Discovering Location Indicators of Toponyms from News to Improve Gazetteer-Based Geo-Referencing

Cleber Gouvêa[1], Stanley Loh[1,2], Luís Fernando Fortes Garcia[2,3], Evandro Brasil da Fonseca[1], Igor Wendt[1]

[1]Centro Politécnico – Universidade Católica de Pelotas (UCPEL)
96010-000 – Pelotas – RS – Brasil

[2]Centro de Tecnologia e Computação – Universidade Luterana do Brasil (ULBRA)
92425-900 – Canoas – RS – Brazil

[3]Curso de Informática – Faculdade Dom Bosco
Porto Alegre – RS – Brasil

cleber AT sindiq.com.br, sloh AT terra.com.br, luis AT garcia.pro.br,
{evandrofonseca08, igor.wendt} AT gmail.com

Abstract. *This paper presents an approach that identifies Location Indicators related to geographical locations, by analyzing texts of news published in the Web. The goal is to semi-automatically create Gazetteers with the identified relations and then perform geo-referencing of news. Location Indicators include non-geographical entities that are dynamic and may change along the time. The use of news published in the Web is a useful way to discover Location Indicators, covering a great number of locations and maintaining detailed information about each location. Different training news corpora are compared for the creation of Gazetteers and evaluated by their ability to correctly identify cities in texts of news.*

1 Introduction

Geo-referencing of texts, that is, the identification of the geographical context of texts is becoming popular in the web due to the high demand for geographical information [Sanderson and Kohler 2004] and due to the raising of services for query and retrieval like Google Earth (geobrowsers). The main challenge is to relate texts to geographical locations. However, some ambiguities may arise [Clough et al. 2004]:

- *Reference Ambiguity*: the same location may be referenced by many names;
- *Referent Ambiguity*: the same term may be used to reference different locations (for example, two cities with the same name);
- *Referent Class Ambiguity:* the same term may be used to reference different kinds of locations (for example, a street and a city with the same name).

For solving ambiguity problems, one of the alternatives is to utilize Word Sense Disambiguation techniques [Li et al. 2003] where co-occurrence of terms or collocations are identified in training texts. These associations are stored in structures called Gazetteers, that relate locations and references such as names of geographical entities,

the kind of location (city, street, state, etc.), synonyms and also geographical coordinates [Hill et al. 2000].

Although the existence of predefined Gazetteers like Geonames [Geonames 2008] and Getty Thesaurus of Geographic Names [Tgn 2008] , they fail in coverage, lacking information about some countries, and they also fail by weak specialization, lacking detailed references to locations (fine granularity) as for example names of streets, squares, monuments, rivers, neighborhoods, etc [Leveling et al. 2006]. This last kind acts as indirect references to geographical locations and is important because texts about locations frequently utilize this kind of information instead of the explicit name of the location (for example, textual news in the Web). [Leveling and Hartrumpf 2007] call this kind of information as "Location Indicators" and identified some types:

- Adjectives: Rio de Janeiro ➔ "Wonderful city";
- Synonyms: Rio de Janeiro ➔ Rio;
- Codes and acronyms: BSB as the acronym for the airport of the city of Brasília;
- Idiom variations: São Paulo and San Pablo;
- Other geographical entities: names of authorities, highways, squares, airports, etc.

Location Indicators include non-geographical entities, like very important people related to the location, historical events or even temporary situations. The problem is that most of these indirect references are dynamic and may change along the time, and for this reason they do not appear in traditional Gazetteers, because pre-defined Gazetteers are manually created and maintained by people. According to [Delboni et al. 2007], the quality of Gazetteers depend on regular updates (global Gazetteers suffer about 20 thousand modifications per mouth [Leidner, 2004]). For this reason, these Gazetteers do not cover a great number of locations neither have much specific information.

The current work proposes the automatic creation of Gazetteers as a way to cover a great number of locations and for maintaining detailed information about each location. The idea is to utilize news published in the Web to generate and maintain a Gazetteer with detailed information, including indirect references (Location Indicators). Although the existence of works that utilize automatic techniques for supervised learning, these works usually demand manual annotation of the training corpus and are applicable only in specific idioms.

