CO-253

GEOSPATIAL REASONING IN A NATURAL LANGUAGE PROCESSING (NLP) ENVIRONMENT

BITTERS B.
Science Applications Int'l Corporation, STERLING, UNITED STATES

ABSTRACT

We outline a strategy to capture explicit, implicit and vague references to space in text documents with special emphasis on documents containing unstructured and semi-structured text. This extraction process is a prelude to performing follow-on analysis of characteristics pertaining to events and more complex spatio-temporal reasoning operations. Recognizing common use spatial expressions, involves automatic recognition and understanding of explicit, implicit and vague references to both space and time. These references are not only in the form of explicitly named geographic entities, but are also in the form of geographic named entities suffixed and prefixed by common-use spatial predicates – modifiers that can drastically change the spatial meanings of referent terms and expressions. Capturing the essence of these common-use modifiers and adjusting the spatial locations of referent terms and expressions accordingly is a primary goal of this research.

INTRODUCTION

In its simplest sense, an event is anything that happens; anything that one could plot on a timeline. Events can be continuing occurrences or they can be conceptually instantaneous. In written discourse, what defines an event is often dependent on the application and domain, but generally events must be measurable in time and they must be conveyed in verbal or written form using finite verbs and other identifiable characteristics. According to the Merriam-Webster Dictionary an event is:

"*Something that happens, an occurrence*;" or in a more deterministic sense: "*The fundamental entity of observed physical reality represented by a point designated by three coordinates of place and one of time in the space-time continuum postulated by the theory of relativity*."

For researchers analyzing language in the news media (Bell, 1991), an event is characterized by: who, what, when, where and some form of attribution. Note that, unlike some definitions of "event," these definitions do not specify that an event involves a change of state, nor does it attempt to distinguish events from processes or states. It should also be noted, in each of these definitions, there are assumed temporal and spatial elements. Those essential characteristics describing an event are: Event type, Actors, Location, Date/Time/Duration, Topic/Purpose, Status and Outcome. Automated interpretation of events at this level of detail is currently beyond the capabilities of most NLP systems (Ferro, et al, 2000). However, using current technology, it is possible to individually extract many of these descriptive elements from a discourse and then reassemble them into relevant descriptions of events, situations and/or processes. In this research we have concentrated on the extraction of geographic entities and the refinement of their derived geographic coordinates.

BACKGROUND AND OBJECTIVES

Geospatial reasoning is concerned with intelligent processing of location information. In the case of natural language processing, geospatial reasoning involves transforming words into specific geographic coordinates. Our objective to go beyond commercially available NLP capabilities and compute refined precise and accurate ground-space positions from spatial expressions. Of primary concern is extracting and understanding spatial information from unstructured and semi-structured text. Structured text, data stored in a formal row/column format, is usually designed in such a way that data type, format, and spatial characteristics are relatively simple to discover in an automated setting. However, the majority of the world's information is stored as natural language text in the form of semi-structured text (i.e. Web pages) or unstructured text (all other forms of written communications) (McCallum, 2005). Rule based approaches to deciphering semantic and syntactic detail in texts have been explored extensively and generally have been shown to be too cumbersome for automated processing. However, in recent years statistical and machine learning techniques have provided robustness in the extraction and understanding of written discourse (Wang, 2005; McCallum, 2003; Kleinberg, 2002). Statistical machine learning methods have revolutionized the process of recognizing and extracting textual references to geographic locations. However, disambiguation of geographic named entity extraction results is still problematic. A wide variety of heuristic evidence is available that can be used to disambiguate each named geographic entity and this is an area of intense research. Another area of concern is resolving vague and inexact

references to spatial locations. Computing imprecise spatial references into precise and accurate locations on the Earth's surface is still a dilemma.

In recent years, numerous researchers have proposed approaches to this dilemma (Amitay, 2004; Leidner, 2007; Li, 2007; Lieberman, 2010; Martins, 2008; Overell, 2009; Purves, 2007; Rauch, 2003; Teitler, 2008; and Volz, 2007). Individually these approaches do not provide consistently satisfactory results. However, very promising but computationally expensive results can be attained when these approaches are incorporated.

Named entity extraction is concerned with deriving precise and accurate geo-locations for named entities. This not only involves identifying named entities and their location; it also involves capturing expressions and terms associated with named entities – expressions and terms containing spatial modifiers. In common discourse, these spatial modifiers can radically change the spatial meaning of a referenced place name. For instance, the spatial reference to "northeast of Washington, DC" is far different than the location derived using most commercial named entity extraction software. Extraction software would position the named entity in the geographic center of mass of Washington, DC; when, in fact, the actual referenced location (including the spatial modifier) is outside the DC city limits.

