matched to a gazetteer with any ambiguity. This refers to the case that a place name that matches to multiple gazetteer entries, and the human analyst is able to identify one of them as the appropriate answer in the given context. Depending on the disambiguation strategy used by the analyst, such cases can be further coded as B1, B2, B3, B4, and B5. These codes reflect the heuristics developed in the literature [2, 17, 18, 4]:

- **B1**: *Prominence rule* prefers features that are more prominent in the geographic world. Prominence is commonly measured by the size of the area, the population that reside in the area, and how central a place is in relation to other places. This is the most popular heuristic when dealing with globally known places [1, 6, 19, 22, 24, 25, 30, 27].
- **B2**: *Ontology rule* prefers features that share the same geographical context defined by a hierarchical geographic ontology. Examples of applying this rule include [1, 30, 25, 24, 22].
- **B3**: *Distance rule* prefers locations that are close to other places previously mentioned in the text, measured by the geographical distance.
- **B4**: *Default sense* prefers entities that are more frequently referred to in the whole collection of text.
- **B5**: *Ontology first and distance refine* uses ontology focus as the first heuristic for resolving ambiguity. If ambiguity remains after considering ontology, human analysts will switch to distance-related heuristics. This is a mix of B2 and B3 heuristics.

C: Place references that are not successful in finding a spatial footprint from gazetteers. In this case, gazetteer is not helpful in suggesting the correct footprint. Human annotator will create the footprint by drawing shapes in the MapView.

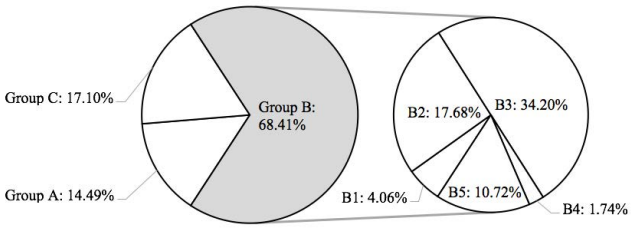

Figure 3: Categorizing place ambiguities and resolution strategies

3. FINDINGS

There are 1957 place names recognized from the 600 articles. Figure 3 shows the how these place names fall into different situations defined by the coding scheme. Now we state our findings related to the two research questions: (1) *What kinds of spatial language are used in the locality*? (2) *What are the common strategies humans use to resolve ambiguity in local spatial references*?

3.1 Findings: Patterns of spatial language

To answer the first question, we discovered the following patterns:

- **Finding 1**: *There are significant portion of place names in local news that are vernacular, vague places, or finer granularity places that were not found in gazetteers.* In our analysis, 83% of the identified place names (case A + case B) directly match with GeoAnnotator gazetteers through either local or global search, and one of the matches is chosen as the referencing footprint. These place references are definitive place names and have well-defined location and boundaries. The rest 17% place references (case C) did not find match, and they tend to be complex prepositional phrases, vague regions, or locally-known vernacular places. Much of the latter group also refer to places of finer granularity that is beyond what generic gazetteers capture.

- **Finding 2**: *The majority of those place names found matches from gazetteers are found to be ambiguous.* Among 83% of the place names (case A + case B) that found matched in GeoAnnotator, only 14.49% of place names (case A) found reasonable footprints without any ambiguity, while the rest 68.41% (case B) were found to be ambiguous.

- **Finding 3**: *For those place names that were not able to find proper footprints in gazetteers, humans can create footprints with ease when supported well.* For those 17% place references that were not successful in finding a spatial footprint from gazetteers, we observed researchers creating footprint by drawing shapes in the map view. Because GeoAnnotator helped human in locating the area and providing good spatial context, drawing the footprint in the map view is relatively straightforward.

- **Finding 4**: There are high degree of verbatim repetition of place references across local news articles over time. Human coders found that they repeatedly encounter common place references, some of which more frequent than others. If the system has the ability to learn such vocabularies from human's manual annotation, we are likely to end up with a local gazetteer that is enriched over time.

