Towards ontology-based disambiguation of geographical identifiers

Raphael Volz
FZI Research Center for Information Technologies
Karlsruhe, Germany
volz@fzi.de

Joachim Kleb
FZI Research Center for Information Technologies
Karlsruhe, Germany
kleb@fzi.de

Wolfgang Mueller
FZI Research Center for Information Technologies
Karlsruhe, Germany
wmueller@fzi.de

ABSTRACT

Geographic names have always been important identifiers. People typically use names and not coordinates to identify geographic features. Therefore to establish identity beyond coordinates, name disambiguation is required to identify the exact geographic feature that is denoted by a name. This paper introduces an ontology-based approach to disambiguate geographical names in texts. The ontology defines the central conceptual basis of our approach and is used to rank geographic features based on disambiguation rules that take into account structural information contained in the ontology (e.g. population of a town), as well as textual indicators contained in the text at hand. Our first evaluation on a subset of the well-known Reuters 21578 corpus indicates promising results both in terms of precision and recall.

Categories and Subject Descriptors

H.1.m [**Miscellaneous**]: ontology; H.3.3 [**INFORMATION STORAGE AND RETRIEVAL**]: Information Search and Retrieval: selection process; I.7.m [**DOCUMENT AND TEXT PROCESSING**]: Miscellaneous: Document Assignment

General Terms

Algorithms, Languages , Performance, Design

Keywords

ontology, semantics, geographic references, disambiguation, candidate identification

1. INTRODUCTION

Geography has become a popular theme on the Web. Applications like Google Maps are easy to use and incentivise people to publish data tagged with geographical identifiers, i.e. coordinates in form of longitude and latitude, to visualize their data in maps and other representations of geography.

When talking about geographic identifiers, however, people use geographic names to denote a certain coordinate and look up information pertaining to this coordinate using names. Geographic names, like all names, are often highly ambiguous. For example, the name *San Jose* refers to 1724 different coordinates in the two largest publicly available database of geographic names GeoNet and GNIS (cf. Figure 1).

WWW2007, May 8–12, 2007, Banff, Canada.

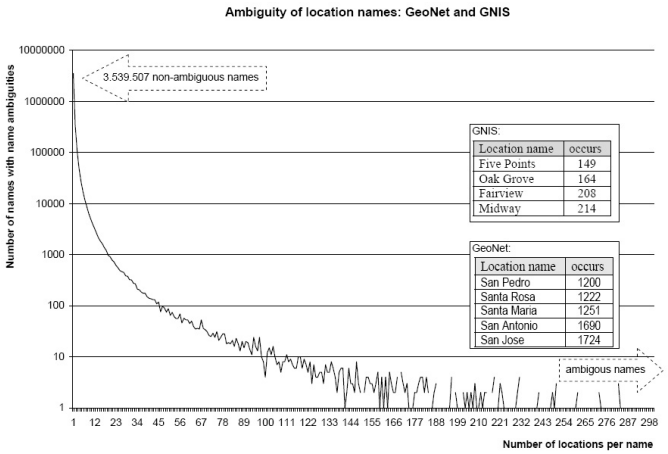

Figure 1: Ambiguity of geographic names in GNS and GNIS

Figure 1 shows the disambiguation of geographic names is a concrete and relevant task that is needed *(I)* to attach the appropriate identifier to data containing a geographic name *(II)* to resolve the appropriate identifier from a geographic name when users retrieve information.

Our paper presents a novel method to disambiguate geographical names based on an ontology. The ontology incorporates data from publicly available geographic gazetteers - the Geonames databases GNS and GNIS (cf. section 2.1) - as well as common world / linguistic knowledge that has been obtained from WordNet. The ontology is used as a gazetteer to map references to (multiple ambiguous) geographic feature candidates from text.

Our disambiguation approach chooses among those feature candidates by dealing with three types of ambiguity: (I) *multi-referent* ambiguity, when two different geographic locations share the same name; (II) *name variant* ambiguity, when the same location has different names; and (III) *geoname-non geoname* ambiguity, where a location name could also stand for some other word such as a person name or nouns, e.g. *Metro* as the city in Indonesia vs. *Metro* as the subway system.

