COMPUTATIONAL SCIENCE AND TECHNIQUES
Volume 2, Number 1, 2014, 229-237

Publisher: Klaipėda University
http://journals.ku.lt/index.php/CST
Online ISSN: 2029-9966
DOI: http://dx.doi.org/10.15181/csat.v2i1.13

AUTOMATIC EXTRACTION OF GEOGRAPHIC CONTEXT FROM TEXTUAL DATA

Jurijs Nikolajevs, Gints Jekabsons
Riga Technical University, Faculty of Computer Science and Information Technology, Institute of Applied Computer Science, Latvia
juryn@inbox.lv, gints.jekabsons@rtu.lv

Abstract. The amount of information on the internet grows exponentially. It is not enough anymore just to have a general access to this huge amount of data, instead it is becoming a necessity to be able to use different kinds of automatic filters to retrieve just the information you actually want. One solution for the information filtering and retrieval is context analysis in which one of the contexts of interest is the geographic context. This paper studies the problem and methodology of geoparsing – recognition of geographic names in unstructured textual content for the aim of extracting geographic context. A prototype implementation of a geoparsing system, capable of automatically analyzing unstructured text, recognizing geographic information and marking geographic names, is developed. Empirical evaluation of the system using articles from real-world news showed that the average quality of its geographic name recognition varies around 75-100%. Possible applications of the developed prototype include automated grouping of any texts by their geographic contexts (e.g., in news portals) and location-based search. Preliminary results of empirical evaluation showed that the average rate of its geographic name recognition varies around 75-100%.

Keywords: geocoding, geoparsing, geographic context, natural language processing.

Introduction

Nowadays, there are a lot of technologies and simple (web) services allowing to manually geotag photos, videos, texts, and other information, but there is still a lack of systems that can extract the geographic context automatically. This study tackles with automatic extraction of geographic context from unstructured text documents.

The great majority of textual data in the web is not directly linked to geographic context. Such linking would be very useful for information searching and structuring. Automated geoparsing (Goldberg, 2008; Keller, Freifeld, Brownstein, 2009; Abascal-mena, López-ornelas, 2009; Nikolajevs, 2011) can help solving this problem by efficiently automating the linking of the textual data to geographic identifiers (e.g., codes or geographic coordinates expressed in form of latitude-longitude). These links can then be used in search engines and other applications that can operate with geographic data. The geocoded texts can also be linked with descriptions of the places as well as their photos, videos, etc. Such geoparsed texts can provide reader with additional information about the geographic context of the text. The user can obtain the additional information about geographic objects mentioned in the text by simple click on the objects name in the text. Sets of geoparsed texts can be structured according to the locations mentioned in them allowing their readers to select the most appropriate information in the shortest time (one

of the examples could be the news portal that structures sets of articles by their geographic content). Identified geographic places can also be displayed on a geographic map.

The previous researches in this sphere have rather specific applications. There are a number of commercial products with geoparsing capabilities. Companies like MetaCarta extract information about place and time, Digital Reasoning (GeoLocator), Lockheed Martin (AeroText), and SRA (NetOwl) (Caldwell, 2009) extract places along with other entities, such as persons, organizations, time, money, etc. Many of the existing services analyze only their specific fields. For example, HealthMap is part of a new generation of online systems designed to monitor and visualize, on a real-time basis, disease outbreak alerts as reported by online news media and public health sources (Keller, Freifeld, Brownstein, 2009). Marc Wick's and Torsten Becker's system works only with RSS data (Scharl, Tochtermann, 2007). These ideas can be combined together and used as a base of a new extraction algorithm.

This paper studies the methodology of geoparsing as well as related problems and solutions thereof, describes main principles and implementation of geoparsing ideas into a prototype, and presents the results of its empirical evaluation.

1. Methodology

Geoparsing is the process of assigning geographic identifiers (e.g., codes or geographic coordinates expressed as latitude-longitude) to words and phrases that occur in unstructured textual content. It can be divided into two main phases (Goldberg, 2008; Nikolajevs, 2011): (1) recognition of geographic names in text and (2) assignment of the most probable geographic identifiers to the words designating the real world objects (geocoding).

1.2. Recognition of geographic names

To recognize potential geographic names in an unstructured text, the so-called external (or gazetter-based) approach can be used (Goldberg, 2008; Mikheev, Moens, Grover, 1999). This approach requires a gazetteer database that contains geographic names (including their variations) and their locations (geographic coordinates). According to study in (Mikheev, Moens, Grover, 1999), the external approach gives higher recognition results in comparison with internal approaches which recognize geographic names using only the information given in text.

