Brazilian Journal of Cartography (2016), Nº 68/6, Special Issue GEOINFO 2015: 1123-1132
Brazilian Society of Cartography, Geodesy, Photgrammetry and Remote Sense
ISSN: 1808-0936

MINING INFLUENTIAL TERMS FOR TOPONYM RECOGNITION AND RESOLUTION

Minerando Termos Influentes para a Descoberta e Resolução de Toponímias

Caio Libânio Melo Jerônimo, Cláudio E. C. Campelo & Cláudio de Souza Baptista

Universidade Federal de Campina Grande – UFCG
Centro de Engenharia Elétrica e Informática – CEEI
Departamento de Sistemas e Computação – DSC
Rua Aprígio Veloso, 882 – Bairro Universitário – CEP 58429-900
caiolibanio@copin.ufcg.edu.br, {campelo,baptista}@dsc.ufcg.edu.br

Received on December 5, 2015/ Accepted on February 19, 2016
Recebido em 5 de Dezembro, 2015/ Aceito em 19 de Fevereiro, 2016

ABSTRACT

The detection of toponyms present in text has appeared as a useful resource for many different applications, such as for social network analysers and for geographic search engines. The variety of ambiguities present in the geoparsing process represents one of the main challenges related to the process of detecting toponyms, bringing the need for treating this problem with careful attention. One important technique to detect toponyms is based on the presence of influential terms, which are terms that could indicate the existence of geographical references in the text. This paper presents an approach to automatically identifying relevant influential terms for a given language, as well as a set of attributes relating these terms with toponyms. The technique presented here was validated with an existing geoparser, using a training set based on online news. The results indicate the technique is effective in identifying influential terms, and has shown that the geoparser's capabilities of detecting toponyms have improved by using the generated list of influential terms.

Keywords: Toponym Detection, Toponym, Geoparser, Influential Terms.

RESUMO

As detecções de toponímias presentes em textos têm se mostrado um recurso importante em diferentes tipos de aplicações, como em análises de dados oriundosde redes sociais e também em sistemas de buscas geográficas. Os diferentes tipos de ambiguidades presentes no processo de geoparsing representam um dos principais desafios relacionados ao processo de detecção de toponímias, fazendo-se necessárias diferentes abordagens para o tratamento deste problema. Uma das técnicas mais importantes para a detecção de toponímias se baseia na presença de termos influentes, que consistem em termos que indicam a presença de referências geográficas em um documento. Este trabalho apresenta uma abordagem para detecção automática de termos influentes, bem como de atributos que possibilitam o relacionamento entre estes termos e suas toponímias. A técnica apresentada neste trabalho foi validada com um geoparser existente, utilizando-se um conjunto de treinamento formado por notícias colhidas da internet. Os resultados obtidos indicam que a técnica é eficiente na identificação de termos influentes relevantes, melhorando a capacidade de detecção de toponímias do geoparser utilizado no estudo.

Palavras chave: Detecção de Toponímias, Toponímias, Termos Influentes.

1. INTRODUCTION

Since the early days of the internet, many changes have occurred, especially in the way people perform searches on the network. In this scenario, new technologies related to Information Retrieval (IR) have emerged. It has been found that users often describe some kind of geographic context within their queries when performing a search in an information retrieval system (GAN *et al*. 2008). A study with the logs of the Excite search engine showed that one fifth of all queries were geographical (GUILLÉN, 2007). Thus, it is possible to realize the importance that the geographical context has in the current internet usage, which made possible the rise of new technologies related to geoparsing methods (JONES & PURVES, 2015; DHAVASE, 2014).

The geoparsing process consists of analysing documents, in order to find geographic references in it. The core difficulty associated with the geoparsing process consists of the different kinds of ambiguity associated with natural languages, including ambiguities related to toponyms (LEIDNER & LIEBERMAN, 2011). This kind of ambiguity refers to the cases where it is not possible to determine if a specific name is related to a geographic term or to another kind of reference, such a person's name. This kind of ambiguity can lead to undesired results, especially in geographic search engines, due to the fact that these systems rely on geographic references found in documents (by the geoparser) to satisfy the user's geographical query.