3.2 Findings: Disambiguation Strategies

For those in case B, we further observed the strategies human analysts use in resolving ambiguity. This effort revealed the following patterns:

- **Finding 5**: *Ontological focus.* In our observations and interview with the human analysts, we found that humans mentally maintain the awareness of what is the general geographic area that the current discourse is focused on. As a discourse evolves and unfolds, such geographical focus is also moving. For example, in "In Benner Township, residents may leave their Christmas trees at the *township building*", the referenced place *Benner Township* is an ontological hypernym of *township building*. This provides additional clues for analysts to disambiguate "township building". Footprints that are more coherent with the ontological focus of the current discourse is more likely to be the correct choice. The number of observations we have for the case of using ontological focus for disambiguation is labeled as B2 in Figure 3, which account for 17.68%.

- **Finding 6**: *Ontology first, distance refines.* Human analysts tend to use ontology focus as the first heuristic for resolving ambiguity. If ambiguity remains after considering ontology, human analysts will switch to distance-related heuristics. The number of observations we have for the case of using the distance heuristic for disambiguation is labeled as B5 in Figure 3, which account for 10.72%.

- **Finding 7**: *Local vs global match.* Humans tend to search local gazetteer first; if a match is found, no further search will be attempted. If local search fails to return satisfactory answer, they will do a global search. The number of observations we have for the case of matching to local features is labeled as B3 in Figure 3 which accounts for 34.2%. This makes sense, since local readers are unlikely to be aware of most of these interpretations offered in global gazetteers or databases, and thus there is little or no need to consider them as possibilities in the geo-disambiguation step. The assumption made here is that newspaper reader's spatial lexicon – those locations that they identify and place on the map without any evidence – is very limited. The reader's local spatial lexicon differs from place to place. In most cases, the local lexicon supersedes the global lexicon.

- **Finding 8**: *Prominence for global features, but not for local features.* For matches returned from global search, human coders reported that they consider prominence of the places more than anything else. However, people do not consider prominence for local features. The number of observations we have for the case of matching to global features is labeled as B1 in Figure 3 which accounts for 4.06%.

In conclusion, heuristic rules that human coders use in disambiguation of place references in local newspaper are

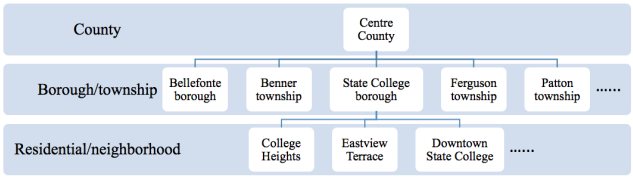

Figure 4: A geographic ontology for a locality

very different from those used in resolving ambiguities in global news. In particular, we found dominant uses of ontology rule and distance rule, either individually or in combination, while prominence rule and default sense rule are rarely used. This makes sense, because if we mainly rely on prominence rule and default sense rule, it would be nearly impossible for the less prominent places to emerge as correct interpretation, since these smaller places tend to have low occurrence in existing news articles corpora.

If place ontology places such an important role in local place disambiguation, what are the place ontologies that local people have or maintain? Our study here did not answer this question well, but it seems that human coders maintained an ontology that has three levels of granularity: county, townships, and neighborhoods. An example of the local ontology is shown in Figure 4, which corresponds to the ontology of Centre County, Pennsylvania.

From our analysis, it becomes obvious that generic gazetteers are not adequate for geoparsing and geomatching for local news. This is consistent with the work of Davies et al [7] who found that none of the existing geoparsing methods can easily pick up locally-known vernacular places.

4. TOWARDS HUMAN-CENTERED GEO-REFERENCING INFRASTRUCTURE

Based on the insights we obtained from analyzing place references in local newspaper articles, we conclude that local news are unique enough to require special geo-referencing infrastructure. This calls for locally valid gazetteers and disambiguation heuristics. Our solution is to extend GeoAnnotator in two directions:

Incremental bottom-up gazetteer enrichment. Many applications, especially those involving local spatial language, suffer from lack of adequate gazetteer. Most gazetteers follow a strict top-down approach, i.e. the gazetteer data is administered by the organization running the gazetteer. Only this toponymic authority can add places or place types to the gazetteer and correct erroneous entries, which slows down updates and hampers the inclusion of local (and often tacit) knowledge. Moreover, in most gazetteers information on geographic footprints is limited to a single coordinate pair, representing the center of a city, administrative district or street. One promising use of VGI is the enrichment of gazetteers with vernacular names and vague places [7]. Keßler and colleagues [13, 14] argued that next generation gazetteers are more likely to be generated from VGI. They showed how a gazetteer can be built using a bottom-up approach based on geotagged photos harvested from the web. VGI-based gazetteers are likely to have polygonal footprints for these gazetteer entries. This inspired us to develop a progressive geospatial referencing framework in Section 6.