The approach proposed in this paper extracts Location Indicators from news, based on co-occurrence of proper names, without manual annotation and for whatever location and language (since it is possible to understand which terms represent proper names and since there are news about the location). The main focus in this paper is the application of the approach for referent disambiguation (cities with the same name) and for geo-referencing of texts where the name of the location is not present (*indirect reference ambiguity*, as defined in this work). To do that, the work utilized different corpora for identification of Location Indicators to be used in Gazetteers. The different corpora were tested and compared among them and against a baseline Gazetteer created with names of streets and neighborhoods for all Brazilian cities.

The paper is structured as follows. Section 2 discusses related work and defines the problem that is focus of this work, section 3 presents the different methods tested for

automatic creation of Gazetteers (identification of Location Indicators), section 4 presents and discusses experiments and evaluations and section 5 presents concluding remarks.

2 Related Work

Some works utilize supervised learning to create Gazetteers, identifying names that are related to geographical locations from a training corpus. [Overell and Ruger 2007] utilize Wikipedia as source for identifying terms related to toponyms. The technique analyzes pages (entries) related to names of cities. The main goal is to obtain synonyms. [Popescu et al. 2008] also utilize Wikipedia to extract references to cities. [Buscaldi et al. 2006] combine Wikipedia and Wordnet as information sources; Wikipedia is useful to identify terms related to locations and Wordnet is used to identify kinds of locations and to identify in Wikipedia only the pages related to locations, eliminating ambiguities as pages related to non-geographical entities with the same name. The work of [Rattenbury et al. 2007] extracts relations between locations and terms, analyzing semantic tags registered in Flickr (http://www.flickr.com) associated to locations.

The problem is that Wikipedia, Wordnet and Flickr depend on human effort for updates. This may cause the lack of coverage (locations without information) in the Gazetteer or lack of specialized indicators (few indirect references).

The work of [Borges et al. 2003] obtain geographical information from Web pages. The technique finds indirect references as telephone numbers, zip codes and locations names in Web pages related to one city, using a tool for generation of wrappers, that has to be trained with manually selected examples.

An alternative solution is to use textual news published in the Web as source for creating and maintaining Gazetteers. The dynamic characteristic of news allows the identification of specific and up-to-date references and covering a greater number of locations.

[Ferres et al. 2004] utilize machine learning methods over news, obtaining co-referent named entities (for example, “Smith = John Smith”) and acronyms (“USA = United States of America”). [Maynard et al. 2004] utilize similar techniques over annotated corpus. [Kozareva et al. 2006] retrieve toponyms and person names using positioning expressions. They do not identify correlation between the terms and the toponyms. [Garbin and Mani 2005] utilize news to identify collocations between terms and locations. However, the window for analyzing collocations is limited to a distance of 20 terms (they do not utilize relations in the whole text). [Smith and Mann 2003] also analyze collocations in news, however they do not consider the degree of importance or weight of the relations between terms.

The problems of the cited works that utilize news include:

- the need for selecting and preparing a training corpus of news;
- the analysis of relations in windows with limited distance between terms;
- the use of relations without weight, disregarding the relative importance of the relations between terms and locations.

The contributions of the proposed work include:

- the use of news text for the training step, that is, to discover relations between terms (the discovery of Location Indicators), without the need for manually annotating a training corpus; the work does not discuss how to capture news, only suggesting the use of news texts without the need of manual annotation;
- the use of a greater window of words, considering also relations between locations and indicators present in different sentences;
- the use of a weight to determine the importance of the relations identified.

The work also evaluates the proposed approach for constructing Gazetteers in a real geo-referencing process and compares the approach with a Gazetteer created with names of streets and neighborhoods. Furthermore, the work discusses and compares different training corpus composed by news, in order to determine whether choices in the corpus selection influence the results or not.

3 The Approach for Discovering Location Indicators from News

The main goal of this work is to test an approach that identifies Location Indicators related to geographical locations, by analyzing texts of news published in the Web. The work is based on the assumption that the majority of news has some kind of Location Indicator inside the text and that statistical analysis may be utilized for retrieving news according to location data. Gazetteers are created with the identified relations and then they are utilized for geo-referencing of news. Different corpora of news are evaluated for the creation of Gazetteers (and these are evaluated by their ability to correctly identify cities in texts of news).

The first step is to collect news in the Web. In the approach, this step consider an random selection, that is, the capture the text of every news published in Web pages, without filtering. The approach does not indicate a special website or a specific technique for this selection but recommends to use websites that publish news with certainty. The suggestion is the use of well-known and reliable information sources.