Our approach establishes a ranking among feature candidates recognized in a text. To this extent, we are leveraging further information about the geographical name in the

given document as well as scoring rules that prefer certain concepts in the ontology over others, e.g. geographic names for cities dominate geographic names for forests or lakes. Thirdly, data available from the ontology, particularly the population of cities is used to give preference to certain candidates to prefer large towns over small towns. These rules reflect natural heuristics as employed by humans. For example, "Paris" - without further information - more likely refers to the capital of France rather than the small town Paris in Texas.

The paper is organized as follows: Section 2 describes the ontology as well as the data sets used to build the ontology. Section 3 describes our approach of disambiguating geographic names using the ontology. Section 4 describes our evaluation corpus and discusses the results of our evaluation. We conclude with a discussion of related work in section 5 followed by a summary of our contribution and an illustration of next steps in Section 6.

2. GEONAMES ONTOLOGY

2.1 Geographic data sources

Our ontology is derived from two publicly available data sources for geographic names:

- The GEOnet Names Server (GNS) [3] provides a database of non-US geographic feature names and is maintained by two US government authorities (US National Geospatial-Intelligence Agency (US NGA) and US Board on Geographic Names (US BGN)). GNS provides approximately 5.5 million names for roughly 4.0 million geographic features outside the US.

- The Geographic Names Information System (GNIS) maintained by US Board of Geographic Names complements GNS with US-specific names. The database holds the officially recognized name of each feature and defines the feature location by state, county, US Geological Survey (USGS) topographic map and geographic coordinates. Other attributes include alternate names and alternative spellings for the official name, feature designations, historical and descriptive information and - for some features - geometric boundaries.

These data sets have been frequently used as gazetteers by the NLP community. Besides providing a list of names, the data contains several additional information items (that can be used for disambiguation). All names are classified in feature types (e.g. city, park, forest, ...) and information about the containing (political) administrative regions (e.g. county, state and country) is provided. For unique identification, coordinate information stored with the name can be used.

Figure 1 visualizes the place name disambiguation distribution characteristic for the GeoNet- and GNIS databases. While a geographic name has in average 4.4 different meanings, some names derived from Spanish saints refer to more than 1000 different locations. Table 1 summarizes the size of both data sets in terms of populated places (without location name variants) and the full names of locations in both data sets.

WordNet	absolute	relative
nouns	114.648	100%
ambiguous with geo names	13.169	11,5%
not ambiguous with geo names	101.479	88,5%

Table 2: Geo/non-geo ambiguity in WordNet and Geo name databases

2.2 WordNet

WordNet is a well known lexical resource about English words, which is often used in natural-language processing and information retrieval applications. The core concept in WordNet is the synset. A synset groups words with synonymous meaning. Ambiguity of words is represented by the mapping a Word to multiple Synsets. We are using WordNet 2.0 which contains 114.648 nouns[1]. 11,5% of those nouns intersect with the geographic names in GNIS and GeoNet (cf. Table 2) and constitute Geo-Non Geo ambiguities.

2.3 Ontology model

Our ontology is created from the above data sources and is represented in the well-known OWL format. We currently only use a subset of the axioms and class descriptions possible in OWL. For our purpose the features already present in RDF Schema suffice. Our ontology uses the following structures:

DEFINITION 1. *A* ontology *is a structure*

$$\mathcal{O} := (C, \leq_C, R, \sigma)$$

consisting of

- *two disjoint sets C and R whose elements are called* classes *and* properties*, resp.,*
- *a partial order $\leq_C$ on C, called* class hierarchy *or* taxonomy*,*
- *a function $\sigma\colon R \to C \times C$ called* signature *of a property*

All classes are serialized as `owl:Class`, while all properties are serialized as `owl:ObjectProperty` or `owl:DatatypeProperty`.

DEFINITION 2. *For a property $r \in R$, we define its* domain *and its range by* $\mathrm{dom}(r) := \pi_1(\sigma(r))$ *and* $\mathrm{range}(r) := \pi_2(\sigma(r))$.