Suggestive geocoding Geocoding (the tasks in step 2 and

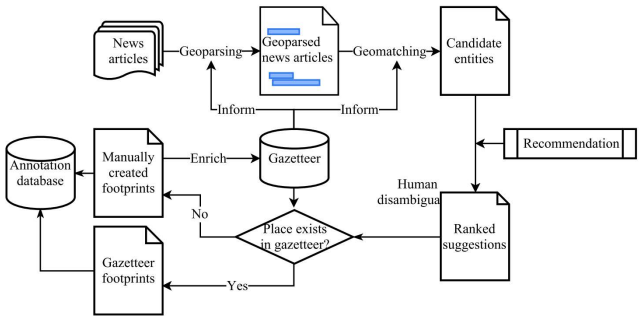

Figure 5: Progressive Geospatial Referencing Framework

step 3) requires deep, human-level knowledge to evaluate a potentially large list of candidate matches and make a choice. These tasks are best done by bringing human and computer to work collaboratively. A simple way to do this is to have computer doing step 2 (generating all the candidate matches) and human take care of step 3 (disambiguation). Alternatively, computer can further reduce human efforts in step 3 by ranking the candidates such that those with higher degree of likelihood are placed on the top of the list, and those candidates that are deemed impossible after considering the context are eliminated from the list. This principle corresponds to the idea of suggestive geocoding.

If we incorporate the above ideas into the gazetteer-based three-step process, the conceptual flow of the progressive geospatial referencing framework can be expressed by Figure 5.

Based on the above framework, we have implemented GeoAnnotator+ as a geo-referencing infrastructure specialized for local news and media.

5. GEOANNOTATOR+: A PROGRESSIVE GEOREFERENCING WORKBENCH

GeoAnnotator+ is the result of extending GeoAnnotator with progressive geocoding and smart footprint recommendation. The fundamental ideas of GeoAnnotator+ can be summarized into a few points:

1. *Start with a VGI-based gazetteer that is best for the target locality.* The quality of this gazetteer does not matter so much, as it only provides a starting point for us to bootstrap the system. In the current implementation of GeoAnnotator+, we bootstrap the system by taking Nominatim[6] as the initial source of gazetteer. Nominatim provides a search engine API for using OpenStreetMap data as a proxy gazetteer to match place names. The quality of the footprints in Nominatim is quite good, as it normally returns matches of local features by good approximation of their real shape (points, lines, or polygons). We choose Nomination as the initial gazetteer because it is considered the best VGI-based geographical data sources with good coverage in most localities worldwide.

2. *Progressive geocoding to enrich gazetteer incrementally.* Provide a semi-automated workbench for human analysts to evaluate all matches and make a choice. If none

[6]https://nominatim.openstreetmap.org

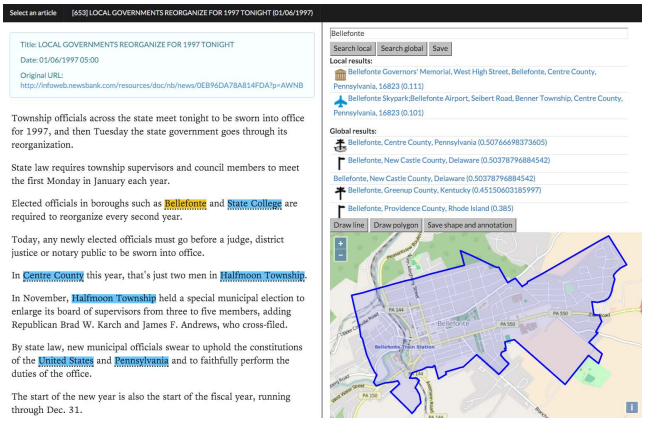

Figure 6: GeoAnnotator+ workbench

of the matches are appropriate, analysts will create the footprint in the workbench. The system not only remembers the result of this human-generated geocode, but also adds a new entry to its gazetteer as enrichment. In this way the system is able to leverage real local and community conversations and outsource this geocoding task to local community members. Hence, new gazetteer entries learned from human reflect local spatial language and local knowledge.