The second step is the identification of relations between city names and other terms (Location Indicators). This step demands a pre-processing of the news. As "location indicators" are usually represented by proper names (PNs), the first task is to identify PNs in the texts of the news. This identification is made by analyzing words that start with uppercase, also considering special cases of multi-words (as for example, New York) and expressions that include prepositions (i.e., Massachusetts Institute of Technology). Regular expressions were defined and utilized in this task. Prepositions and adverbs that start a sentence are eliminated. Following suggestion from [Amitay et al. 2004], we obtained with statistical analysis a list of prepositions and adverbs to be eliminated. These words, that appear frequently in lowercase, are called "geographical stopwords" [Hu and Ge 2007]. A special analysis is when the name of a city is part of an expression (example: New York Mayor or, in Portuguese, Prefeito de Nova Iorque). For these cases, names of cities are extracted from the expressions by considering a list of all city names in Brazil and by analyzing the use of prepositions.

The relations between city names and location indicators are determined by a weight (numerical value, representing the importance or probability of the relation). The weight is calculated by the distance between the terms inside texts of a collection (a training corpus). The idea is to calculate the distance between the terms inside each text

of the collection (local weight) and then to utilize the whole collection to determine the final (global) weight. Relations between cities are eliminated.

For the local weight calculus, the approach consider two distances: the internal distance (between terms inside the same sentence) and the external distance (between terms in different sentences of the same text). A sentence is a set of ordered terms between two final points. Formulas (1) and (2) present the calculus of the internal weight $\boldsymbol{Wi_k}$ (inside a sentence $\boldsymbol{k}$) between a city $\boldsymbol{c}$ and a location indicator $\boldsymbol{r}$.

$$\mathrm{Wi_k}(\mathrm{c,r}) = \sum_{\substack{i=1 \\ d \le 9}}^{n} \sum_{j=1}^{m} \frac{(10 - d_{c_i r_j})}{10} \quad (1) \qquad \mathrm{Wi_k}(\mathrm{c,r}) = \sum_{\substack{i=1 \\ d>9; d \le 18}}^{n} \sum_{j=1}^{m} \frac{(19 - d_{c_i r_j})}{100} \quad (2)$$

Where,

$\boldsymbol{d_{xy}}$ is the number of terms between $\boldsymbol{x}$ and $\boldsymbol{y}$ in the sentence, being that $\boldsymbol{x}$ references a city $\boldsymbol{c}$ and $\boldsymbol{y}$ references a location indicator $\boldsymbol{r}$,

$\boldsymbol{k}$ is the k^{th} sentence in the text, where the terms appear together,

$\boldsymbol{i}$ is an index to the i^{th} appearance of the name of $\boldsymbol{c}$ in the sentence,

$\boldsymbol{j}$ is an index to the j^{th} appearance of the term $\boldsymbol{r}$ in the sentence,

$\boldsymbol{n}$ is the total number of appearances of $\boldsymbol{c}$ in the sentence,

$\boldsymbol{m}$ is the total number of appearances of $\boldsymbol{r}$ in the sentence.

For $\boldsymbol{d}$ >18, the weight $\boldsymbol{Wi(c,r)}$ is fixed to the value 0.01. The internal weight ($\boldsymbol{Wi}$) must be calculated for all pairs of terms (referencing cities and location indicators) that appear together inside a sentence.

Formula (3) presents the calculus of the external weight $\boldsymbol{We_t}$, for relations between a city $\boldsymbol{c}$ and a location indicator $\boldsymbol{r}$ present in different sentences of a text $\boldsymbol{t}$.

$$\mathrm{We_t}(\mathrm{c,r}) = \sum_{\substack{i=1 \\ d \le 9}}^{n} \sum_{j=1}^{m} \frac{(10 - d_{c_i r_j})}{1000} \quad (3)$$

Where,

$\boldsymbol{d_{xy}}$ is the number of sentences between $\boldsymbol{x}$ and $\boldsymbol{y}$ in the text $\boldsymbol{t}$, being that $\boldsymbol{x}$ references a city $\boldsymbol{c}$ and $\boldsymbol{y}$ references a location indicator $\boldsymbol{r}$,

$\boldsymbol{i}$ is an index to the i^{th} appearance of the name of $\boldsymbol{c}$ in the text $\boldsymbol{t}$,

$\boldsymbol{j}$ is an index to the j^{th} appearance of the term $\boldsymbol{r}$ in the text $\boldsymbol{t}$,

$\boldsymbol{n}$ is the total number of appearances of $\boldsymbol{c}$ in the text $\boldsymbol{t}$,

$\boldsymbol{m}$ is the total number of appearances of $\boldsymbol{r}$ in the text $\boldsymbol{t}$,

$\boldsymbol{t}$ is the text for which the external weight is being calculated.

