Journal of Computer Science 6 (6): 653-659, 2010
ISSN 1549-3636

Toponym Disambiguation by Arborescent Relationships

Imene Bensalem and Mohamed-Khireddine Kholladi
Department of Computer Science, Faculty of Engineering Science,
Mentouri University, MISC Laboratory, Algeria

Abstract: Problem statement: The way of referring to a place in the geographical space can be formal, based on the spatial coordinates, or informal, which we use in natural language by using toponyms (place names). A toponym can represent several geographical places. This ambiguity made problematic its conversion towards a unique formal representation. Toponym disambiguation in text is the task of assigning a unique location to an ambiguous place name in a given textual context. **Approach:** Several toponym disambiguation heuristics assumed a geographical proximity between the toponyms of the same context. This proximity can be in terms of spatial distance or in terms of arborsecent relationships, i.e., proximity in the hierarchical tree of the world places. This study presented a new toponym disambiguation heuristic in text based on the quantification of the arborescent proximity between toponyms. This quantification was done by a new measure of geographical correlation that we call the Geographical Density. **Results:** Our method was compared to the state of the art methods using GeoSemCor corpus and it has outperformed them in term of recall (87.4%) and coverage (99.0%). The results showed that the toponyms of the same context are much closer in terms of arborescent relationships than in terms of spatial relationships. **Conclusion:** We believe that the quantification of arborescent relationships between toponyms of the same textual context is a good way to improve the recall of TD task. However, all the arborescent relationships' types must be considered and not only the meronymy, which is the relation the most exploited in the existing TD methods.

Key words: Toponym disambiguation, arborescent relationship, geographical density, referent hierarchical path

INTRODUCTION

The geographical space is ubiquitous. All our activities, experiences, knowledge and decisions are related to places on the geographical space. A reference to a place in this space can be either formal (for instance based on the spatial coordinates) or informal that we use in natural language using toponyms (place names). The formal presentation is the basis of all spatial processing that can be performed by the machine (e.g., spatial analysis and geometric calculation). However, spatial processing is not possible using toponyms (Hill, 2006).

With the increasing number of websites and digital libraries, the text in natural language has become an important source of geographical information (Any information related to a place in the earth is a geographic information) (Borges *et al*., 2003; Morimoto *et al*., 2003; Smith and Crane, 2001). This later is obtained using the automatic Processing of Natural Language techniques (NLP), but unfortunately, it cannot be exploited effectively by machines unless the geographical locations are represented in a formal way, which is not often the case in textual documents. In fact, it has been estimated that at least 70% of the textual documents contain references to geographic locations in the form of toponyms (Hill, 2006).

The conversion of the geographical locations from the informal to the formal representation is a necessity to take advantage of the geographic information extracted from texts such as news stories, historical texts and biographies. However this conversion is problematic because of the ambiguity of toponyms.

In fact, there are two types of toponyms ambiguity, the geo/geo ambiguity and the geo/non-geo ambiguity (Amitay *et al*., 2004).The geo/geo ambiguity arises when a toponym represents several places; for example, in TGN gazetteer (http://www.getty.edu/research/conducting_research/vocabularies/tgn (last visit 20/08/2009)), Tripoli is the name of 16 places in the

Corresponding Author: Imene Bensalem, Department of Computer Science Faculty of Engineering Science, Mentouri University, MISC Laboratory, Algeria

world. The geo/non-geo ambiguity appears when a place name refers also to either a non geographic entities (e.g., Arafat is a place name and also a person name) or has other senses (e.g., Java is a programming language and an Indonesian island).

Toponym Disambiguation (TD) (a.k.a. toponym resolution) addresses the geo/geo ambiguity and it represents the task of assigning a unique location to an ambiguous place name in a given textual context. Once a toponym is disambiguated, it can be represented formally (for example by the latitude and the longitude) to be exploited by machine.

Disambiguation of toponyms is an important task in many domains, such as geographical information retrieval (Overell and Ruger, 2007) and information extraction (Li *et al*., 2003).

This article addresses the problem of toponym disambiguation by proposing a new method based on measuring the geographical correlation between toponyms that appear in the same text (the same context).