In simplest case of external approach, each text word that begins with a capital letter is checked for its existence in the gazetteer database (this will work with majority languages where the proper names starts with capital letter). Additionally, the database is also queried for such word combinations where the second word indicates that the previous one is a geographic object (e.g., where the second word is "lake", "river", "bay", etc.) as well as two or more consecutive words that start with capital letters. Also, to be able to recognize old place names or place names in other languages, gazetteer database must contain such alternate names. Additionally, because of linguistic features found in many languages it might be necessary to transform all words to their basic primary form before querying the database (it can be done by

analyzing the ending of each word and, according to the grammar rules, returning all possible basic forms of the word or word combination; for this purposes language grammar tables, that are available in any language dictionary, can be used).

The result of this phase is an unfiltered list of (potentially many) geographic names that potentially could refer to actual geographic places mentioned in the text.

1.3. Assignment of the most probable geographic identifiers

Previous phase returned a list of potential geographic names from the gazetteer. At this point it is not clear, whether a potential place name is really the place mentioned in the text. Therefore, we need to develop heuristics for ranking of the place names found in gazetteer. But this is not a trivial task. For example, there are also many names of geographic objects, that could are used as regular words in other languages, or are homonymic with person names, e.g., city of Paris and person Paris Hilton. Also many place names are not unique – a number of European cities have duplicates in the "New World" where the settlers named their colonies in honor of their countries or cities in Europe (for example, the city of Newcastle in England and Newcastle in Australia). Street names also can be homonymic to city names. To solve some of these problems, context analysis or special stop-lists can be created. Using context analysis, some other place names found in the text can point to the region where the place is located (e.g., USA or UK in case of Newcastle).

In the developed prototype, each place name is evaluated according to three criteria: First, every name found in the database should be compared to the original name, which was queried to the database. If the returned queried word exactly matches a place name, that word gets the highest rating – the highest level of reliability. If it matches an alternative name (for example it was queried Dinaburg that is old name of Daugavpils city) of the place it gets medium level of reliability. If there is only a partial match of the names (it was queried word combination, but database returned only one word from it or queried word is part of other word) it gets lower level of reliability. Let R_1 be normalized reliability:

$$R_1 = \frac{L_{cur}}{L_{\max}} \tag{1}$$

where L_{cur} – reliability for the current place;
L_{max} – maximal possible reliability.

Let R_2 be normalized frequency of the name occurrence in the text (if the name appears in the text more often, the probability that the text is about this geographic object increases):

$$R_2 = \frac{F_{cur}}{F_{\max}} \tag{2}$$

where F_{cur} – frequency of occurrence for the current place;
F_{max} – maximal frequency of occurrence (frequency of occurrence of the most repeatable place name).

Let R_3 be normalized evaluation of mutual arrangement of the referred geographic objects on geographic map (the whole text is most likely related to the region where most of the found places are situated):

$$R_3 = \frac{\sum_{i=0}^{n} d(a_{cur}, a_i)}{\sum_{i=0}^{n-1} \sum_{k=i+1}^{n} d(a_i, a_k)} \tag{3}$$

where d – distance between two point on the Earth surface;
a – one of the geographic objects found in the text (a_{cur} points to the current object);
n – total amount of geographic objects found in the text.

Now the final rating of the current geographic object is:

$$R_{final} = \frac{1}{3}(R_1 + R_2 + R_3). \tag{4}$$

This evaluation is applied for all potential geographic objects found in the text and finally the object with the highest rating is selected to be linked to the text. For the primary tests, described in this paper, the weights of each criteria are equal. In the future it is planned to make criteria weight value optimization using Genetic algorithm.

1.4. Architecture of a geoparsing system

Architecture of a geoparsing system can be divided into two stages: data preparation and data presentation (Figure 1). The data preparation stage is performed periodically, independently from user activities. For example, every hour the system connects to news feeds of external news sources and downloads their articles. These articles are then sent to the geoparser in order to identify potential geographic names which are then sent to geocoder. The goecoder queries the gazetteer database and retrieves place name identifiers for the analyzed text words that were found in the database. The geocoder then performs data ranking and links the words to the most highly ranked results. Finally, the texts with the linked geographic information are stored in a database until further retrieval for presentation purposes.