After extraction of associated spatial modifiers, they must be qualified and applied as topologic and geometric transforms to the gazetteer derived geographic coordinates. Our previous spatial relationship research (Bitters, 2009) will allow adjustment of geographic positions by exploiting spatial modifiers commonly used in day-to-day discourse. Through automated extraction of named geographic entities and their spatial modifiers, we will perform a new form of region connection calculus (RCC) to adjust gazetteer-derived positions to more realistic locations on the ground.

APPROACH & METHODS

Social network analysis is a process for measuring and graphically analyzing relationships and flows between people, groups, organizations, animals, computers or other information/knowledge processing entities. Nodes in a network are the people, objects or groups while the links show relationships or flows between the nodes. Social network analysis provides a means for both visual and mathematical analysis of the relationships between different groups or entities. This research effort attempts to apply the basic principles of social networking to the preparation and augmentation of future GIS databases with a particular emphasis on real-time simulation databases. The significance of this portion of the research effort is that it has:

- Produced semantic models of common-use spatial modifiers and expressions for use in geographic named entity extraction.
- Produced a probabilistic knowledgebase to derive new feature content in natural and cultural landscapes.
- Produced a semantic model of toponyms.
- Demonstrate the power of probabilistic geospatial ontologies in inference.
- Demonstrate a common data structure to store logical descriptions geospatial data.

Geographic Named Entities. A *toponym* (from the Greek: τόπος, tópos "place" and ὄνομα, ónoma, "name") is any generally accepted name for a real or imaginary place, as well as names derived from places or regions. Various types of toponyms are available for a particular place: the official name, native-language official name, common-use names, foreign language variants, transliterations, local-use variants, and slang and colloquial forms. Any toponym that differs from that used in the official or well-established language is termed an exonym (from the Greek: ἔξω, éxō, "out" and ὄνομα, ónoma, "name".) Any toponym used by native speakers is called an endonym, (from the Greek ἔνδον, éndon, "within" or αὐτό, autó, "self" and ὄνομα, ónoma, "name".) This category includes both official, legally sanctioned, native-language toponyms, official, English-language, legally sanctioned transliterations, and any common-use, native-language toponyms in use. Therefore, any place name other than the officially sanctioned, legal, native-language name would be an exonym. For example, India, Germany, Greece, Japan, and Korea are the English exonyms corresponding to the endonyms Bharat, Deutschland, Ellas, Nippon/Nihon, and Hanguk/Chosun. Table 1 provides a representative sample of toponyms for the country name Afghanistan. This is only a partial list of the toponyms for Afghanistan, and does not include many of the foreign language forms of the country name. However, it does illustrate the wide variaty of place names available for a particular location. To ensure efficient operation and success in geographic named entity extraction operations, a comprehensive archive of available place name variants must be readily available – an archive including not only official toponyms but also as many unofficial, slang variants, colloquial forms, even foreign language forms of place names.

Augmenting gazetteer records with all available name variants is a first step in preparing for geographic named entity extraction. The more toponym variants available within a gazetteer knowledgebase, the higher the likelihood is of successful extraction. However, this does not insure that each extracted reference will be disambiguated to a single discrete named entity. Other heuristics are necessary to insure full disambiguation of each geographic name to a single unique location.

Table 1. Partial Toponym List for the Country of Afghanistan.

Description	*Toponym*	*Toponym Type*
Official Persian Name	جمهوری اسلامی افغانستان	**Endonym**
Official Pashto Name	د افغانستان اسلامي جمهوريت	**Endonym**
Official Transliterated Persian Name	**Jomhūrī-ye Eslāmī-ye Afġānistān**	**Endonym**
Official Transliterated Pashto Name	**Da Afġānistān Islāmī Jomhoriyat**	**Endonym**
Official English Language Name	**Islamic Republic of Afghanistan**	**Exonym**
Formal Dutch Language Name	**Islamitische Republiek Afghanistan**	**Exonym**
Formal Pashto Language Name	**Dowlat-e Eslami-ye Afghanestan**	**Exonym**
Formal French Language Name	**République islamique d'Afghanistan**	**Exonym**
Formal Norwegian (Bokmål) Name	**Den islamske stat Afghanistan**	**Exonym**
Formal Norwegian (Nynorsk) Name	**Den islamske staten Afghanistan**	**Exonym**
Formal Portuguese Language Name	**República Islâmica do Afeganistão**	**Exonym**
Formal Spanish Language Name	**República de Afganistán**	**Exonym**
Common-Use Native-Language Name	افغانستان	**Endonym**
Common-Use Foreign Language Name	**Afġānistān, Afghanestan, Afganistán, Afeganistão**	**Exonym**
Common-Use English Name	**Afghanistan**	**Exonym**
Obsolete Variants	**Gandhara, Upaghanistanaha**	**Endonym**

.