3. *Smart footprint recommender* ranks the matched candidates by their likelihood of being correct. Based on our understanding of how humans disambiguate multiple interpretations of place names, a computer can play the role of a smart recommender by automating a set of heuristic rules. By presenting the candidate list in the order of likelihood, human annotators are more likely to find the answer by exploring only the top few candidates. This creates savings to human cognitive effort.

Figure 6 shows the interface design of of GeoAnnotator+ workbench. Although its layout of three views (DocumentView, CandidateListView, and FoortprintView) is similar to GeoAnnotator in Figure 2, the CandidateListView in GeoAnnotator+ is now showing the ranked recommendation of footprint candidates according to the a *Smart footprint recommender*.

6. LOCURECOMMENDER: A SMART FOOTPRINT RECOMMENDER

We have implemented a smart footprint recommender: *LocuRecommender*. It works within GeoAnnotator+ and ranks gazetteer-matched candidates according to an ordered set of disambiguation strategies. The general framework for LocuRecommender is described in **Algorithm 1**.

The question is: *what are these disambiguation strategies for local news?* Although we still do not have a comprehensive understanding of the way humans evaluate candidate footprints, there are several sources of knowledge that inform our design of LocuRecommender.

First, we roughly know from previous studies that there are four commonly-used heuristics when building automated disambiguation tool. They are default sense rule, prominence rule, ontology rule, and distance rule [2, 17, 18, 4]. However, we do not know to what degree such heuristics are

Algorithm 1: LocuRecommender algorithm: general framework

```
1  Function Recommend(text):
       Data: place name as a string of text text
       Result: recommendation as a list of sorted entities
               result
2      result ← empty list;
3      previous_annotations ←
       InitializePreviousAnnotations();
4      previous_annotation_results ←
       SearchInPrevious(text);
5      local_results ← SearchInLocal(text);
6      global_results ← SearchInGlobal(text);
7      result ← concatenate(previous_annotation_results,
       local_results, global_results);
8      return result
9  Function InitializePreviousAnnotations():
10     previous_annotations ← empty hash table;
11     for annotation ∈ previous_annotations do
12         Increase the count of occurrences where
           annotation.text is annotated as entity
           annotation.place_id;
13         Update previous_annotations with the new
           count;
14     return previous_annotations
```

applicable for local news and whether they reflect the ways that people in a local community disambiguate local spatial language. Addressing this lack of understanding was one of the motivation for us to conduct content studies in section 2. In section 3.2, we have found that interpretation of geographical references in local news follows the principles of "ontology first, distance refine." In other words, people tends to use ontology focus as the first heuristic, followed by distance-related heuristics. Default sense rule and prominence rule do not apply to location disambiguation of local news. These principles are reflected in our design of search algorithm in LocuRecommender (see **Algorithm 2**).

In summary, LocuRecommender within GeoAnnotator+ ranks gazetteer-matched candidates incorporate the following strategies:

- **Prioritize previous annotations**. The system maintains a list of previously mentioned place names in the whole corpus along with the entities they have been resolved to. Given a place name, the system first searches through this list and examine how the exact same name has been previously resolved. This is based on *Finding 4*.

- **Exploit local gazetteer**. We incorporate two gazetteers in the system: a global gazetteer that contains all knowledge from Nominatim, and a local gazetteer that only contains the Centre County entities from Nominatim. In the recommendation algorithm, results given by the local gazetteer are always prioritized over those from the global one. This is based on *Finding 7*.