Overview of the existing toponym disambiguation methods: Despite the fact that toponym disambiguation methods are very different in spirit, they have common factors (Leidner, 2007). Most TD methods include two main phases for each toponym:

- Extracting the candidate referents: In this phase, for each toponym, the possible referents are extracted from a geographical knowledge resource (ex. a gazetteer or ontology)
- Choosing the correct referent: This phase involves applying a set of heuristics to determine among all the candidates the referent the most likely to be the meaning intended by the ambiguous toponym. TD heuristics rely mainly on the context and the knowledge resources as a source of evidence

We classify the existing heuristics of toponym disambiguation into two main categories: Preference rules-based heuristics and context-based heuristics.

Heuristics of the first category depend mainly on human's preferences and intuition. For instance, assigning to the ambiguous toponym the referent with the largest population (Amitay *et al*., 2004; Pouliquen *et al*., 2004; Rauch *et al*., 2003) or choosing the most frequent referent, for example if the toponym to be resolved is Gaza, applying this heuristic, the referent "Gaza>Palestine" will be chosen instead of "Gaza>USA" because the former is the most known (Stokes *et al*., 2008).

Heuristics of the second category seek evidence clues in the textual environment where the ambiguous toponym occurs; this makes the task of toponym disambiguation similar to the Word Sense Disambiguation (WSD) (Navigli, 2009) which is a common task in NLP domain.

Among this category works, we refer to Leidner *et al*. (2003) who attribute to the ambiguous toponyms of the same context the referents that reduce the bilateral distances to occupy together the smallest possible geometric space. This heuristic takes into account all the candidate referents for each toponym and optimizes using the proximity as criterion.

Clough (2005) proposed a heuristic based on calculating the overlap score between the context and the referent hierarchical path (i.e., the number of toponyms in common). More the score is high; more the referent is likely to be correct. There is also a similar method which seeks in the text the eventual mention of the root place (i.e., the referent's direct holonym). For example, searching Lebanon or Libya if the ambiguous toponym is Tripoli, this heuristic is used by Pouliquen *et al*. (2004) and Li *et al*. (2006).

Smith and Crane (2001) proposed a heuristic that consists in calculating the geographical centroid of toponyms' candidate referents and then remove all referents located more than two standard deviations away from the center. A similar method is proposed in by Rauch *et al*. (2003).

The method of Buscaldi and Rosso (2008a) is based on the calculation of WordNet conceptual density for each referent candidate of the ambiguous toponym. The referent that maximizes the conceptual density is then allotted to the ambiguous toponym. Conceptual Density (CD) is a measure of correlation between the sense of a word and its context. It was presented in the domain of WSD by Agirre and Rigau (1996) and then was reformulated by Rosso *et al*. (2003). This latter is then adapted to the disambiguation of toponyms by Buscaldi and Rosso (2008a). The conceptual density is calculated based on the hierarchical paths of toponyms candidate referents. The hierarchical paths in this method are obtained from WordNet.

The heuristic that we propose is context based and, like Buscaldi and Rosso (2008a) method, it uses toponyms' hierarchical paths obtained from WordNet as a primary knowledge for disambiguation.

The arborescent relationships: By observing the context-based heuristics of toponym disambiguation, we notice that most of this class heuristics are basing on the intuition that assumes the existence of a certain geographical proximity between the toponyms'

referents of the same context. In the methods presented above, (Leidner *et al*., 2003; Smith and Crane, 2001) and (Rauch *et al*., 2003) suppose a distance proximity between the toponyms referents and (Clough, 2005; Pouliquen *et al*., 2004; Li *et al*., 2006) and (Buscaldi and Rosso, 2008a) assume a proximity in the hierarchical tree of world places, that we call arborescent proximity.

In a world hierarchical tree (Fig. 1 shows a part of this tree) we can distinguish two types of arborescent relationships between places: Hierarchical relationships and non-hierarchical relationships.

The hierarchical relationship exists between the components of the same branch in the tree. For example, between a country and each city that it contains (e.g., between Africa, Algeria and Constantine in Fig. 1). The non-hierarchical relationship exists between the nodes that are in different branches but have one (or several) common root (e.g., Algeria and Morocco in the Fig. 1). The common root can be direct (e.g., Andalusia for Seville and Cordoba) or indirect (inherited) (e.g., Africa for Constantine and Marrakech).