Gazetteer Databases. Several different open-source, worldwide gazetteer databases are readily available for free download from the Internet. The most comprehensive are National Geospatial-Intelligence Agency' (NGA) Geographic Names Server (GNS) and the gazetteer available from Geonames.org. In addition, US Geological Survey provides comprehensive names data for the US and its territories and GeoBase Canada provides authoritative place names for all of Canada.

Preprocessing gazetteer data to incorporate the best of all available source data is an essential first step to optimize geographic named entity extraction operations. Gazetteers include many more types of feature than just populated places. Significant reductions in processing time can be attained by limiting the size of names databases to only those regions discussed within each document. Rather than using a gazetteer of the entire world, a comprehensive names database partitioned by country reduces memory requirements. Further, maintaining a separate set of country files containing only populated place names will often reduce the size of gazetteer data by as much as half. Our design will include two sets of country files providing worldwide gazetteer data, one containing only populated places and the second containing all other named entities.

Ambiguity in Geographic Names. Even with a comprehensive archive of toponyms, a geographic feature extraction software tool must contend with vague and ambiguous references to geographic locations in text documents. For a variety of reasons an explicit reference to a named geographic entity will often not be fully understandable by extraction software. Table 2 provides an example of the occurrence of the most frequent populated place names in the United States. The name Fariview occurs 218 times in Alabama, Arizona, Arkansas and 27 other states. Deciphering in which state an isolated text reference to "Fairview" might apply is impossible without additional information.

Table 2. Top Ten Populated Place Names in the U.S. and the Number of Occurrences of each. (Source: U.S. Census Bureau, 2011)

Rank	*Populated Place Name*	*No. of Occurrences*
1	Fairview	218
2	Salem	98
3	Georgetown	87
4	Springfield	65
5	Greenville	55
6	Clinton	51
7	Franklin	52
8	Arlington	47
9	Washington	39
10	Madison	31

Place Name Recognition. Extracting raw geographic place names from a corpus is a relatively trivial process. Using lexicon of place names derived from gazetteer data it possible to transform captured place names into discrete point locations on the ground. Our approach relies on a form of geographical entity extraction to identify nouns and noun phrases based on semantic resources in the form of ontologies and hierarchical geographic networks mapped to ontologies. We have developed a lexicon of all forms toponyms providing a knowledgebase of both native language, foreign language and transliterated forms of geographic names.

Additionally, we have developed a hierarchal data structure that identifies the child-parent relationship of geographic named entities tracing their location in the chain of worldwide administrative divisions. Assigning global and floating location variables for each article, section, paragraph even sentence can assist in resolving ambiguities. These variables would then narrow down the area of the world for which the article is concerned, identifying: World Region, Country and various other levels of administrative division. This allows the identification of the relative location of a named entity within its parent administrative divisions – its "geographic neighborhood". Therefore, in a document about Massachusetts, we can surmise to a certain level of probability that an isolated reference to "Cambridge" will probably refer to "Cambridge, Massachusetts". The concept of "geographic neighborhood" involves resolving ambiguity of a place name based on its hierarchy of parent administrative divisions. By keeping a sequential running tally of those areas/regions/administrative divisions of the world being discussed in a discourse (a "changing window" pointing to a region or area of the world) it is often possible to resolve many ambiguities of vague, implicit and incomplete geographic name references.

One very common technique is to initially search text for place names of especially large or populous places (e.g., country names, big cities) and resolve them immediately. Using this detail to initially load data to "geographic neighborhood" variables, adds context to named entity extraction operations. Another common strategy is to recognize "object/container" pairs of toponyms within the text, so called comma-groups (e.g., "Paris, France"). This works well for text with formal place name entries. When comma-group references are present, they are an ideal means to establish a "geographical neighborhood" of a document. However, comma-groups are usually the exception rather than the norm.