For $\boldsymbol{d}$ > 9, the weight $\boldsymbol{We(c,r)}$ is fixed to the value 0.001. The external weight ($\boldsymbol{We}$) must be calculate for all pairs of terms (referencing cities and location indicators) that appear in the text, in different sentences. The formulas and predefined values for $\boldsymbol{d}$ were defined by empirical analysis of samples of texts. The weights were established to

give more relevance to closer relations (inside a sentence) but without disregarding far relations (for example, in different sentences of the text).

The local weight of a relation between ***c*** and ***r*** is calculated as the sum between the internal weight (***Wi***) and the external weight (***We***), for each text (one at each time), as exposed in formula (4). Local weight must sum all internal weights of a relation between ***c*** and ***r***, remembering that internal weights are calculated for each sentence.

$$\mathrm{Wl_t}\,(c,r) = \left[\sum_{k=1}^{n} Wi_{kt}(c,r)\right] + We_t(c,r) \quad (4)$$

Where,

$Wl_t(c,r)$ is the local weight between ***c*** and ***r*** for the t^{th} text in the collection,

t is an index for all texts in the collection,

k is an index for all sentences in the text ***t*** where ***c*** and ***r*** appear together,

n is the total number of sentences inside the text ***t*** where ***c*** and ***r*** appear together,

$Wi_k(c,r)$ is the internal weight between ***c*** and ***r*** for the k^{th} sentence in the text ***t***,

$We(c,r)$ is the external weight between ***c*** and ***r*** for the text ***t***.

The local weight considers relations inside each text. A global weight was defined to consider the whole collection and is calculated as exposed in formula (5).

$$\mathrm{Wg(c,r)} = \frac{\sum_{i=1}^{n} Wl_i(c,r)}{z} \quad (5)$$

Where,

Wl_i is the local weight between ***c*** and ***r***, considering the text ***i***,

i is an index to the texts of the training collection,

n is the total number of texts in the training collection,

z is the total number of cities ***c*** that are related to ***r*** in the collection.

This formula normalizes the weight by dividing the sum by the total number of cities that are related to the term ***r***, considering that a term ***r*** may be related to more than one city. The argument is to give more importance for terms that are related to few cities; general relations or terms (that are related to more cities) will receive a smaller weight.

Other formulas were tested, as for example utilizing simple frequency for the relations between cities and location indicators (without weights) and not utilizing normalization (without dividing the global weight by ***z***). However, results of formal tests (previously carried out) led us to conclude that the formulas presented in this paper generates better results (for example, gains of 15% in precision).

The resulting Gazetteer to be used in posterior geo-referencing processes is composed by a set of cities, each one with a list of Location Indicators (single terms or expressions). Between the city and the indicator, there is a weight (the global weight), representing the relative importance of the relation for identifying the city when the indicator is present in the text (in the case of this paper, texts are news).

4 Experiments and Evaluations

Experiments were carried out to test the approach, including the method utilized for calculating the weight of relations between cities and Location Indicators, and also to compare different training corpus utilized for identifying these relations and thus for creating the Gazetteers.

The evaluation process consists in constructing different Gazetteers with different training corpus and then performing geo-referencing of news from a test collection captured in the Web, analyzing the ability of each Gazetteer in correctly identifying the city associate to the news, through measures like precision and recall. Each Gazetteer has the same structure: a set of cities, each one associated to a list of location indicators. Each association between a city and a Location Indicator has a weight, that is, the global weight calculated as explained in the early section of this paper.