For the geographic recognition in text, there are three main families of methods: Gazetteer Lookup Based, Rule-Based and Machine Learning Based (LEIDNER & LIEBERMAN, 2011). The Gazetteer Lookup Based consists of analysing texts elements (words or characters) and search for this references in a predefined set of real geographic place names, in order to verify if the searched term exists in a predefined set. The Rule Based method uses a set of rules in a DSL (domain specific language) encoding decision procedures, allowing an interpreter to decide whether a word is a geographic term. The Machine Learning Based method basically consists in analysing texts in order to find specific patterns that could indicate a presence of geographic terms in the text, based in previously learned information.

One important type of pattern, that is useful to detect different types of named entities (NE) in text (including geographic ones) is related to the detection of influential terms (ITs). These terms generally appear near the named entity. For example, the term "city of" suggests that the next term probably refers to a city name, making this term a relevant case of influential term. Ratinov and Roth (2009) included this kind of feature in a set of "context aggregation features" to develop a Named Entity Recognition (NER) method that automatically detects token's context based on existent terms in a distance window, in order to determine their context in documents. Combined with other approaches, this technique shows itself useful in determining the context of a correlated NE. In a geoparser, this method helps determine the geographic scope of a term (city names, state names, street references), and also helps in the disambiguation process, allowing to decide if a reference is a real toponym or just a person's name, by applying a contextual meaning to the NE.

The objective of this research is to propose an innovative method of automatically detecting ITs for a given language. To achieve this goal, we developed a set of heuristics with the objective of learning the relations between geographical terms (toponyms) and other terms that could indicate the presence of geographical content in a document. The heuristics presented here have been implemented within an existing geoparser (CAMPELO & BAPTISTA, 2009), which is part of a search engine prototype, making possible to analyse the effectiveness of the presented method, based on the number of ITs detected and based on observed improvements of the geoparser effectiveness, in terms of corrected toponyms detected, false positives and false negatives.

To execute both the training and parsing process, we used Brazilian news from a major online newspaper (Globo G1 - www.g1.globo.com). News often have a strong geographic component, since this type of documents frequently have associations with the place where the readers live (LIEBERMAN & SAMET, 2011), making this kind of documents a good scenario for both training and geoparsing

processes. In order to analyse the results, we compared the detection rate between two cycles of training, collecting informations about the toponyms detected correctly and incorrectly. It was discovered a total of 1,211 ITs and, using these new terms, we observed the geoparser was able detect more toponyms in the analysed documents, with a p-value = 0.00001924.

This article is based on (JERÔNIMO *et.al.* 2015) previously presented at GEOINFO conference. The remainder of this paper is organised as follows. The next section presents related work. Section 3 presents our proposed approach to identifying influential terms. Then the experiments conducted to validate our approach is presented in Section 4. In Section 5 we have the discussion of the results. Finally, Section 6 concludes the paper and points to future directions.

2. RELATED WORK

Toponym recognition and resolution have been studied in different contexts, such as in information retrieval, social media geoparsing, geographic information systems and in a large variety of different applications where detecting geographic references could play a relevant role.

Many research papers related to this topic uses the idea of influential terms in toponym detection. Gelernter and Balaji (2013) presented a method of geo-parsing microtext with the objective to make these texts more readily usable for tracking news events, political unrest, or disaster response by providing a geographic overview. Their presented technique included the use of special cues (in, near, to, west, south) to identify possible location abbreviations in microtext.

Keller *et al.* (2008) describes an automated approach to discover geographic references taking the context into account, which relies on a window of words surrounding the word to parse, providing a generalization of the gazetteer's rule-based geoparsing. This approach has the objective of geo-parsing texts from media reports to track global disease outbreaks. Other methods also consider the context given by neighbor terms in geographic references analysis (RAUCH *et al.* 2003; AMITAY *et al.*, 2004).

Campelo and Baptista (2009) used a similar technique, where the occurrence of an influential term could increase the confidence that there is a geographic reference in a given document. Domingues and Eshkol-Taravella (2015) proposed a solution to detect toponyms in custom-made maps also using predefined patterns based on verbs, locative nouns and locative prepositions (e.g., to leave, departure, arrival, beside, alongside, close to).