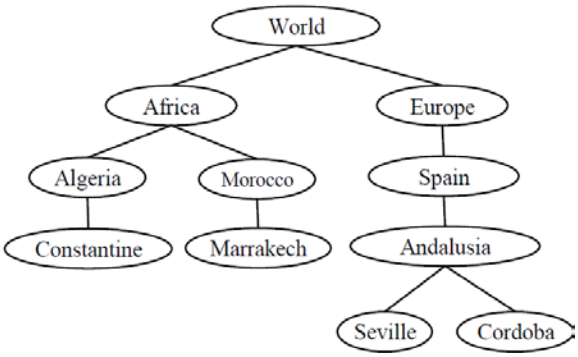

Fig. 1: A part of the hierarchical tree of world places

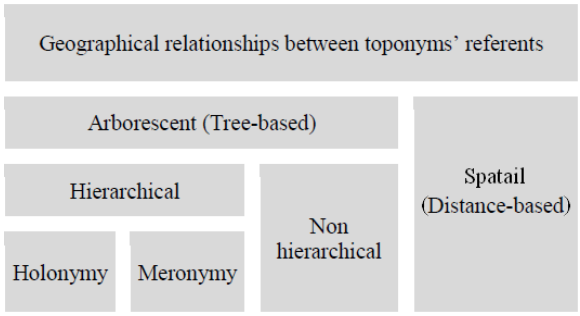

Fig. 2: The different types of geographical relationships that may exist between toponyms' referents of the same context

There exist two sorts of hierarchical relationships: Meronymy that is "part-of" relationship and holonymy that is "has-parts" relationship. For example, we say that Andalusia is a holonym of Cordoba and Cordoba is a meronym of Andalusia.

A hierarchical path of a place is composed of place names interrelated by the holonymy/meronymy relationship. A hierarchical path represents then a branch in the hierarchical tree of the World places. For example, the hierarchical path of Marrakech is "Africa>Morocco>Marrakech" and we say that Morocco is holonym of Marrakech and Africa is a direct holonym for Morocco and inherited holonym for Marrakech.

The types of the geographical relationships are summarized in Fig. 2.

We notice that most of the existing methods based on arborescent proximity are able to resolve toponyms by searching its holonyms in the same context (i.e., its meronymy relationships) (Since the meronymy is the relation "is-part-of", then looking for meronymy relationships for a place name consists in finding its holonyms). Clough (2005) quantifies the existence of this relationship by the Overlap Score (OS) between the context and the referent hierarchical path. The OS allows discovering the occurrence of all the referent's holonyms, either direct or indirect, in the context. Pouliquen *et al*. (2004) and Stokes *et al*. (2008) methods seeks holonyms in the text without calculating the OS.

However, to the best of our knowledge, the only method that exploits other arborescent proximity (i.e., not only the meronymy relationships) is the conceptual density-based method proposed by Buscaldi and Rosso (2008a). For example, if we consider {Georgia, Atlanta, Savannah, Texas} (These toponyms are taken from the file br-a01 of GeoSemCor corpus) to be the context toponyms, this method disambiguates Georgia to "Georgia>USA" instead of "Georgia>Asia", because the former has a common root -which is USA-with the other context toponyms (non hierarchical relationship) and also it contains Atlanta and Savannah as parts (meronymy relationship) (It should be noted that this explanation is ours and it represents our own point of view about Buscaldi and Rosso (2008a) method and it is not taken from what is described in their article).

MATERIALS AND METHODS

We propose in the rest of this article a new toponym disambiguation heuristic. Our heuristic is able to discover all arborescent relationships (hierarchical and non hierarchical) between the toponyms of the same context and it is based on a new correlation measure between the toponyms that we call the geographical density.