The following Gazetteers were constructed:

(C1) 3000 NP X NP *Old:* the training corpus was composed by 3000 news published in the site Folha Online (http://www.folha.com.br), between the years 2001 and 2006; only relations between proper names were considered;

(C2) 3000 NP X NP *New:* the training corpus was composed by 3000 news published in the site Folha Online, between the years 2007 and 2008; only relations between proper names were considered; the idea is to compare this Gazetteer (with recent news) to the previous Gazetteer (with old news), but both with the same quantity of texts;

(C3) 6000 NP X NP (New+Old): the training corpus was formed by the union of both previous Gazetteers; the idea is to test if a greater collection of texts can generate better results;

(C4) 3000 NP X NP (SA): this Gazetteer was constructed from a training corpus with 3000 news recently published in Folha Online; however, the difference to the previous corpus is that this one was composed only by news that are related to only one city; the idea is to evaluate if training news with only one city result in better performance;

BASELINE: this Gazetteer was composed by location indicators corresponding to names of streets and neighborhoods of the cities. This corpus was created from a special database containing all Brazilian cities and their respective streets and neighborhoods. For this case, the global weight of the relations was not calculated and the value 1 was assumed for all relations.

For evaluating the quality of the Gazetteers (and indirectly the quality of the each corpus utilized), a collection with 230 news published in the web was utilized as a test corpus (news were randomly captured from different years from the Folha Online). No common news were utilized in training and test collections. Only 9 Brazilian cities were considered for the test, including the greatest cities and some medium cities with more than 100 thousand habitants. Each test news references only one city and has at least one

location indicator. The goal is to evaluate if each Gazetteer is useful for identifying the city inside the text of a news.

Due to those restrictions (news with the presence of location indicators and published in different time period), this work utilized training and test corpora especially created for the experiments, instead of using pre-existing corpora as for example GeoCLEF[1] and HAREM[2]. The set of news utilized in the experiments are available for other authors[3].

The evaluation process consists in identifying proper names in the test texts and to compare these terms to the ones stored in the Gazetteer, remembering that it is possible that one term is associated to more than one city in the Gazetteer. Using a probabilistic reasoning, the evaluation process determines the probability of each city be present in the text. Only the more probable city is considered associated to each test text.

The probabilistic reasoning works as following:

- for each city present in the Gazetteer, the steps below are performed;
- for each term associated to the city in the Gazetteer, the presence of this term is verified in the text;
- if the term is present in the text, its weight (global weight as stored in the Gazetteer, associated to the city in question) is summed to the total probability of the city to be present in the text;
- the final sum is utilized as the probability of the city to be present in the news;
- this process is repeated for each city in the Gazetteer and for each text in the test collection;
- only the city with greater probability is considered the unique city associated to the text.

This evaluation process was done for each of the 5 Gazetteers described early.

For each text in the test collection, only one city was associated by the approach being tested. After that, the measures Precision, Recall and F1 (that combines precision and recall, with the same weight) were applied for each Gazetteer.

Results are presented in the table 1. Lines are ordered by the value of F1. The last column presents the total number of relations between a city and a term, present in each Gazetteer. Figure 1 presents the results of precision and recall in a graphical figure.

4.1 Results analysis

This sub-section analyzes the results and discusses the main points.

Comparing the four Gazetteers created by the approach against the baseline Gazetteer (created with names of streets and neighborhoods), we can note that the approach generates better results: all the four Gazetteers performed better than the

[1] http://ir.shef.ac.uk/geoclef/

[2] http://www.linguateca.pt/aval_conjunta/HAREM/

[3] http://gpsi.ucpel.tche.br/~cleber/geoinfo2008/

baseline in both precision and recall measures. We then conclude that news are useful for the creation of Gazetteers and also improve geo-referencing processes. News can help the identification of location indicators that are not related to streets and neighborhoods. A special analysis found that, considering the 100 location indicators with more weights for each tested city in the four Gazetteers constructed by the approach, only 19% of the terms were present in the baseline Gazetteer.

Table 1. Precision, Recall and F1 for each Gazetteer

Gazetter	Prec	Rec	F1	N. Rel.
(C3) 6000 NP X NP	100%	40%	0.5714	9159
(C2) 3000 NP X NP *new*	100%	39%	0.5612	5783
(C1) 3000 NP X NP *old*	100%	36%	0.5294	6945
(C4) 3000 (SA) NP X NP	99.3%	35%	0.5176	4757
Baseline (Streets and Neighbors)	91%	22%	0.3543	119184