Although there have been proposed many approaches to detecting toponyms that rely on the presence of influential terms in the geoparsing process, a common drawback of these existing works is that they do not describe precisely how the list of influential terms are built. Moreover, in most cases, the ITs are set manually by a user, rather than automatically detected by a learning process.

3. IDENTIFYING INFLUENTIAL TERMS

There are two important factors that should be considered while designing a mechanism for automatically identifying influential terms for toponym recognition: the type of places and the types of georeferences the geoparser can deal with. The former refers to the place types such as cities and states. Existing geoparsers normally define a hierarchy of place types that it can deal with (such as city → state → country). The latter refers to types of georeferences usually found in text that can be used to infer locality, such as postcodes, phone numbers or place names. While place names can be used in a text to refer to any place type, other georeference types may be mapped to specific place types (such as phone area codes, which are usually associated with countries or states). Influential terms, in turn, are usually employed in association with one or more place types or georeference types. Thus, we say that an IT can be mapped to specific tuples <georeference type, place type>. For example, "city of" may be an influential term for cities when it is referred by its place name, but not by a postcode.

As the solution developed in this research was implemented in an existing geoparser (CAMPELO; BAPTISTA, 2009), the influential terms identified were based on the types of georeferences and places it is able to recognize. The types of places the system is able to detect is based on a 5-level administrative hierarchy (i.e., city, microregion, mesoregion, state

and region). On the other hand, the system can process place names, phone numbers and postcode as georeference types. In order to preserve the capabilities of the validating geoparser, the method presented here will only allow the identification of ITs related to the types of georeferences and the types of places that are currently supported by the system. However, it should be highlighted that our proposed solution is general enough to keep identifying additional ITs as the geoparser improves its capabilities.

The aim of our method is to generate a dataset consisting of a table containing all the influential terms identified and a set of related attributes, as follows:

- Term: an identified IT. There may be multiple rows in the table for a given IT, each of which associated with different attributes.
- Distance: distance in text (in number of words) from the IT to the correlated toponym.
- Place Type: the place type of the toponym that the IT is associated with (city = 1, microregion = 2, mesoregion = 3, state = 4 and region = 5);.
- Georeference Type: the type of the georeference that the IT is associated with (place name = 1, telephone reference = 2 and postcode reference = 3).
- Confidence: the calculated confidence for the IT. This is a number between 0 and 1 that quantifies the influence of a term to the tuple <georeference type, place type> when the distance between them in the text is equals to the value given in the field Distance.

Another important challenge in developing a learning based method of identifying influential terms is that it leads to a problem like the classic "the chicken or the egg" dilemma: the identification of influential terms must rely on known toponyms present in the text. However, for training a corpus containing thousands of news, annotating those toponyms manually would not be feasible. On the other hand, for automatically detecting toponyms, it is crucial to apply heuristics based on the presence of influential terms. We have overcome this challenge by executing the previous version of our geoparser on the training set for detecting toponyms. Afterwards, we could apply our approach to identifying ITs. This process is illustrated in Figure 1 (the training variables will be described later in this section).

The previous version of our geoparser relies on an IT table containing just 32 rows manually inserted based on human observations. Nonetheless, one could argue that such geoparser would detected a significant number of false positives, which would consequently affect the training process for identifying ITs. However, in this geoparser, each detected toponym is associated with a confidence value (from 0 to 1), and only those above a certain threshold is accepted. In a previous work (CAMPELO; BAPTISTA, 2009), we found that the parser performs reasonably well (less than 30% false positives) for a threshold of 0.5. Thus, in this task, we increased this threshold to 0.6, making the results much more reliable (with approximately less than 10% false positives). The counter effect of this is that less toponyms are detected (more false negatives), which would not be acceptable for a geographic search engine, for example. However, in this case, the decrease in the number of toponyms detected per document does not affect the efficacy of the training process for identifying ITs, as the training set of toponyms is still large enough for this task.

3.1 The It Identification Algorithm

The process of identifying ITs is showed in Figure 1 (last box).

The Figure 2 illustrates process of identifying ITs with more details.