Our approach incorporates initial scans of each document for explicitly named entities followed by "global neighborhood" analysis, and advanced filtering heuristics based on geographic name variants and additional ancillary geographic information. In addition, an extensive data model of common-use spatial modifiers is used to capture the full essence of spatial expression. These spatial modifiers are then used to spatially adjust raw geographic locations into more precise positions on the ground.

Place Name Alignment. Because of inherent ambiguity in verbal discourse, it is often difficult to determine which ontology concept best characterizes the entity referenced in a complex semantic resource. To resolve these ambiguities, our approach incorporates an alignment phase that augments gazetteer data with additional detailed information to assist in geographic names filtering operations. The added value from this ancillary information will resolve most ambiguities. What spatially relevant information is available to assist in the identification of named entities? Besides a worldwide knowledgebase of gazetteer data, we will exploit the following resources during the alignment phase to produce formal lexicons of specialty geographic information:

• Detailed lists of toponyms – English language, native, and foreign language name variants - to include slang, obsolete and current variant forms of geographic names.

• A knowledgebase of legal administrative divisions, their English equivalent names, their hierarchical structure and their use in nations of the world provides a means to isolate the location of incomplete references to geographic named entities. For example, for an isolated reference to "Cambridge", if the surrounding context discusses only the United States this situation can be resolved to mean "Cambridge, Massachusetts" rather than "Cambridge, UK."

• Identifying the parent child relationship of gazetteer data through topological analysis of spatial extent versus point location information. Using spatial footprints of current administrative divisions and urban areas will generate a worldwide hierarchal network of place names relative to their parent administrative division.

• A knowledgebase documenting relationships between various "standardized" country and administrative encoding systems.

• Other key elements of information available for many areas of the world that can assist in resolving ambiguities include population statistics, capital city information and spatial extent.

Geographic Feature Associations. Waldo Tobler proposed the "1st Law of Geography" with the statement, "Everything is related to everything else, but nearby things are more related than distant things" [Tobler, 1970]. This bold statement has been the under-pinning premise of modern geographic information science (GISci) and most forms of spatial analysis. This statement is the foundation for the concept of spatial association. A spatial association is any commonly occurring co-existence between two objects. If you live in the suburbs, merely walking outside the front door will reveal some of these more subtle spatial associations. The existence of the obligatory flowering tree in the front yard is a prime example of an association - one mandated by many municipality building codes. In most modern American subdivisions, each residence will have one. A mailbox – most every residential dwelling will have one. On the edge of the right-of-way notice the utility access box – if there is public water for each residence, there will probably be a water valve in the utility box. These are all examples of commonly occurring feature associations. As in Figure 1, these associations can be portrayed as a directed graph. In isolation these graphs are hierarchal in nature but when combined with other graphs, tend to exhibit small world properties, a situation not yet fully explored. Of importance is that links between nodes describe probabilistic, recurring and quantifiable spatial relationships. These relationships define how objects might be spatially oriented and separated in the real world. By applying corresponding elements of this knowledgebase to vague and implicit named geographic entity information it is possible to predict potentially refined and more accurate geographic locations of features.

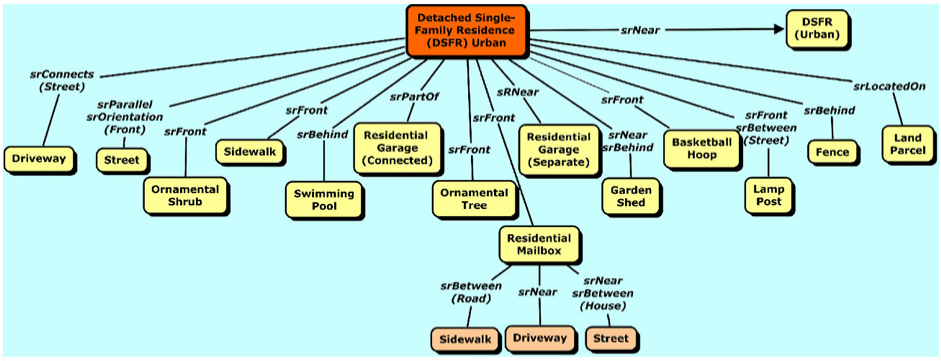

Figure 1. A probabilistic spatial relationship network showing commonly occurring spatial associations and relationships between a Detached Single-Family Residence and various other feature classes.