If $c_1 \leq_C c_2$, for $c_1, c_2 \in C$, then c_1 is a subclass of c_2 *and c_2 is a* superclass of c_1.

If $c_1 <_C c_2$ and there is no $c_3 \in C$ with $c_1 <_C c_3 <_C c_2$, then c_1 is a direct subclass of c_2 *and c_2 is a* direct superclass of c_1. *We note this by $c_1 \prec c_2$.*

DEFINITION 3. *A* knowledge base *is a structure*

$$KB := (C_{KB}, R_{KB}, I, \iota_C, \iota_R)$$

consisting of

- *two sets C_{KB} and R_{KB},*
- *a set I whose elements are called* instances *(or* identifiers *typically denoted by URI's),*

[1][11] presents further WordNet 2.0 statistics

Database	number of names	ambiguous	relative	non-ambigous	relative	mean ambiguity
GNIS	177.983	95.435	54%	82.548	46%	4,8
GEOnet	5.808.675	2.351.716	40,5%	3.456.959	59,5%	4,0
Sum:	5.986.658	2.447.151	40,4%	3.539.507	59,6%	4,4

Table 1: Ambiguity in data sources GNIS and GEOnet

- *a function* $\iota_C \colon C_{KB} \to \mathfrak{P}(I)$ *called* class instantiation*, where* $\mathfrak{P}$ *denotes the powerset of the set of instances.*
- *a function* $\iota_R \colon R_{KB} \to \mathfrak{P}(I \times I)$ *called* property instantiation.

The knowledge base is stored as RDF triples, any element of ι_C is exported as `(instance,rdf:type,class)`. Elements of ι_R are exported as triples `(instance,property,instance)`

Definition 4. *A* lexicon *for an ontology* $\mathcal{O} := (C, \leq_C, R, \sigma)$ *is a structure*

$$Lex := (S_C, S_R, Ref_C, Ref_R)$$

consisting of

- *two sets* S_C *and* S_R *whose elements are called* names for classes *and* properties*, resp.,*
- *a property* $Ref_C \subseteq S_C \times C$ *called* lexical reference for classes*, where* $(c, c) \in Ref_C$ *holds for all* $c \in C \cap S_C$*.*
- *a property* $Ref_R \subseteq S_R \times R$ *called* lexical reference for properties*, where* $(r, r) \in Ref_R$ *holds for all* $r \in R \cap S_R$*.*

Based on Ref_C*, we define, for* $s \in S_C$*,*

$$Ref_C(s) := \{c \in C \mid (s, c) \in Ref_C\}$$

and, for $c \in C$*,*

$$Ref_C^{-1}(c) := \{s \in S \mid (s, c) \in Ref_C\} \ .$$

Ref_R *and* Ref_R^{-1} *are defined analogously.*

An ontology with lexicon *is a pair*

$$(\mathcal{O}, Lex)$$

where $\mathcal{O}$ *is an ontology and Lex is a lexicon for* $\mathcal{O}$*.*

Definition 5. *An* instance lexicon *for a knowledge base* $KB := (C_{KB}, R_{KB}, I, \iota_C, \iota_R)$ *is a pair*

$$IL := (S_I, R_I)$$

consisting of

- *a set* S_I *whose elements are called* names for instances*,*
- *a property* $R_I \subseteq S_I \times I$ *called* lexical reference for instances*.*

A knowledge base with lexicon *is a pair*

$$(KB, IL)$$

where KB is a knowledge base and IL is an instance lexicon for KB.

The lexicon is serialized as RDF triples using `rdf:label` as the predicate in the triple.