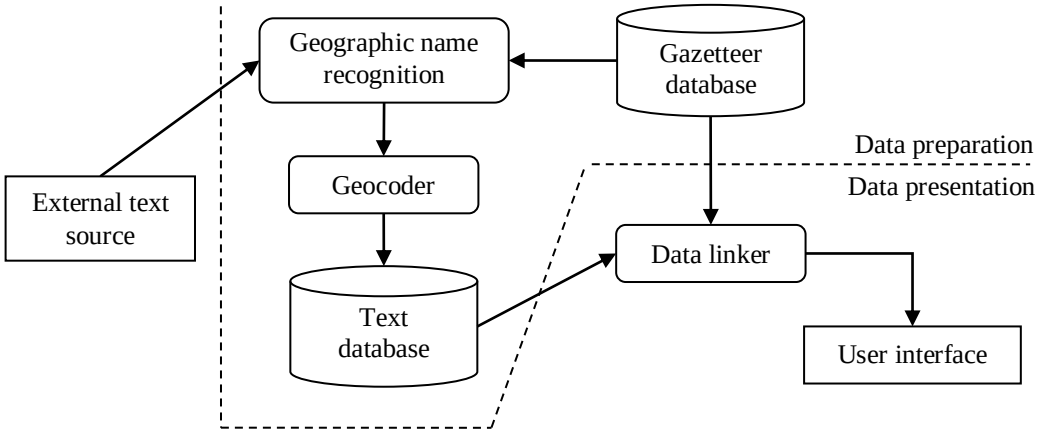

Figure 1. Overall architecture of textual data geoparsing system

In the data presentation stage, the texts from the database are retrieved, linking the additional information to the identifiers got in the previous stage. The text with attached additional information can then be shown in the user interface. The operations of the presentation stage are performed only on user request.

2. Developed prototype and it's empirical evaluation

In this study, the developed geoparsing system prototype is oriented to the analysis of texts in Latvian language mentioning geographic places of Latvia. The used gazetteer consists of geographic names of Latvia and most popular place names in the world. The prototype supports both perfect and partial match database search methods, as well as potential geographic name filtering, word basic form generation, and ranking of found names. The prototype is implemented using Java programming language while as database storage open-source MySQL database management system was chosen.

For empirical evaluation of the developed prototype, a dataset of 20 real-world news articles (in Latvian and about Latvia) was created. The system was evaluated in two ways: recognition accuracy of individual geographic names and recognition accuracy for a whole article (in this case the whole text was to be linked to one geographic place to which the text refers the most).

Table 1 summarizes the results of the performed empirical evaluation. The table contains the following columns: the amount of correctly identified geographic names (in percents); the amount of correctly identified geographic names which exist in the database (in percents) – this is an important indicator to evaluate the recognition rate regardless of the completeness of the gazetteer database; the amount of incorrectly identified geographic names in relation to the total number of geographic names in the text; column with the answer "Yes" or "No" that shows whether the geoparser linked the whole text to the one correct geographic object. The results are calculated for both recognition methods – exact match (only the results that exactly matched the query was returned from database) and partial match (database was able to return results that partly match the query, for example, was queried "Piņki" but database was returned both "Piņki" and "Lielie Piņki"), and using both, main names and alternative names from the gazetteer. In cases when exact match and partial match methods gave different results for the whole text linking, the last column contains two different answers: "(1)" when the result was obtained with exact match method and "(2)" when it was obtained with partial match method.

The results show that using the exact match method on average 80.14% of geographic names in the analyzed texts were found correctly. This percentage is considerably lower than 100% because geoparser made mistakes but most of the mistakes are due to the fact that the used gazetteer did not contain many of the places mentioned in the texts. If recognition rate is calculated only from those geographic names stored in the gazetteer, almost all of the names were identified correctly – an average of 98.33%. This suggests that the basic form generation algorithm is working very well – it has generated all basic forms correctly, but for future experiments and practical applications the gazetteer should be made more complete.

The amount of incorrectly identified geographic names (false positives) on average was 31.52% – the system incorrectly identified objects that actually were not mentioned in the text. This was because it identified places with the same name but different coordinates. For example, Gulbarga (in Gulbene district) and Gulbarga (in Daugavpils district) or Abava (in Ventspils district) and Abava (in Tukums district), etc. Note that in cases when the amount of such incorrectly identified names is larger that the total amount of geographic names in text, the calculated level may exceed 100%. Some news articles also mentioned more specific places, for example, one mentioned "Rīgas zooloģiskais dārzs". As the gazetteer database does not contain such name, the system only recognized city "Rīga" which in this case is only partly correct.