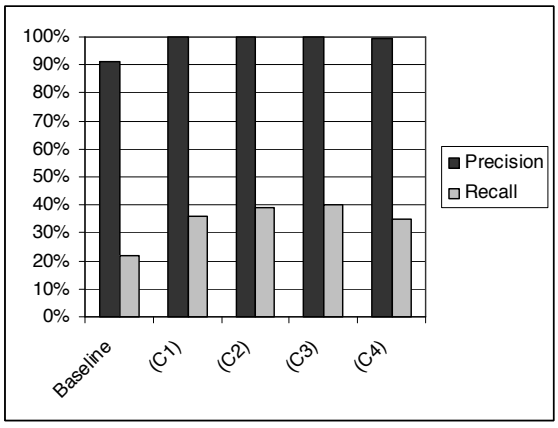

Figure 2. Graphical results of Precision and Recall for each Gazetteer

Comparing the four Gazetteers among them, we first can note that Gazetteers created with news published in different time periods (C1 vs. C2) had a small difference in performance, with a little advantage in recall measure to the Gazetteer created from more recent news (recall: C2 = 39% vs. C1 = 36%). We can conclude from this examination that more recent news are better, but we do not have to capture real-time news or even up-to-date news, because news published one year later can serve for the construction of Gazetteers with relative good performance.

Comparing training collections with different sizes (C3 vs. C1 and C2), we can note that the corpus with greater size (C3) has a better performance but with a small improvement (1.7%). This leads us to conclude that the size of the corpus is important but it may have a limit of performance. Future tests must analyze the size of the training collection composed by news.

Analyzing the performance of the Gazetteer C4, constructed from a corpus where only a city was present in each training text, we can note that this kind of corpus does not bring improvement in the performance. The initial idea was to improve recall, however this did not happen. Our explanation for this poor performance is that this kind of corpus generates a smaller number of relations than the other training collections, that is, identifying less location indicators.

5 Concluding Remarks

The main contribution of this work was to demonstrate that the construction of Gazetteers with Location Indicators instead of using names of cities, streets and neighborhoods are useful to improve geo-referencing processes. This is special important because the texts about locations frequently utilize this kind of information instead of the explicit name of the location (for example, textual news in the Web).

Furthermore, the paper demonstrated that these Location Indicators may be discovered by the analysis of news published in the Web. News can bring different Location Indicators, as for example related to very important people as mayors and authorities, related to entities as hospitals, airports, museums, universities, related to geographical places as highways, parks, constructions, buildings and so on. Most of Location Indicators are dynamic and may change along the time, and for this reason they do not appear in traditional Gazetteers.

Other contribution of the paper is that the creation of Gazetteers may be quite automatically done, by capturing news in the Web and applying the proposed approach. This may cover a great number of locations and maintain up-to-date detailed information about each location with little effort.

Furthermore, news has a special advantage that is to be more accessible than names of streets and neighborhoods. Databases with names of streets and neighborhoods are difficult to be found or must be paid. In addition, such databases, if available, may not consider new cities or changes in the existing cities (as cities that grow fast).

However, we should remember that it is necessary the existence of news about the location for the identification of Location Indicators (related terms). We believe that even small cities have newspapers or local informative vehicles (electronic or in paper) that can be used as a training collection for the Gazetteer construction.

The approach was tested with news written in Portuguese, but other languages may be utilized. The requisite is that be possible to identify proper names in the language. The rest of the approach, including the formulas, remain equal for all languages.

The paper also analyzed different corpus of news as training collections for the automatic construction of Gazetteers (evaluated by the ability of Gazetteers in correctly identifying cities in texts of news). The conclusion is that it is important to maintain the Gazetteer along the time, utilizing more recent news to update the location indicators and the corresponding weights. Although the update of the Gazetteer is important, it can be done one time per year. This is an important finding because the maintenance of the Gazetteer demands efforts and costs.

Future works include the evaluation of different sources, such as Wikipedia, scientific articles and websites for the semi-automatic construction of the Gazetteers and the evaluation of the size of the training collection.

6 Acknowledgements

This work is partially supported by CNPq and CAPES (entities of the Brazilian government for scientific and technological development).

7 References

Amitay, E., Har'el, N., Sivan, R. and Soffer, A. (2004) Web-a-where: Geotagging Web Content. In *Proceedings of the 27th SIGIR*, pages 273–280.

Borges, K. A. V., Laender, A. H. F., Medeiros, C. B., Silva, A. S. and Davis Jr., C. A. (2003) The Web as a data source for spatial databases, *V Simpósio Brasileiro de Geoinformática - GeoInfo*, Campos do Jordão (SP).