Table 1: Notation conventions in the geographical density heuristic

T: all the toponyms that appear in a document D. Each toponym appears one time in T. n is the number of toponyms.	$T = \{t_i \in D, i = 1...n\}$
G: a gazetteer. Each referent r_{id} is represented by a set of characteristics which differ according to the used gazetteer. In this heuristic we need for each referent: His ID, his toponym and his hierarchical path. We say that the place r_{id} is a referent of t_i if t_i is the name of r_{id}.	$G = \{r_{id}, r_{id}$ is a geographical location in the Earth$\}$
h_{id} is the hierarchical path of r_{id} in the hierarchical tree of G. Each h_{id} node is a reference $r_{id\cdot k}$, where the first node $r_{id\cdot 1}$ is the extreme inherited holonym and the last node $r_{id.l}$ is $r_{id\cdot}$, where l is the length of the hierarchical path	$h_{id} =$" $r_{id\cdot 1}>r_{id\cdot 2}>\ldots>r_{id.l}$"
Comp (h_{id}) are the referents that compose a hierarchical path h_{id}.	Comp (h_{id}) = { r_{idk}, k=1..l}
R_i the referents set of the toponym t_i.	$R_i = \{r_{id} \in G, t_i$ is the name of rid$\}$
H_i a set composed of the hierarchical paths of referents in R_i	$H_i = \{h_{id}, r_{id} \in R_i\}$
R is the set of sets R_i of all toponyms t_i that appear in a document D	$R = \{R_i, i = 0..n\}$
H is the set of hierarchical paths of all referents of all toponyms ti that appear in a document D	$H = \{H_i, i = 0..n\}$
Comp (H_i): is the set of the components of all $h_{id} \in H_i$ without duplication of elements	Comp (H_i) = $\cup$ Comp (h_{id}), $h_{id} \in H_i$

Notation: Table 1 contains notations used to define the geographical density notion.

Principle: Our heuristic is based on the assumption that toponyms that appear together in the same document are related geographically with arborescent relationships.

The proposed heuristic resolves a toponym by the referent which is:

- The most linked geographically to the referents of other toponyms in the World places hierarchical tree, i.e., its hierarchical path has relatively many referents in common with the hierarchical paths components of the referents of the other toponyms in T i.e., with Comp (H-Hi)elements, (we can say that is an indirect relationship with the context)
- The most linked to the context, i.e., its hierarchical path and the context contain relatively many names in common

These two features are quantified by calculating what we call the Geographical Density (GD) which is defined as a measure of the arborescent correlation between a referent of a toponym and the toponyms of the context in which it appears.

Toponym Disambiguation by the geographical density consists of the following steps:

- Extract all the toponyms of the document at hand D
- Eliminate duplications (applying one sense per discourse assumption). T is the set of toponyms of the document D without duplications
- Determine the list of candidate referents R_i for each toponym ti. Each candidate referent rid must be represented by its hierarchical path h_{id}
- Calculate the geographical density for each candidate referent in R_i, i = 1..n
- Allocate to each toponym t_i the candidate referent rid with the maximum geographical density GD (r_{id}, T)

The geographical density: The Geographical density calculation is essentially based on the candidate referents hierarchical paths of all toponyms in the context (the hierarchical paths of all R elements i.e., H). The hierarchical path of a referent is composed of the referent itself and its holonyms i.e., its direct and indirect roots (With a view to brevity, henceforth, when we say referent's holonyms we mean all holonyms (direct and inherited) that compose its hierarchical path).

The GD of a referent r_{id} of an ambiguous toponym t_i increases when: (a) This referent is among holonyms of other referents in R-R_i and/or (b) its holonyms are among the candidate referents of the other toponyms (i.e., among R-R_i elements) and/or (c) its holonyms are also holonyms for other referents in R-R_i and (d) its inherited holonyms are partially or wholly in the context.

(a), (b) and (d) indicate the presence of a hierarchical relationship between the target referent r_{id} and some referents of other toponyms and (c) indicates the presence of a non-hierarchical relationship.

(a), (b) and (c) are quantified by calculating the frequency of the referent r_{id} and its holonyms ($r_{id.1}$, $r_{id.2}$,…, $r_{id.l}$) in the set R. The frequency of a reference $r_{id.k}$ is the sum of its weight in each R_i (Eq. 2). The weight W is a Boolean function which indicates the existence or the absence of a referent $r_{id.k}$ in the set Comp(H_i) (Eq. 3). The greatest value that can take a frequency is n: the number of the sets R_i in R.

(d) is quantified by calculating the overlap score of the hierarchical path of the referent r_{id} with the context T, this is represented by the value OS (h_{id}, T).