Geographic Feature Relationships. A primary function of a geographic information system is determining those factors that dictate the location, distance and proximity between features. For example, the distance separating areas of fast-food establishments relative to urban area features is an example of applied spatial relationships. Spatial relationships are attributes defining either absolute and/or relative locations of two or more objects. Spatial relationships can be in the form of distances and proximities between objects, direction of an object from other objects, relative movement of objects, or topological relations of two or more objects (inside, outside, intersecting, etc.). Traditional geographic information science (GISci) is concerned with a limited set of spatial relationship such as the eight basic topologic relationships of the Region Connection Calculus (RCC-8) (Renz, 2003): Equal, Disjoint, Intersects, Touch, Overlap, Cross, Within, and Contains. However, as can be seen in Table 3, there are significantly more spatial

relationships used in common English language discourse than the few traditionally used in GISci. When used in conjunction with spatial associations, it is possible to qualify and quantify the probable location of new feature content based on the location of known features. Formal identification of these "casual" spatial relationships has resulted in the preparation of detailed property definitions and equivalent and synonymous word lists for each.

Table 3. Commonly Used Spatial Relationships.

Spatial Term	*Spatial relationship*
above	Write your name above the line.
across	The house is across the street.
against	She leans against the tree.
ahead of	The truck is ahead of the car.
along	The river bank is along the river.
among	He is standing among the trees.
around	The fence is around the yard.
behind	The shed is behind the house.
below	Write your name below the line.
beneath	He sat beneath the tree.
beside	The girl is standing beside the boy.
between	She is between two trees.
from	He came from the house.
in front of	The mail box is in front of the house.
inside	He is inside the house.
nearby	There is a tree nearby the house.
off	His hat is off.
out of	He came out of the house.
through	She went through the door.
toward	She is walking toward the house.
under	He is hiding under the table.
within	Please mark only within the circle.

An on-going effort has been to define algorithmic approaches for these spatial relationships to allow automated computation of precise locations for new feature content. In this way, after determining that a new feature might exist relative to existing feature content, the position of the probabilistic feature can be computed and added to a GIS database. A detailed discussion and example of this process can be found in (Bitters, 2009). This association/relationship network knowledgebase currently contains only a limited set of feature associations. An on-going effort is underway to expand this set of recurring associations and their relationships to allow future qualification and quantification of probabilistic feature content.

Beyond Basic Named Entity Extraction. Formal topology does not recognize the multitude of relative positioning of objects as we do in verbal or written discourse. It reduces the collection of common-use relationships into a limited set of mathematical situations that can be used to compute common spatial relations encountered in the real world (Renz, 2003). However, in everyday verbal and written discourse a wide variety of spatial predicates and modifiers are used to express the spatial separation of distinct objects. To date our research has identified in excess of 200 potential spatial relationships based on common-use English language usage. If, from verbal discourse, we could extract those terms and expressions commonly used to express spatial relationships – spatial modifiers and predicates; and also identify and extract the geographic object to which they refer, it would then be possible to use appropriate spatial analysis overlay functions to compute adjusted positions of geographical objects and situations.

As an example, from the text - "John and Mary met across from the Post Office", a typical named entity extraction software package would extract the named entities "John and Mary" and "Post Office". Follow-on logic would place "John and Mary" at the derived geographic coordinates of the center of mass of the "Post Office". However, the actual position could in fact be hundreds of meter away from the actual location where "John and Mary" actually met. If the named entity extraction software package were able to recognize the spatial modifier "across from", the computed new position would take into account the offset and the resulting geographic coordinate would be a more realistic geographic position describing this situation.

To illustrate this approach to location refinement, the process shown in Figure 2 assumes the existence of GIS data containing a street network for the region of interest. From the geographic entity extraction process, the coordinates of the "Post Office" have been captured and plotted relative to the street network (the cross at center of mass of the building footprint in Figure 2-a). Using basic spatial analysis functions, it is possible to refine the location of "John and Mary's meeting" to a more precise set of geographic coordinates.

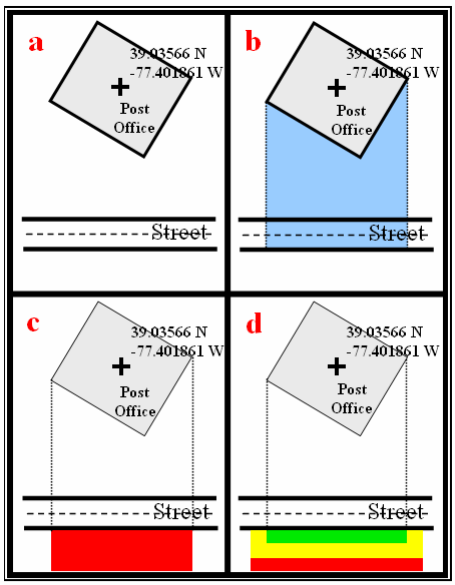

Figure 2. A spatial analysis function for the spatial relationship – Across From.