Table 1. The results of empirical evaluation

Article	Exact match method			Partial match method			Whether the system has correctly linked the whole text
	Correctly identified geographic names, %	Correctly identified geo. names that exist in gazetteer, %	Incorrectly identified geographic names, %	Correctly identified geographic names, %	Correctly identified geo. names that exist in gazetteer, %	Incorrectly identified geographic names, %	
1	50%	100%	50%	50%	100%	100%	Yes
2	66,67%	100%	0%	66,67%	100%	100%	Yes(1),No(2)
3	100%	100%	8.33%	100%	100%	50%	Yes
4	50%	100%	0%	50%	100%	200%	Yes
5	100%	100%	200%	100%	100%	200%	Yes
6	100%	100%	0%	100%	100%	50%	Yes
7	50%	100%	0%	50%	100%	0%	Yes
8	100%	100%	33.33%	100%	100%	33.33%	Yes
9	100%	100%	0%	100%	100%	100%	Yes(partly)
10	83.33%	100%	0%	83.33%	100%	0%	-
11	100%	100%	0%	100%	100%	20%	Yes(partly)
12	60%	100%	40%	60%	100%	140%	No
13	66.67%	66.67%	33.33%	100%	100%	66.67%	Yes(2),No(1)
14	100%	100%	75%	100%	100%	125%	No
15	100%	100%	0%	100%	100%	50%	Yes
16	66.67%	100%	33.33%	66.67%	100%	33.3%	No(partly)
17	92.86%	100%	7.14%	92.86%	100%	14.29%	Yes(partly)
18	66.67%	100%	0%	66.67%	100%	33.33%	Yes
19	50%	100%	50%	50%	100%	100%	Yes
20	100%	100%	50%	100%	100%	25%	Yes
Average:	**80.14%**	**98.33%**	**31.52%**	**81.81%**	**100.00%**	**72.05%**	

The percentage of correctly recognized geographic names using the partial match method is slightly higher (81.81%) because there were identified some geographic names due to their alternative names. Most recognition problems here are the same which were described in analysis for perfect match method. However, all the geographic names that were stored in the gazetteer database were successfully recognized.

This method shows much worse results in the incorrectly identified geographic names – the average percentage is 72.05%. The system has found almost as many incorrect objects as correct ones. This happens because many cities have lakes or other object names which at least partly match the name of the city (for example, a search for "Riga" also found "Riga Reservoir").

In the experiments, where the system was set to link the whole text to a one most appropriate geographic object, in 11 cases out of 20 the system perfectly identified the object using both methods. In two cases out of 20 one of the methods was wrong with the object choice. In three cases, the methods identified only partially correct answer but they can also be counted as correct answer because the task was to link the text to only one object while human would actually link the articles to two or more objects (and one of them the system identified correctly). In three cases, the system identified the object incorrectly. In one case, the system identified the village Abava while human would link this article to the river Abava (note that river names are not included in the used gazetteer). In one case, the system linked text to one particular city while a human could not link this article to any particular geographic object at all. To sum up all the results, the average percent of correct linking is equal to (11+1+3)/19 = 0.7895 or 78.95% but, taking in account that the database did not include information about rivers, the rate is (11+1+3)/18 = 0.8333 or 83.33%.

In order to quantify the differences in efficiency of the two word matching methods even more clearly, contingency tables were constructed (Table 2). From the tables, measures of relevance can be calculated, namely Precision (the fraction of retrieved instances that are relevant) and Recall (the fraction of relevant instances that are retrieved). For exact match method Precision is TP / (TP + FP) = 84.33% and Recall is TP / (TP + FN) = 80.46%. For partial match method Precision is 85.54% and Recall is 61.21%. As can be seen, Precision for this dataset for both methods is almost the same while for the partial match method Recall is considerably lower. The partial match method retrieves too many irrelevant results and therefore its usage is not recommended.

Table 2. Contingency table for recognition results of a) exact match method and b) partial match method

a)	Geographic objects linked by system	
Geographic objects to be linked	True Positive TP = 70	False Positive FP = 13
	False Negative FN = 17	True Negative TN = 66

b)	Geographic objects linked by system	
Geographic objects to be linked	True Positive TP = 71	False Positive FP = 12
	False Negative FN = 45	True Negative TN = 38

Conclusions

The developed approach is a combination of ideas mentioned in previous researches of automated data extracting. The empirical evaluation of the developed geoparsing system prototype (on the example of Latvian language) showed that the average quality of its geographic name recognition varies around 75-80% if one takes into account the incompleteness of the gazetteer database used in experiments, while the percentage is close to 100% if one considers only the place names stored in the database. This suggests that for practical applications the developed geoparser could give sufficiently high degree of recognition but the quality of the used gazetteer database can significantly worsen this result. Increase in the amount of news articles used in experiments would increase the accuracy of the evaluation but the results would likely remain within the existing limits.