As described above, our algorithm identifies toponyms using the previous version of the geoparser. After this process, the mechanism stores information about their preceding terms (Figure 1 - box 3), called IT candidates, along with information about the correlated toponym, such as: the distance between the preceding term and the toponym; the toponym classification (georeference type and place type). A rule implemented in this stage asserts that each IT candidate will be associated with the nearest toponym only. In other words, the rule ensures that there will be no other toponym between an IT and its related toponym. By implementing this rule, we observed a significant decrease in the number of false positives for IT identification, that is, the cases where identified ITs were not syntactically related to the correct toponym. In this process, if a repeated IT candidate is found, a counter associated with this IT is incremented,

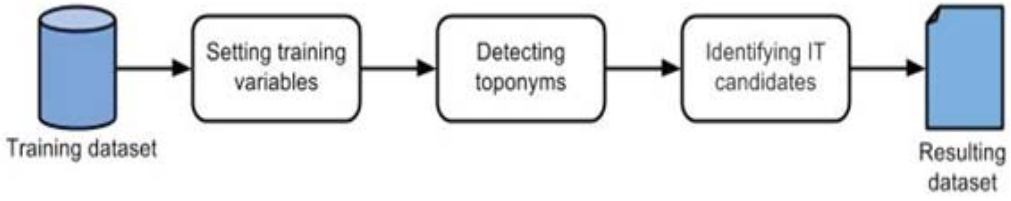

Fig. 1 - The flow process of the training algorithm.

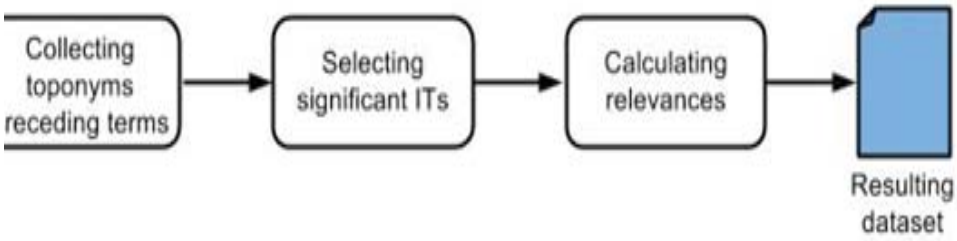

Fig. 2 - The process of Identifying Its.

which represents the number of times that the term is found in correlation to a toponym, for the same classification and distance.

The collection of preceding terms depends on a training variable called MAX_DISTANCE. This variable represents the maximum distance that the algorithm should consider when looking for influential terms, always in backwards direction from the toponym. For example, in the text fragment "The next olympics will be held in the city of Rio de Janeiro", if the toponym was "Rio de Janeiro", and the value for this variable was 3, the terms selected for composing IT candidates will be "the", "city" and "of".

Our approach allows the identification of ITs composed by more than one term, such as "south of" and "close to". For that, another training variable should be considered: MAX_COMBINATION. This variable expresses the maximum number of words that should be combined to form IT candidates (i.e., the maximum length of an IT). For the example of the text fragment given above, if the value of MAX_COMBINATION was 2, the IT candidates generated would be "the", "city", "of", "the city" and "city of". We found by empirical analysis that MAX_DISTANCE = -6 and MAX_COMBINATION=3 provide best results. Hence, these are the values adopted to perform our experiments. In this paper, we refer to ITs composed by just one term as "atomic", whilst those made by two or more terms are called "composed".

After acquiring IT candidates, the algorithm performs a selection process over these terms (Figure 2 - box 2) in order to reject terms that are not in conformance to the predefined rules. This rules consists in rejecting terms with less than two characters in length, terms formed by numbers, terms matching a specified set of stop words. These rules have shown to be effective in removing candidate terms that could decrease the quality of IT identification. Moreover, in this stage, the selected ITs must be in conformance with another training variable called ACCEPTANCE_THRESHOLD.This variable represents the minimum number of occurrences of an IT candidate in the training dataset. That is, if the number of times an IT appears in the training set is less than the value for this variable, it will be rejected and its confidence value will not be calculated. We adopted ACCEPTANCE_THRESHOLD = 3 in our experiments.