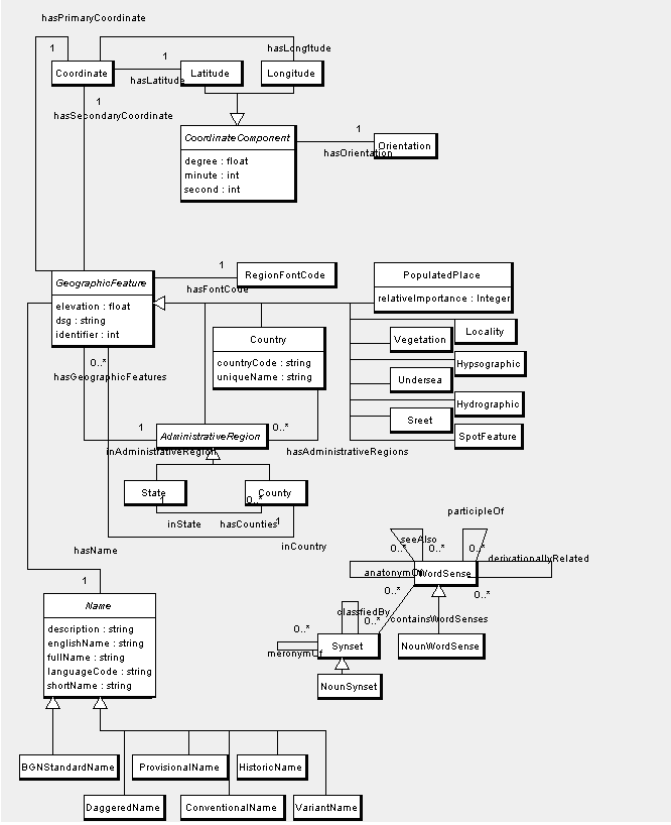

Figure 2: Geoname Ontology (UML-Presentation)

2.4 Geoname Ontology

Figure 2 visualizes the top level class hierarchy of our geonames ontology. The class *Geographic Feature* is the central class of the ontology. This class has several subclasses to differentiate between the different forms of geographic features that are found in GNIS and GeoNet, e.g. *PopulatedPlace* is the superclass of all country capitals, cities and villages. Another example is the class *Hydrographic*[2], which is used to represent all water-related geographic features. Every geographic feature is associated with a coordinate that is decomposed in Latitude and Longitude. Geographic features can be associated with political *Administrative Regions*, in particular its subclass *Country*. Besides geographic information the ontology also contains four classes to represent the content of WordNet, i.e. *Synset*, *WordSense*, *NounWordSense* and *NounSynset*. This representation of WordNet follows the official W3C proposal[3].

[2]Which itself has further subclasses such as *Sea, Stream* or *River* (not visible in Figure 2)

[3]`http://www.w3.org/TR/2006/WD-WordNet-rdf-20060619/`

The knowledge base of the ontology is created by iterating over all WordNet noun word senses and populating the *NounWordSense* and *NounSynset* classes with the exception of word senses of hyponyms of the synset "city" and "country"[4]. Additionally we iterate over all GeoNet and GNIS entries, while instating the appropriate subclass of *GeographicFeature* that is indicated by the feature classification of the entry at hand. At the same time properties are established to the related classes such as *Country*, *Coordinate* etc.

During the iterations to establish the knowledge base, the (instance) lexicon of the ontology *IL* and *Lex* is created from all names (and name variants) in the geographic databases as well as all nouns in WordNet. Additionally, we have populated the lexicon with names that reference classes, e.g. the word "Sea" establishes a reference to the class "Sea", as well as names that reference instances, e.g. the ISO codes for countries reference the respective instance that represents the country.

3. GEO-NAME DISAMBIGUATION

Our approach for disambiguation establishes a ranking among feature candidates recognized in text. We are leveraging information about the geographical name in the given text and use the ontology as a gazetteer for instance identification. Additionally we utilize scoring rules that prefer certain concepts in the ontology over others, e.g. geographic names for cities dominate geographic names for forests or lakes. The scoring rules reflect natural heuristics as employed by humans and are based on definitions available in the ontology.

The disambiguation process involves several steps:

1. Spotting candidates for geographical identifiers in text utilizing the ontology as a gazetteer
2. Narrowing candidates using surrounding textual information (textual disambiguation)
3. Ranking candidates using ontological information (ontology-based disambiguation)

3.1 Spotting candidates

To spot candidates, we first transform a given text into a bag of word representation. Therefor a document is represented as a vector D consisting of the terms $(t_1, \ldots, t_n)$ of the document. The original ordering of the terms within the document is preserved. In the formation of D we rely on the built-in gazetteers of the NLP tool, which offer means to spot person names, organization names and stop words[5]. Such elements spotted by the NLP tool are not part of D. Additionally, we have defined several grammars that suppress the recognition of consecutive terms. For example, the recognition of a person's first name through the NLP tool will automatically consider the following term as a last name and not add both terms to D.