The geographical density GD (r_{id}, T) of a candidate referent r_{id} is the sum of these two values described above (frequency of each Comp(h_{id}) elements in R and the overlap score of this later with the context) (Eq. 1):

$$GD(r_{id}, T) = \sum_{k=1}^{l} (Frequnecy(r_{id.k}, R)) + OS(h_{id}, T) \quad (1)$$

$$Frequnecy(r_{id}, R) = \sum_{i=1}^{n} W(r_{id}, R_i) \quad (2)$$

$$W(r_{id}, R_i) = \begin{cases} 0, \text{if the number of } r_{id} \text{ in Comp}(H_i) = 0 \\ 1, \text{if the number of } r_{id} \text{ in Comp}(H_i) \neq 0 \end{cases} \quad (3)$$

Experimentation: The evaluation of the toponyms disambiguation methods requires the use of two main resources that are: Textual corpora and sense inventories, for instance gazetteers. The evaluation is still problematic in this area due to lack of standard resources that enable the comparison between the performances of different methods. Leidner (2004; 2006) addressed this problem but unfortunately his data are not freely available (according to a personal communication with Jochen Leidner).

Buscaldi and Rosso (2008a) have evaluated their method based on the conceptual density using the WordNet ontology (http://wordnet.princeton.edu/) as a senses inventory and the corpus GeoSemCor.

We choose to evaluate our method using the same resources used by Buscaldi and Rosso (2008a): WordNet ontology and GeoSemCor corpus. This choice has two reasons: On one hand, these resources are the only resources of DT freely available and on the other hand this will allow us to compare our method with that of Buscaldi and Rosso (2008a).

The GeoSemCor corpus presented for the first time in (Buscaldi and Rosso, 2008a) is a version of SemCor where each toponym can be assigned to its correct referent in WordNet. This corpus is freely available on Buscaldi's hompage (http://users.dsic.upv.es/~dbuscaldi/resources/geosemcor2.0.tar.gz). The Table 2 provides some information about it.

Table 2: Information about the GeoSemCor corpus

Total number of toponyms	1210.00
Ambiguous toponym	498.00
Document number	123.00
Average number of toponyms per document	9.84
Toponyms number without duplications in the same document	693.00
Average number of toponyms per document without duplication	5.20
Number of duplication with different senses in the same document	13.00

WordNet (Miller, 1995) is an electronic lexical database of English but also available for many other languages. Words included in WordNet are linked to each other by a variety of semantic relations, among them holonymy (and its inverse meronymy), which is the most significant relationship between place names. Words in WordNet are grouped in 4 categories: Nouns, verbs, adjectives and adverbs. The nouns are in turn divided into 26 classes. Toponyms are founded among the nouns of these two classes: Location and Object. Location class contains nouns denoting a spatial position; however, the Object class contains nouns denoting natural objects.

RESULTS

Experiments in this study aim to understand the role of the discovered arborescent relationships between toponyms of the same context in the TD task and the evaluation of our method by comparing it with others.

Performance estimation of toponym disambiguation methods is done with metrics used in Information Retrieval (IR) and automatic Processing of Natural Languages (NLP) area. These metrics are: Precision, recall, coverage and f-measure. They are calculated in the TD domain as follows:

$$\text{Precision} = \frac{\text{Number of toponymes resolved correctly}}{\text{Number of resolved toponymes}}$$

$$\text{Recall} = \frac{\text{Number of toponymes resolved correctly}}{\text{Total number of toponyms}}$$

$$\text{Coverage} = \frac{\text{Number of resolved toponymes}}{\text{Total number of toponyms}}$$

$$\text{F\_measure} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

Experimentation results are in the Table 3.

GD line represents the results of our method based on the geographical density. This latter-as explained above-is the sum of the referent frequency and the overlap score of its hierarchical path with the context.

Table 3: Experimentation results of GD method and other methods using GeoSemCor corpus

	Precision (%)	Recall (%)	Coverage (%)	F-measure
GD (freq + OS)	88.2	87.4	99.0	0.878
OS	90.8	78.3	86.3	0.841
CD	89.9	77.5	86.2	0.832
Map	87.9	70.2	79.9	0.781

The OS line represents the results of experiments with the overlap score only. The line called CD represents the results of the conceptual density-based method of Buscaldi and Rosso (2008a). Map shows the results of Smith and Crane (2001) method, which is distance proximity-based. The results of these 4 methods were obtained using the corpus GeoSemCor. CD and Map results are taken from (Buscaldi and Rosso, 2008a; 2008b) by considering all toponyms of the document as a context.

DISCUSSION

The experiments results show that OS method has the highest precision; this means that the occurrence of toponym's holonyms in the context is the most accurate indicator of its sense.