In order to improve the quality of place name recognition, it is necessary to improve the quality and completeness of the gazetteer database (in some cases the prototype was not able to recognize the place name, because information about it was not stored in gazetteer). It could also be helpful to add list of person name "stop words" as during the recognition process some of names and surnames mentioned in the articles were identified as geographic names. Also it would be helpful to enhance the gazetteer database with such large-scale objects as countries and municipalities and such small-scale objects as streets.

Future studies may include improvements of place ranking heuristics, improvements of gazetteer completeness, and comparisons of the system to other existing geoparsing systems.

References

Abascal-mena, R., López-ornelas, E., 2009. Structured Data Analysis of Travel Narratives by Using a Natural Language Processing Tool. *In: Proceedings of the IADIS International Conference on WWW/Internet.* 19-22 November 2009 Rome. Rome: IADIS, 242 – 246 p.

Caldwell, D., 2009. *Geoparsing Maps the Future of Text Documents.* Directions Magazine. Available from: http://www.directionsmag.com/articles/geoparsing-maps-the-future-of-text-documents/122487 [Accessed 11 April 2012].

Goldberg, D.W., 2008. *A Geocoding - Best Practices Guide.* Springfield: NAACCR, pp. 261.

Keller, M., Freifeld, C., Brownstein, J., 2009. *Automated vocabulary discovery for geoparsing online epidemic Intelligence.* BMC Bioinformatics. Available from: http://www.biomedcentral.com/1471-2105/10/385 [Accessed 3 April 2012].

Mikheev, A., Moens, M., Grover, C., 1999. Named Entity Recognition without Gazetteers. *In: Proceedings of EACL'99.* Stroudsburg, 1-8 p.

Nikolajevs, J., 2011. Automated geocoding of textual data. *In: Proceedings of Research innovations fundamentals 2011.* Klaipeda: Klapedos Universiteto, 12 – 16 p.

Scharl A., Tochtermann K., 2007. The Geospatial Web: How Geobrowsers, Social Software and the Web 2.0 are Shaping the Network Society. London: Springer, 282p.

J. Nikolajevs graduated from Riga Technical University receiving bachelor's degree in 2011. The topic of his thesis was Automatic geocoding of textual data. In 2011 he was accepted into the Gold Fund of RTU graduating students. He now studies in Riga Technical University master's program. Currently his research interests include geocoding, geoparsing, geographic information systems, deductive databases, distance learning, concept maps.

G. Jekabsons graduated from Riga Technical University receiving master's degree in 2005. In 2009 he received a PhD at Faculty of computer systems and information technology, Riga Technical University. The topic of his thesis was Heuristic methods in multidimensional regression model building. The PhD thesis won the Verner fon Siemens excellence award in year 2009. His research interests include location-based services, geographic information systems, machine learning, and statistics.

AUTOMATINIS GEOGRAFINIO KONTEKSTO ATPAŽINIMAS NESTRUKTŪRUOTAME TEKSTE

Jurijs Nikolajevs, Gints Jekabsons

Santrauka

Šiame straipsnyje nagrinėjama geografinių duomenų analizės problema ir aprašomas geografinių duomenų analizės sistemos prototipo įgyvendinimas ir jo empirinis įvertinimas. Prototipas automatiškai analizuoja nestruktūruotą tekstą atpažindamas geografinę informaciją ir žymėdamas geografinius vardus. Geografinių vardų atpažinimo tikslumas naudojant šį prototipą siekia 75%-80%, jei eksperimentui yra naudojama nepilna vietovardžių duomenų bazė. O tuo atveju, kai atpažinimui ieškomi tik vietovardžiai įtraukti duomenų bazėje, tikslumas yra artimas 100%. Atlikto tyrimo rezultatai ir sukurtas prototipas tinkami taikyti automatiniam teksto grupavimui ir paieškai pagal vietovardžių geografinį kontekstą, pavyzdžiui naujienų portaluose.

Pagrindiniai žodžiai: geografinis kodavimas, geografinė analizė, geografinis kontekstas, natūraliosios kalbos apdorojimas.