The ontology is used as a gazetteer utilizing the instance lexicon *IL* and obtaining references to candidate instances $i \in I$ using the relation R_I. We obtain the set of candidate geographic identifiers in the document D using a function $cand(t_i) := \{i \in I | (t_i, i) \in R_I\}$[6]. In a similar fashion we identify a set of concepts by utilizing the *signs for concepts* S_C in the lexicon *Lex*: $con(t_i) := \{c \in C | (t_i, c) \in Ref_C\}$. Note, that $cand(t_i)$ will also include references to the non-geographic names which have been incorporated from WordNet.

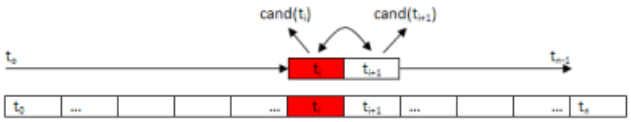

Figure 3: Traversal of term vector to identify relations between candidates.

3.2 Narrowing candidates

The second step is textual disambiguation where textual patterns allow to narrow candidates of ambiguous geographic names. We first traverse the term vector D with a window of two consecutive terms $(t_i, t_i + 1)$ (cf. Figure 3). If candidates for t_i refer to instances of *geographic feature* and candidates for $t_i + 1$ refer to instances of *administrative regions* or vice versa, the combination of t_i and t_{i+1} is used to narrow the set of candidates for t_i to those geographic features that are located in the given administrative region t_{i+1}. For example, the text "Paris, France" will thereby lead to the selection of the French capital.

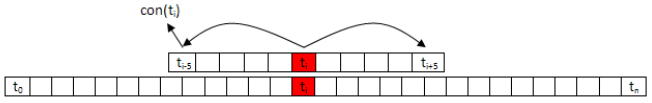

Figure 4: Traversal of term vector to identify relations between concepts and candidates.

In a next step we traverse the term vector D for a second time with a larger window of 11 consecutive terms $(t_{i-5}, \ldots, t_i, \ldots, t_{i+5})$ (cf. Figure 4) and narrow candidates based on geographic feature classes. If candidates for a given term t_i are instances of any class referenced by $con(t_j)$, where $i - 5 \leq j \neq i \leq i + 5$, the set of candidates is narrowed to the instances of these classes. For example, the text 'Notthingham forest' will lead to the selection of the forest in the UK over the city of Notthingham.

3.3 Ranking candidates

The remaining candidate sets are then ranked utilizing weights that are attached to the classes of the ontology. These weights are propagated to the instances in the candidate set and used to rank those instances. The instance with the highest rank is chosen. If the weight of an instance is negative, we disregard the term as geographic name.

Table 3 shows the weights attached to concepts as used in our evaluation:

[4]Otherwise we would duplicate this information in the ontology

[5]We utilize the well-known stop word list of the SMART project `ftp://ftp.cs.cornell.edu/pub/smart/english.stop`

[6]This is a simplified definition. References to instances constituted by multiple consecutive nouns in the text are supported by combining those nouns.

Concept	Weight
Country	+3000
Populated Place	+3000
Administrative Region	+1000
Locality	+1000
Hydrographic	+10
Hypsographic	+10
Spot Feature	+10
Vegetation	+10
Street	+5
Undersea	-10
WordSense	-10000

Table 3: Class Weights

Class weights from table 3 are transitively propagated to their subclasses via $\leq_C$. Class weights are then propagated to class instances via the instantiation function ι_c. For all instances of the class *PopulatedPlace* the number of inhabitants (divided by thousand) is added to the weight of the instance.

Note, that we use negative weights for the WordNet classes. This ranks names from WordNet down and at the same time disregards all terms that have only a WordNet meaning. Finally, we select the instance $i \in cand(t_i)$ with the maximum score (or randomly choose an instance if there are several instances with the equal maximum score).