Buscaldi, D., Rosso, P. and Garcia, P. P. (2006) Inferring geographical ontologies from multiple resources for geographical information retrieval. In *Proceedings of the 3rd Workshop on Geographical Information Retrieval (GIR),* pages 52–55, Seattle, USA.

Clough, P., Sanderson, M. and Joho, H. (2004) Extraction of semantic annotations from textual web pages. Technical report, University of Sheffield.

Delboni, T.M., Borges, K.A.V., Laender, A. H. F. and Davis Jr., C.A. (2007) Semantic Expansion of Geographic Web Queries Based on Natural Language Positioning Expressions. Transactions in *GIS*, 11(3): 377-397.

Ferres, D., Massot, M., Padro, M., Rodriguez, H. and Turmo, J. (2004) Automatic Building Gazetteers of Co-referring Named Entities. *Proceedings of the 4th International Conference on Languages Resources and Evaluation* (LREC). Lisbon, Portugal.

Garbin, E. and Mani, I. (2005) Disambiguating toponyms in news. In *Proc. Human Language Technology Conference (HLT-EMNLP),* pages 363–370, Vancouver, BC.

Geonames. http://www.getty.edu/research/conducting_research/vocabularies/tgn/ (accessed September 02, 2008).

Hill, L. (2000) Core elements of digital gazetteers: Placenames, categories and footprints Borbinha, J. and Baker, T. (Eds.) *Research and Advanced Technology for Digital Libraries, proceedings*.

Hu, Y. and Ge, L. (2007) A Supervised Machine Learning Approach to Toponym Disambiguation. In: Scharl, A., Tochtermann, K. (eds.): *The Geospatial Web: How Geobrowsers, Social Software and the Web 2.0 are Shaping the Network Society*. Springer, London, 3-14.

Kozareva, Z. (2006) Bootstrapping Named Entity Recognition with Automatically Generated Gazetteer Lists, In *Proceedings of EACL student session (EACL),* Trento, Italy.

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

Leveling, J. and Hartrumpf S. (2007) University of Hagen at GeoCLEF: Exploring location indicators for geographic information retrieval. In *Results of the Cross-Language System Evaluation Campaign, Working Notes for the CLEF Workshop*. Budapest, Hungary.

Leveling, J., Hartrumpf, S. and Veiel, D. (2006) Using semantic networks for geographic information retrieval. In Peters C., Gey F. C., Gonzalo J., Jones G. J. F., Kluck M., Magnini B., Muller H., de Rijke M., editors, Accessing Multilingual Information Repositories: *6th Workshop of the Cross-Language Evaluation Forum, CLEF*, Vienna, Austria, LNCS. Springer, Berlin.

Li H., Srihari R. K., Niu C., Li W. (2003) InfoXtract location normalization: a hybrid approach to geographical references in information extraction. In *Workshop on the Analysis of Geographic References*, *NAACL-HLT*, Edmonton, Canada.

Maynard, D., Bontcheva, K. and Cunningham, H. (2004) Automatic Language-Independent Induction of Gazetteer Lists. In *Proceedings of 4th Language Resources and Evaluation Conference (LREC)*.

Overell, S. E. and Ruger, S. (2007) Geographic Co-occurrence as a Tool for GIR. In *Proceedings of the Workshop On Geographic Information Retrieval (GIR),* Lisboa, Portugal.

Popescu, A., Grefenstette, G. and Moëllic, P. A. (2008) Gazetiki: automatic creation of a geographical gazetteer. In *Proceedings of the 8th ACM/IEEE-CS joint conference on Digital libraries table of contents*, Pittsburgh PA, PA, USA.

Rattenbury, T., Good, N. and Naaman M. (2007) Towards Automatic Extraction of Event and Place Semantics from Flickr Tags. In *Proc. of SIGIR*, Amsterdam, Netherlands.

Sanderson, M. and Kohler, J. (2004) Analyzing geographic queries. In *Proceedings of the Workshop on Geographic Information Retrieval*, Sheffield, UK.

Smith, D. and Mann, G. (2003) Bootstrapping toponym classifiers. In *Workshop on the Analysis of Geographic References, NAACL-HLT*, Edmonton, Canada.

TGN (Getty Thesaurus of Geographic Names).
http://www.getty.edu/research/conducting_research/vocabularies/tgn/ (accessed September 02, 2008).