- **Rank search results with regard to a moving focus**. The system keeps track of the moving focus of an article as we proceeds with testing. We use the

Algorithm 2: LocuRecommender algorithm: searching functions

```
1  Function SearchInPrevious(text):
       Data: place name text
       Result: recommended results based on previous
               annotations
2      results ← previous_annotations[text], a list of all
       (entity, count) tuples that represent the frequency
       of previous resolutions of text;
3      sort results by count in descending order;
4      return results
5  Function SearchInLocal(text):
       Data: place name text
       Result: recommended results based on local
               gazetteer
6      results ← all entities from the local gazetteer with
       the name of text;
7      if an immediate annotation preceding text is found
       then
8          current_focus ← immediate annotation
           preceding text;
9      else
10         current_focus ← Centre County;
11     for entity ∈ results do
12         entity.ontological_distance ← the distance
           between entity and current_focus in the
           ontology tree in the gazetteer;
13         entity.physical_distance ← the Haversine
           distance between entity and current_focus;
14     sort results by their ontological_distance in
       ascending order. for results with the same
       ontological_distance, sort by physical_distance in
       ascending order;
15     return results
16 Function SearchInGlobal(text):
       Data: place name text
       Result: recommended results based on local
               gazetteer
       /* gather results and current focus in the
          same fashion with SearchInLocal, but
          using global gazetteer instead          */
17     for entity ∈ results do
18         entity.ontological_distance ← the distance
           between entity and current_focus in the
           ontology tree in the gazetteer;
19         entity.prominence ← the importance score of
           entity;
20     sort results by their prominence in descending
       order. for results with the same prominence, sort
       by ontological_distance in ascending order;
21     return results
```

immediate preceding annotation of the current place name as the current focus, and gazetteer matches for the current place name are sorted by their distance to the focus. This design choice incorporates *Finding 5*.

- **Use different ranking strategy for local and global entities**. For local search results, the system primarily ranks them by their distance to the current focus in terms of geographical ontology tree. For example, the distance between a place and its geographic hyponym is close than the distance between a place and its siblings. For global search results, the system leverages the prominence score of entities from the gazetteer in ranking. The prominence score is calculated by considering the number of Wikipedia articles that have link to the place entity. When the article does not have a Wikipedia entry, the default score is based on the object rank (state, county, city, etc.). This is based on *Finding 8* and *Finding 6*.

We conducted a preliminary evaluation of LocuRecommender to find out if adding LocuRecommender does produce savings on human efforts. We downloaded another 60 articles of Centre Daily Times newspaper from NewsBank (2/16/1997–2/28/1997, 10% the size of our annotation corpus) and conducted geocoding using GeoAnnotator+ with LocuRecommender. We annotated 175 place names and analyzed which one in the ranked list is judged as correct one by human. The result is summarized in Figure 7.

From these results, we can make following observations:

- With the help of LocuRecommender, 54.3% of place references were resolved by reusing previous annotations. A match to previous annotation allows a quick decision on the footprint without extensive search and evaluation by humans. This is on average ten times cheaper in terms of saved human effort.
- 84.4% of the place references were resolved by selecting one of the top 3 recommendations (whether they are previous annotations, or results of local/global search). 77.1% of the place references were resolved by consulting the top 1 recommendation.
- 13.7% of the place references were not resolvable. These place names are not encountered in previous annotations, nor can be retrieved from the gazetteer. However, compared with 17.1% unresolvable in Figure 3, the result does show that the gazetteer was enriched in the process of geocoding the first 600 articles. We expect that the trend of increasing matching rate is going to continue as GeoAnnotator+ is used more over time.

7. CONCLUSION

The need to unlock the rich local knowledge from the vast text-based local news media and community social media requires establishing geospatial referencing capability for hundreds of localities. We present a progressive geo-referencing approach to address the need for geospatial referencing of news articles on local and community level. Our approach is informed by analyzing local newspaper use of spatial language and by observing human annotators in their heuristic decision-making for resolving ambiguities. If successful