The recall using the geographical density is higher compared to that of OS. This confirms that to disambiguate the largest number of toponyms it is more efficient to detect all types of arborescent relationships rather than the meronymy relationships only.

The coverage of our method is the higher one. This shows that the geographical density is the most discriminant geographical correlation measure compared to the overlap score with the context, the conceptual density and the spatial distance measures.

By comparing our method with Map method we can deduce that the toponyms of the same context are much closer in terms of arborescent relationships than in terms of spatial relationships.

The recall of our method is considerably high compared to that of CD-based method (+9.9%). However, its precision is a little smaller (-1.7%) compared to that of CD-based method. This is due to the coverage of our method which is higher than the CD-based method.

CONCLUSION

We present in this article a new toponym disambiguation heuristic which is based on the assumption of the existence of an arborescent geographic relationship between toponyms of the same context. So, it resolves toponyms ambiguity by choosing the referents the most related between them in the hierarchical tree of the world places. In addition, we have classified the arborescent relationships in two classes: Hierarchical and non-hierarchical relations.

To quantify the degree of arborescent proximity we have introduced a measure of geographical correlation that we have called the Geographical Density (GD), this is by analogy to the Conceptual Density (CD) used for word sense disambiguation and applied by Buscaldi and Rosso (2008a) for the TD.

The evaluation results insure the validity of our assumption. In addition, it shows that the search of meronymy relationship which is a heuristic used in a lot of DT methods is precise but insufficient to disambiguate all toponyms of a context.

The performance of the GD-based method has exceeded CD-based method performance in terms of recall and coverage. CD-based method is similar to ours in that it uses referents hierarchical paths as the main knowledge.

Comparing our method with that of Smith and Crane (2001) shows that toponym disambiguation relying on arborescent proximity is more accurate and more efficient than disambiguation based on proximity in terms of distance.

Finally, we recognize that GeoSemCor corpus and WordNet allow us to evaluate our method and compare it to others, but in reality these two resources are not really dedicated to the task of toponym disambiguation. There is no doubt that the use of resources tailored to TD task will enable us to make a more precise evaluation.

ACKNOWLEDGEMENT

We would like to thank Davide Buscaldi for sending us an original copy of his article (Buscaldi and Rosso, 2008a) and also for sharing GeoSemCor corpus in the Web. We are also most grateful to Simon Overell who kindly suggested to us the evaluation of our method using the GeoSemCor corpus.

REFERENCES

Agirre, E. and G. Rigau, 1996. Word sense disambiguation using conceptual density. Proceeding of the 16th Conference on Computational Linguistics, Aug. 5-9, Association for Computational Linguistics, Copenhagen, Denmark, pp: 16-22. DOI: 10.3115/992628.992635

Amitay, E., N. Har'El, R. Sivan and A. Soffer, 2004. Web-a-where: Geotagging web content. Proceeding of the 27th Annual International ACM SIGIR Conference on Research and Development In Information Retrieval, July 25-29, ACM Press, Sheffield, UK., pp: 273-280. DOI: 10.1145/1008992.1009040

Borges, K.A., A.H. Laender, C.B. Medeiros, A.S. Silva and C.A. Davis, 2003. The web as a data source for spatial databases. Proceeding of the Anais do V Brazilian Symposium on Geoinformatics, Nov. 3-5, Campos do Jordao, Brazil. http://www.geoinfo.info/geoinfo2003/papers/geoinfo2003-38.pdf

Buscaldi, D. and P. Rosso, 2008a. A conceptual density-based approach for the disambiguation of toponyms. Int. J. of Geogr. Inform. Sci., 22: 301-313. DOI: 10.1080/13658810701626251

Buscaldi, D. and P. Rosso, 2008b. Map-based vs. knowledge-based toponym disambiguation. Proceeding of the 2nd international Workshop on Geographic Information Retrieval, Oct. 29-30, ACM Press, Napa Valley, California, USA, pp: 19-22. DOI: 10.1145/1460007.1460011

Clough, P., 2005. Extracting metadata for spatially-aware information retrieval on the Internet. Proceeding of the ACM Workshop on Geographic Information Retrieval, Nov. 4, ACM Press, Bremen, Germany, pp: 25-30. DOI: 10.1145/1096985.1096992

Hill, L.L., 2006. Georeferencing: The geographic associations of information. MIT Press, ISBN: 978-0-262-08354-6, pp: 260.