For example, the term Lancaster (without any further information such as county or country) would yield a score of 3129 for Lancaster (CA) (a populated place with 129 tsd. inhabitants in California) and a score of 3047 for Lancaster (UK) (the well-known city in the UK with 47 tsd. inhabitants), while all Lancaster counties in the US would only get a score of 1000, and the Lancaster trough in the Lancaster sound in the north of Canada receives -10 points being a subclass of undersea. Hence, the algorithm chooses Lancaster, CA as the instance that provides the most likely geo reference.

To additionally allow to determine a page focus, i.e. choose one geographic reference among all geographic references in the text to describe the main geographic focus of the text, we simply count term occurrences and multiply the weight of the instances with the number of occurrences in the text. We then choose the instance with the maximum weight among all candidate sets.

4. EVALUATION

While Leidner [4] describes a reference corpus of toponym resolution, the data set is not publicly available. We therefore compiled two corpora from the well-known Reuters 21578 data set to evaluate our approach to disambiguation of geographic names:

- Corpus 1: 250 documents hand-annotated with *all* geo-references for geo-name disambiguation
- Corpus 2 : 100 documents hand-annotated with *one* country reference identifying page focus

We evaluate our approach on both corpora and consider a result as correct, if it corresponds exactly to the user's input.

Test Run	I	II	III
Precision	40,1%	68,9%	67,9%
Recall	86,7%	86,7%	86,7%
Weights			
WordNet senses	0	-20000	-10000
Administrative Region	1000	1000	1000
Country	3000	3000	3000
Hypsographic	10	1	10
Locality	1000	500	1000
PopulatedPlace	3000	10000	3000
Road	5	10	5
Spotplace	1000	2500	1000
Hydrographic	10	10	10
Undersea	-10	1	-10
Vegetation	10	20	10

Table 4: Results of evaluation 1 - precision, recall and class weights used

4.1 Evaluation 1 - geo-name disambiguation

We evaluate the success of our approach using the classical precision and recall measures.

- precision = $\frac{|G_{res} \cap G_{rel}|}{|G_{res}|}$
- recall = $\frac{|G_{res} \cap G_{rel}|}{|G_{rel}|}$

Here, G_{res} is the set of geographic references identified by the algorithm and G_{rel} is the set of relevant geographic references of the corpus as identified by the annotator. We have carried out three test runs with different weights.

The first run did not weight WordNet terms negatively and therefore leads to a low precision of 40%. This is due to the fact that many terms are incorrectly considered as geographic names. These *false positives* are constituted by common words, such as the aforementioned example 'Metro'. The recall is at an acceptable level of 86,7%. Despite the use of such a large gazetteer as the one constituted by our ontology, annotators where (semantically) correctly tagging several organizations such as the 'Bank of England' as well as tagging adjectives such as 'French' with a geo-reference to the country, while this information is not utilized in our approach.

In the second run, we weight all WordNet terms with a high negative score. Similarly, we choose to increase the weight of populated places, as cities are very common in the corpus. This improves the precision tremendously to almost 70%.

The third run tests the sensitivity of precision to the weight of populated places and WordNet WordSenses and indeed the precision goes back slightly when decreasing the weights of those two classes. Unfortunately, we were not able to improve the precision above the level of test run II in several iterations with other weights.

Looking at the results, we could observe that the identification of country names has much greater precision due to absence of ambiguities. Here, only misspellings and other noise in the data prevent 100% precision.

4.2 Evaluation 2 - Page focus

Inspired by the good results for country name assignment, we were curious about the results for page focus assignment

limited to country names. While we can achieve 100% precision our recall is limited to 90,9%. Hence, our simple approach to count the occurances of country names already leads a very high recall, only in seldom cases the human annotators chose to identify countries with lower frequency as the page focus of the annotated texts.