Resolution status	Count	Percentage
Previous annotation		
Rank 1	94	53.7%
Rank 2–3	1	0.6%
Rank 4+	0	0%
Previous annotation total	95	54.3%
Local search result		
Rank 1	30	17.1%
Rank 2–3	8	4.6%
Rank 4+	3	1.8%
Local search result total	41	23.4%
Global search result		
Rank 1	11	6.3%
Rank 2–3	2	1.1%
Rank 4+	2	1.1%
Global search result total	15	8.6%
Not resolvable	24	13.7%
Grand total	175	100%

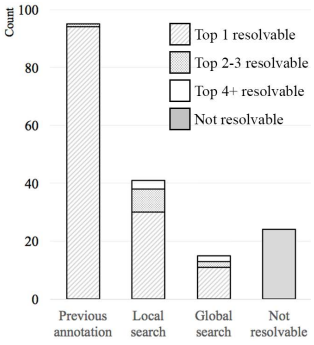

Figure 7: Performance of LocuRecommender

along this direction, the infrastructure like GeoAnnotator+ can provide geo-referencing services to many localities similar to State College / Centre County, PA.

While a system that follows the progressive geospatial referencing framework is expected to enrich its gazetteer and improve its performance over time, it does require human effort to deal with uncertainties, and complex geographical descriptions that have no gazetteer match (e.g. a long prepositional phrase representing a place). The key challenge here is to reduce the human effort to the minimal and accelerate the process of incremental improvement. This can be accomplished partially by presenting / suggesting to the user the best ranking of candidate matches.

There are a number of research questions that remain to be addressed by future work. First, we need to show how effective of the proposed progressive geo-referencing approach in generating locally specialized gazetteers. This can be done by using GeoAnnotator+ to conduct more human assisted geocoding of local news and assess the improvement in gazetteer matching in LocuRecommender. To ensure reliability of the findings, multiple human coders should be assigned to work on the same set of news articles, so that similarities and differences in coders' judgment can be compared for cross validation. Second, we need to assess the effectiveness of the LocuRecommender in footprints disambiguation and recommendation. This can be accomplished through lab controlled experiments comparing human georeferencing efforts with and without the help of LocuRecommender. We plan to continue our research using articles from the Centre Daily Times in the past 20 years. The effect of alternative heuristic rules for disambiguation can also be compared to refine our findings.

Acknowledgment

This work is partially supported by a grant from the National Science Foundation under award IIS-1211059. The work of the first author is also partially supported by a grant from the Chinese Natural Science Foundation under award 71373108.

References

[1] E. Amitay, N. Har'El, R. Sivan, and A. Soffer. Web-a-where: geotagging web content. *Proceedings of SIGIR '04 conference on Research and development in information retrieval*, pages 273–280, 2004.

[2] I. Bensalem and M. K. Kholladi. Toponym disambiguation by arborescent relationships. *Journal of Computer Science*, 6(6):653–659, 2010.

[3] C. Buchanan. Sense of place in the daily newspaper. *Aether: The Journal of Media Geography*, 4(March):62–84, 2009.

[4] D. Buscaldi. Approaches to disambiguating toponyms. *SIGSPATIAL Letters*, 3(2):16–19, 2011.

[5] D. Butt. Local Knowledge: Place and New Media Practice. *Leonardo*, 39(4):323–326, 2006.

[6] P. Clough. Extracting metadata for spatially-aware information retrieval on the internet. In *Proceedings of the 2005 workshop on Geographic information retrieval - GIR '05*, page 25, New York, 2005. ACM Press.

[7] C. Davies, I. Holt, J. Green, J. Harding, and L. Diamond. User Needs and Implications for Modelling Vague Named Places. *Spatial Cognition & Computation: An Interdisciplinary Journal*, 9(3):174–194, 2009.

[8] D. W. Goldberg, J. P. Wilson, and C. A. Knoblock. From text to geographic coordinates: the current state of geocoding. *URISA Journal*, 19(1):33, 2007.

[9] M. F. Goodchild. Citizens as sensors: The world of volunteered geography. *GeoJournal*, 69(4):211–221, 2007.

[10] S. Harrison and D. Tatar. Places: People, Events, Loci – the Relation of Semantic Frames in the Construction of Place. *Computer Supported Cooperative Work (CSCW)*, 17(2):97–133, 2008.