Leidner, J.L., 2007. Toponym Resolution in Text: Annotation, Evaluation and Applications of Spatial Grounding of Place Names. Dissertation.com, ISBN: 978-1581123845, pp: 289.

Leidner, J.L., G. Sinclair and B. Webber, 2003. Grounding spatial named entities for information extraction and question answering. Proceedings of the HLT-NAACL 2003 Workshop on Analysis of Geographic References, May 31-31, Association for Computational Linguistics, Edmonton, Canada, pp: 31-38. DOI: 10.3115/1119394.1119399

Leidner, J.L., 2004. Towards a reference corpus for automatic toponym resolution evaluation (extended abstract). Proceeding of the Workshop on Geographic Information Retrieval held at the 27th Annual International ACM SIGIR Conference (SIGIR'04), Sheffield, England, UK., pp: 1-6. http://www.geo.uzh.ch/~rsp/gir/abstracts/leidner.pdf

Leidner, J.L., 2006. An evaluation dataset for the toponym resolution task. Comput. Environ. Urban Syst., 30: 400-417. DOI: 10.1016/j.compenvurbsys.2005.07.003

Li, H., R.K. Srihari, C. Niu and W. Li, 2003. InfoXtract location normalization: A hybrid approach to geographic references in information extraction. Proceeding of the HLT-NAACL 2003 Workshop on Analysis of Geographic References, May 31-31, Association for Computational Linguistics, Edmonton, Canada, pp: 39-44. DOI: 10.3115/1119394.1119400

Li, Y., A. Moffat, N. Stokes and L. Cavedon, 2006. Exploring probabilistic toponym resolution for geographical information retrieval. Proceedings of the 3rd ACM Workshop On Geographic Information Retrieval, GIR 2006, Aug. 10-10, Department of Geography, University of Zurich, Seattle, WA., USA., pp: 17-22. http://www.geo.unizh.ch/~rsp/gir06/papers/individual/li.pdf

Miller, G.A., 1995. Word Net: A Lexical database for English. Commun. ACM, 38: 39-41. DOI: 10.1145/219717.219748

Morimoto, Y., M. Aono, M.E. Houle and K.S. McCurley, 2003. Extracting spatial knowledge from the web. Proceeding of the 2003 Symposium on Applications and the Internet, Jan. 27-31, IEEE Computer Society, Orlando, FL., USA., pp: 326-333. DOI: 10.1109/SAINT.2003.1183066

Navigli, R., 2009. Word sense disambiguation: A survey. ACM Comput. Surveys, 41: 2. DOI: 10.1145/1459352.1459355

Overell, S.E. and S. Ruger, 2007. Geographic co-occurrence as a tool for GIR. Proceeding of the 4th ACM Workshop on Geographical Information Retrieval, Nov. 9-9, ACM Press, Lisbon, Portugal, pp: 71-76. DOI: 10.1145/1316948.1316968

Pouliquen, B., R. Steinberger, C. Ignat and T.D. Groeve, 2004. Geographical information recognition and visualization in texts written in various languages. Proceeding of the 2004 ACM Symposium on Applied Computing, March 14-17, ACM Press, Nicosia, Cyprus, pp: 1051-1058. DOI: 10.1145/967900.968115

Rauch, E., M. Bukatin and K. Baker, 2003. A confidence-based framework for disambiguating geographic terms. Proceeding of the HLT-NAACL 2003 Workshop on Analysis of Geographic References, May 31-31, Association for Computational Linguistics, Edmonton, Canada, pp: 50-54. DOI: 10.3115/1119394.1119402

Rosso, P., F. Masulli, D. Buscaldi, F. Pla and A. Molina, 2003. Automatic noun sense disambiguation. Lecture Notes Comput. Sci., 2588: 273-276. DOI: 10.1007/3-540-36456-0_27

Smith, D.A. and G. Crane, 2001. Disambiguating geographic names in a historical digital library. Lecture Notes Comput. Sci., 2163: 127-136. DOI: 10.1007/3-540-44796-2_12

Stokes, N., Y. Li, A. Moffat and J. Rong, 2008. An empirical study of the effects of NLP components on geographic IR performance. Int. J. Geogr. Inform. Sci., 22: 247-264. DOI: 10.1080/13658810701626210