5. RELATED WORK

Ontologies have been used in geographic information system to merge data from different sources [2, 6]. However, when looking at geographic information retrieval, the use of ontologies is novel. The geographic information retrieval community has addressed the disambiguation of geographic names in several papers where tailored algorithms incorporating heuristics to choose among candidates are described:

[1] presents the Web-A-Where System, which uses a custom gazetteer to identify geographic names which are likely to have a non-geographic meaning. Geographic names are identified with a gazetteer that is compiled from cities with more than 50.000 inhabitants in the GNIS and GeoNet database and therefore is by far not as extensive as our data set, which includes all cities and also other geographic features. Simple heuristics which assign confidence values to candidates are used. For example, a default heuristic is that a higher population increases confidence values for candidates. Confidence values are also increased if other text references qualify administrative regions (such as countries or states). If multiple place names occur in one text, the confidence of candidate locations that share the same administrative region are increased. The proposed page focus strategy selects up to four place names that cover most of place names in a page.

The Perseus project [10] disambiguates ambiguous location names by a series of heuristics based on the qualifiers in the vicinity (e.g. state name immediately following the city name), nearby disambiguated place names and general world knowledge.

[8] describes the NewsExplorer application, which automatically builds up knowledge from newspaper articles in 13 languages and identifies places and other named entities. They employ an disambiguation process that distinguishes place names from known person names by consideration of place importance and by estimation of main countries based of the minimum kilo metric distance to other places mentioned in the text.

The disambiguation method described in [7] introduces several steps for disambiguating *wiki-pages*. It results in a so called *Disambiguation Pipeline*, which consists of disambiguation steps based on *templates*, *categories*, *referents* and *text heuristics*. The *templates* in Wikipedia are initially used for disambiguating documents, relating them to classes like *Country* or *Bibliography* etc. The disambiguation by *category* offers the possibility to denote associations between documents. Category tags can identify the country or continent of an article or indicate an article not referring to a place, while the referent disambiguation offers the possibility to express relations between named entities (e.g. between places and parent places: describing a town, mentioning county or country). Disambiguation with *Text Heuristics* offers options to apply concrete rules between entities e.g. describing an important related place. Only places of equal or greater importance are used as referrers.

MetaCarta [9] is a commercial system that uses NLP patterns, capitalization convention, place names found in vicinity, human population and other heuristics to disambiguate place names. To query web pages using a place name, they are scored by a function combining confidence values, positions and prominence of the place name in the web pages.

The disambiguation approach of [12] is rule-based and makes use of both contextual information extracted from the web pages as well as spatial distances between place names. Several rules analyze contextual information to choose the appropriate location candidate, e.g. if the administrative region is found in the context. If no exact place name can be assigned spatial distance to other recognized place names in the text are used to rank those candidates higher, which have the least difference. Similar place name disambiguation methods have been adopted in other research efforts [5],[6].

We have captured many of the ideas for disambiguation heuristics presented in the above-mentioned papers in our disambiguation rules. Our use of ontologies generally increases the flexibility and allows to easily incorporate further knowledge into the disambiguation process. Similarly, it allows to flexibly change disambiguation rules that are defined on top of the ontology, e.g. to provide the recognition of person names.

6. CONCLUSION

6.1 Summary

We have presented a novel approach for disambiguation of geographical names based on ontologies, which allow us to formulate ranking rules based on concepts and utilize the lexicalization of the ontology as a gazetteer. The geonames ontology constructed for our purpose of disambiguation is used as a gazetteer to map textual references to instances and classes in the ontology. Disambiguation rules in form of rankings are based on attaching weights to concepts and propagating weights to the instances of concepts. Data from the ontology can be utilized for the ranking as well (such as the population of towns). The ontology also finally establishes unique references for geographic identifiers in order to be reused in other applications in form of URI's. This provides a basis for both data integration as well as natural extension points of the ontology.

6.2 Outlook

Going forward we see several next steps: First, we want to evaluate our approach large scale by disambiguating geographic references in arbitrary RSS feeds and utilizing this information to filter news articles by regions referenced. Technically, RSS feeds will be tagged by geographic identifiers.