After the selection process, the remaining ITs have their confidence value calculated based on the frequency that each one appeared in the training data (Figure 2 - box 3). The IT confidence is an important parameter when performing the geoparsing process using the discovered influential terms. This value can inform the geoparser how significant the IT is when it appears preceding a specific toponym, in a predefined distance. In order to calculate the IT's confidence, the solution takes into account

the frequency of each influential term, based on the number of occurrences collected during the training process.

For atomic terms, the algorithm also considers the number of occurrences that the IT appears in correlation to non-geographic references, in order to decrease its confidence value. This approach aims to reduce the false positive cases for toponym detection due to a significant reduction of the ITs confidence that are frequently not only vinculated to geographic references. This step is crucial for calculating ITs relevantes, as there are many terms that are frequently used to refer to a toponym, but are also frequently used in other contexts, such as the term "in". The approach to calculating the confidence value for atomic ITs is shown in Equation 1.

$$R_a = \frac{PCount^2}{PCount + NCount} \quad (1)$$

- R_a represents the calculated confidence for an atomic IT.
- $PCount$ represents the occurrence counter for each IT found, in the cases where this term was associated with a toponym.
- $NCount$ represents the occurrence counter for each IT found, in the cases where this term was not associated with a toponym.

This is important to highlight that the system does not count occurrences of ITs (atomic or composed) that are also part of other longer composed ITs, such as the terms "in" and "north" which are both part of the term "in the north of". This changes significantly the way ITs and their attributes are identified, and we observed that the geoparser efficacy improves considerably for detecting toponyms. For example, terms like "north" and "south" are quite frequent, and could be stored in association with distance -2, such as in "north of London". However, as the term "north" rarely appears without the preposition "of", it could lead the system to detect false toponyms where any term is between it and the IT, such as in "...they went to the north. In London, people...". Thus, by considering this rule, the atomic term "north" is discarded and the term "north of" is stored in association with distance "-1". The approach to calculating the confidence value of composed IT is shown in Equation 2.

$$R_c = \frac{CCount}{CSize} \quad (2)$$

- R_c represents the calculated confidence for a composed IT;
- $CCount$ represents the occurrence counter for each composed IT found in correlation with a toponym;
- $CSize$ is the number of composed ITs found during the training stage;

In both cases (and) the values are normalised by the algorithm before completing the process and storing the confidence values in the ITs table.

In Figure 3, we present the general idea of the proposed algorithm for mining influential terms in the form of pseudo-code.

4. EXPERIMENTAL EVALUATION

This section presents the experiments conducted to validate our proposed approach.

4.1 Experimental Units

The performed experiments were based on web news written in Portuguese, from a Brazilian communication vehicle (http://g1.globo.com/). We decided to use this type of document because news usually have strong relations to geographic areas, since these documents often refer to the place where the readers live (e.g., city, state, country), making these documents an excellent dataset of geographic references, both to the training and the parsing processes. The news were collected in the year of 2015 by an automated tool, developed to read an RSS feed and extracting their related news. These news were collected from the "last news" category, due to this kind of subject could bring us a large number of toponyms references, as this documents are frequently associated with a specific place or region.

4.2 The prototype used

To validate the proposed solution, the methods were implemented in an existing geoparser, that is part of a search engine prototype, called GeoSEn. This is a geographic search engine, with the objective of retrieving

```
1. miningInfluentialTerms(trainingNewsList)
2.   n = trainingNewsList.size()  // size of the training set
3.   listIT = []  //list containing the influential terms founded
4.   for (i in n)
5.      listPT = [] //list containing the toponym' precedent terms
6.      //collecting the toponyms from an individual news article
7.      listToponyms = trainingNewsList.getNews(i).getToponyms();
8.      for (tp in listToponyms) //collecting precedent terms from toponym
9.         listPT.addAll(tp.getPrecedentTerms());
10.     for (pt in listPT) //adding each precedent term to the list of ITs
11.        if(! listIT.contains(pt)) listIT.add(pt)
12.        else listIT.get(pt).incrementCouter();
13.  for (itCandid in listIT) //removing precedent terms according to rules
14.     if(! accordingtoRules(itCandid))itCandid.markToRemove();
15.  listIT.removeMarkedTerms();
16.  for (it in listIT)  //calculating confidence
17.     if (it.isAtomic()) it.setConfidence(calculateAtomicConf(it))
18.     else it.setConfidence(calculateComposedConf(it))
19.  return listIT;
```