This is accomplished by first identifying the blue area (Figure 2-b) in front of the post office – that area between the building and the far side of the street. Next, (Figure 2-c) this area is extended beyond the street, perpendicular to the street centerline. Finally, (Figure 2-d) based on a distance-decay algorithm, buffer areas are generated perpendicular to the street and away from the "Post Office" to identify those probable locations of "John and Mary's meeting" - green tint for likely areas, yellow tint for less likely and red tint for least likely. This technique provides more accurate probabilistic geographic location than geo-position derived directly from raw gazetteer data. If other spatial feature association information and their corresponding spatial relations are available it is possible to refine the precision of the geo-location even further.

RESULTS

In this research we have demonstrated an approach to spatial reasoning that has the potential to increase the accuracy and the precision of spatial information extracted from text documents. Results of preliminary testing indicate:

- Gazetteer data augmented with an expanded set of toponyms adds efficiency and increases accuracy when performing named geographic entity extraction processes.
- A lexicon of spatial predicates and equivalent terms and expressions is an effective means to capture spatial modifiers in a named entity extraction environment.
- An expanded set of spatial relationship functions, spatial relationship functions based on common use terms and expression, add precision to derived coordinate values.
- Initial scanning of text documents for explicit named geographic entities provides a basis for detailed named geographic entity extraction.

CONCLUSION AND FUTURE PLANS

Additional testing and refinement of named entity extraction tools are necessary to refine and enhance derived coordinates. Additional spatial relationship functions are necessary to include the subtle spatial nuance of common use expressions and terms. An expanded set of equivalent and synonymous terms, both spatial relationships and toponyms, is essential to capture the fine distinctions in meaning used in everyday verbal communication.

REFERENCES

Amitay, E., N. Har'El, R. Sivan, and A. Soffer. 2004. Web-a-Where: Geotagging web content. In Proc. of SIGIR, pages 273–280, Sheffield, UK.

Bell, A. 1991. The Language of News Media. Wiley-Blackwell, New York

Bitters, B. 2009. Spatial relationship networks: Network theory applied to GIS data. Cartography and Geographic Information Science. 36(1):81-93.

Kleinberg, J. 2002. Bursty and hierarchical structure in streams. ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD).

Leidner, J. L. 2007. Toponym Resolution in Text: Annotation, Evaluation and Applications of Spatial Grounding of Place Names. PhD thesis, University of Edinburgh, Edinburgh, Scotland.

Li, Y. 2007. Probabilistic toponym resolution and geographic indexing and querying. Master's thesis, University of Melbourne, Melbourne, Australia.

Lieberman, M.D., H. Samet, and J. Sankaranayananan. 2010. Geotagging with local lexicons to build indexes for textually-specified spatial data. In Proc. of ICDE, Long Beach, CA.

Martins, B. 2008. Geographically Aware Web Text Mining. PhD thesis, University of Lisbon, Lisbon, Portugal.

McCallum, A. 2005. Extraction: Distilling Structured Data from Unstructured Text. ACM Queue 3(9):48-57.

McCallum, A., and Jensen, D. 2003. A note on the unification of information extraction and data mining using conditional-probability, relational models. IJCAI Workshop on Learning Statistical Models from Relational Data.

Overell, S.E. 2009. Geographic Information Retrieval: Classification, Disambiguation and Modelling. PhD thesis, Imperial College London, London.

Rauch, E., M. Bukatin, and K. Baker. 2003. A confidence-based framework for disambiguating geographic terms. In Proc. of HLT-NAACL, pages 50–54, Edmonton, Canada.

Renz, J. 2002. Qualitative Spatial Reasoning with Topological Information. Lecture Notes in Computer Science 2293, Springer Verlag, Heidelberg.

Teitler, B. E., M. D. Lieberman, D. Panozzo, J. Sankaranarayanan, H. Samet, and J. Sperling. 2008. NewsStand: A new view on news. In Proc. of ACM GIS, pages 144–153, Irvine, CA.

Tobler, W. 1970. 1970. A computer movie simulating urban growth in the Detroit region. Economic Geography 46:234–40.

U.S Census Bureau. 2011. American FactFinder. United States Census Bureau web site. http://factfinder.census.gov/home/saff/main.html. Last retrieved 2011-01-25.