[11] L. L. Hill. *Georeferencing: the geographic associations of information*. MIT Press, Cambridge, Mass, 2006.

[12] L. L. Hill, J. Frew, and Q. Zheng. Geographic names: The implementation of a gazetteer in a georeferenced digital library. *D-Lib Magazine*, 5(1), 1999.

[13] C. Keßler, K. Janowicz, and M. Bishr. An agenda for the next generation gazetteer: Geographic information contribution and retrieval. *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems - GIS09*, (c):91–100, 2009.

[14] C. Keßler, P. Maué, J. T. Heuer, and T. Bartoschek. Bottom-up gazetteers: Learning from the implicit semantics of geotags. *Lecture Notes in Computer Science*, 5892 LNCS:83–102, 2009.

[15] R. R. Larson. Geographic Information Retrieval and Spatial Browsing. In L. Smith and M. Gluck, editors, *GIS and Libraries: Patrons, Maps and Spatial Information*, pages 81–124. University of Illinois, Urbana-Champaign, 1996.

[16] R. Laurini. Geographic ontologies, gazetteers and multilingualism. *Future Internet*, 7(1):1–23, 2015.

[17] J. L. Leidner. *Toponym Resolution in Text*. PhD thesis, University of Edinburgh, 2007.

[18] M. D. Lieberman, H. Samet, and J. Sankaranarayanan. Geotagging with local lexicons to build indexes for textually-specified spatial data. In *IEEE ICDE Conference*, pages 201–212. IEEE, 2010.

[19] B. Martins, H. Manguinhas, J. Borbinha, and W. Siabato. A geo-temporal information extraction service for processing descriptive metadata in digital libraries. *e-Perimetron*, 4(1):25–37, 2009.

[20] P. Moy, M. McCluskey, K. McCoy, and M. Spratt. Political correlates of local news media use. *Journal of Communication*, 54(3):532–546, 2006.

[21] R. Purves. Geographic information Retrieval: Are we making progress? In *NCGIA Specialist meeting on Spatial Search*, pages 1–6, 2014.

[22] R. S. Purves, P. Clough, C. B. Jones, A. Arampatzis, B. Bucher, D. Finch, G. Fu, H. Joho, A. K. Syed, S. Vaid, and B. Yang. The design and implementation of SPIRIT: a spatially aware search engine for information retrieval on the Internet. *International Journal of Geographical Information Science*, 21(7):717–745, 2007.

[23] H. Rantanen and M. Kahila. The SoftGIS approach to local knowledge. *Journal of Environmental Management*, 90(6):1981–1990, 2009.

[24] E. Rauch, M. Bukatin, and K. Baker. A confidence-based framework for disambiguating geographic terms. In *Proceedings of the HLT-NAACL 2003 workshop on Analysis of geographic references*, pages 50–54, 2003.

[25] N. Stokes, Y. Li, A. Moffat, and J. Rong. An empirical study of the effects of NLP components on Geographic IR performance. *International Journal of Geographical Information Science*, 22(3):247–264, 2008.

[26] Y.-F. Tuan. Language and the Making of Place: A Narrative-Descriptive Approach. *Annals of the Association of American Geographers*, 81:684–696, 1991.

[27] M. Vasardani, S. Timpf, S. Winter, and M. Tomko. From descriptions to depictions: A conceptual framework. In T. Tenbrink, J. Stell, A. Galton, and Z. Wood, editors, *Spatial Information Theory - 11th International Conference, COSIT 2013, Scarborough, UK, September 2-6, 2013.*, volume LNCS 8116, pages 299–319. 2013.

[28] E. Wenger. *Community of Practice: Learning, Meaning and Identity*. Cambridge University Press, New York, 1998.

[29] S. Winter and M. Truelove. Talking About Place Where it Matters. In *Cogn. Linguist. Asp. Geogr. Sp.*, pages 121–139. 2013.

[30] W. Zong, D. Wu, A. Sun, E.-P. Lim, and D. H.-L. Goh. On assigning place names to geography related web pages. In *Proceedings of the 5th ACM/IEEE-CS joint conference on Digital libraries - JCDL '05*, page 354, New York, 2005. ACM Press.