Second, as part of this application, we will incorporate means for users to provide feedback on the correctness of the disambiguation. This information will be used to improve our algorithms using statistical machine learning techniques. We will also use the ontology beyond its current gazetteer function and display information items related to ontology instances (such as country and state located in, neighboring geographic features, facts such as population, altitude, alternate spellings in different languages, etc.). We also plan to utilize the extensibility of the ontology by incorporating further information sources such as the dbpedia[7] project, which has created an ontology for facts from Wikipedia. The

[7] `http://dbpedia.org/`

currently used expressiveness power of the ontology will be extended by considering more than the currently used object property relations between concept nodes for determination of the correct meaning of the geographic identifiers.

Thirdly, in the process of spotting the meaningful candidate terms for geographic references, the sliding window algorithm will be further improved. Additional to the estimation of an accurate window size, we also consider the use of further text patterns to improve the introduced algorithm.

Concerning the suggested ranking algorithm, we plan a dynamically adjustment of the ontology weight values. In a next step we are going to an iteratively refinement of the initial measure through consideration of already estimated documents over the corpus. We also plan to improve our page focus algorithm by consideration of spatial distances between the identified geographic entities.

Beyond this application, we will look for further disambiguation objectives. Many cases of incorrect assignment of geographic names are due to organization names such as "Texas Instruments". We will again try to tackle this issue by extending the ontology, using it as a gazetteer and defining appropriate heuristics. We assume this will also allow us to improve the page focus task in many cases, as our annotators have assigned page focus based on organization information, e.g. "Bank of England" → "UK" or "OECD" → "Europe".

Acknowledgments. We thank our colleague Frank Kleiner for his technical support and our students for the annotation support.

7. REFERENCES

[1] E. Amitay, N. Har'el, R. Sivan, and A. Soffer. Web-a-where: geotagging web content. In *SIGIR '04: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 273–280, New York, NY, USA, 2004. ACM Press.

[2] P. Clough, M. Sanderson, and H. Joho. Extraction of semantic annotations from textual web pages. Deliverable D15 6201, EU Project: SPIRIT IST-2001-35047, April 2004.

[3] GEOnet. GEOnet Names Server. `http://earth-info.nga.mil/gns/html/`.

[4] J. L. Leidner. Towards a reference corpus for automatic toponym resolution evaluation. In *Proceedings of the Workshop on Geographic Information Retrieval held at the 27th Annual International ACM SIGIR Conference (SIGIR 2004)*, Sheffield, UK, 2004.

[5] H. Li, R. K. Srihari, C. Niu, and W. Li. Location normalization for information extraction. In *Proceedings of the 19th international conference on Computational linguistics*, pages 1–7, Morristown, NJ, USA, 2002. Association for Computational Linguistics.

[6] D. Manov, A. Kiryakov, B. Popov, K. Bontcheva, and D. Maynard. Experiments with geographic knowledge for information extraction. In *Workshop on Analysis of Geographic References, HLT/NAACL'03*, Edmonton, Canada, 2003.

[7] S. Overell, J. Magalhães, and S. Rüger. Place disambiguation with co-occurrence models. In A. Nardi, C. Peters, and J. L. Vicedo, editors, *CLEF 2006 Workshop, Working notes*, September 2006.

[8] B. Pouliquen, M. Kimler, R. Steinberger, C. Ignat, Tamara, Oellinger, K. Blackler, F. Fuart, W. Zaghouani, A. Widiger, A.-C. Forslund, and C. Best. Geocoding multilingual texts: Recognition, disambiguation and visualisation. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC-2006)*, May 2006.

[9] 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, Morristown, NJ, USA, 2003. Association for Computational Linguistics.

[10] D. A. Smith and G. Crane. Disambiguating geographic names in a historical digital library. In *ECDL '01: Proceedings of the 5th European Conference on Research and Advanced Technology for Digital Libraries*, pages 127–136, London, UK, 2001. Springer-Verlag.

[11] J. Yu, Z. Wen, Y. Liu, and Z. Jin. Statistical Overview of WordNet from 1.6 to 2.0. In *2nd Global WordNet conference*, pages 352–357, 2004.

[12] W. Zong, D. Wu, A. Sun, E.-P. Lim, and D. H.-L. Goh. On assigning place names to geography related web pages. In *JCDL '05: Proceedings of the 5th ACM/IEEE-CS joint conference on Digital libraries*, pages 354–362, New York, NY, USA, 2005. ACM Press.