Fig. 3 - General idea of the proposed technique

web documents based on their geographic scope, usuallyspecified as a parameter in the user's queries. As Campelo and Baptista (2009) describe, GeoSEn's geoparser implements a set of heuristics to detect geographic references in web documents, and the presence of influential terms is one of these heuristics. Each heuristic has a specific weight for the final value of confidence rate. This confidence value is assigned to the toponyms found in the text (by querying a local Gazetteer). In that geoparser, the ITs and their associated attributes must be informed manually by the system administrator.

Structural adaptations were performed in the system's geoparser, with the objective of implementing the entire training phase showed in Figure 1. As part of our implementation strategy, we used the original geoparser with all of its capabilities. Then, by implementing the method of discovering ITs proposed in this research, we obtained a new geoparser with the ability of detecting a large range of toponyms. The original geoparser was used to detect toponyms (Figure 1 - box 2), making possible to execute the first learning cycle based on this initial set of detected toponyms, generating the first set of discovered influential terms (Figure 1). Then, this set of detected influential terms can be used to detect toponyms more reliably, which can be the basis for further executions of the training cycle (i.e., discovering ITs based on more reliable toponyms), allowing an incremental process of learning influential terms.

4.3 Design of Experiments

The experiments performed in this research have the objective of answering the research question Q, as follows:

- *Research question (Q)*: Have the discovered influential terms improved the toponyms detection rates?
- *Null hypothesis (H-0)*: The discovered influential terms does not improve the toponyms detection rates.

To answer this question, our proposed experiment consists in training the parser initially with 1,000 aleatory news (set D1) and, after the training process, executing the trained geoparser with a test set of 5,000 news documents. After this first experiment step, the process was repeated with 9,000 news as a training set, totaling 10,000 news documents (set D2).

For sets D1 and D2, the system returned a copy of the analysed documents containing the detected and rejected toponyms coloured in green and red respectively, as HTML files. An example of this coloured output file is shown in Figure 4.

We selected an aleatory subset of 100

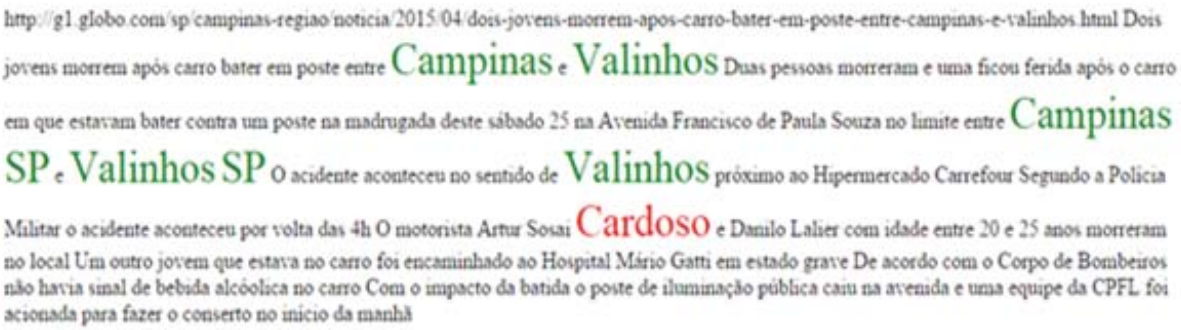

Fig. 4 - Example of a coloured output file.

of such coloured documents that showed differences between the number of accepted and rejected toponyms, totalizing 50 documents for each D1 and D2 (paired analysis), and submitted them for human examination by a group of volunteers, in order to judge the quality of the toponyms detection, in terms of correct detections (true positives), correct rejections (true negatives), false negatives and false positives. These categories of detection were considered separately for a better statistical analysis. The main steps of this experiment are summarised as follows:

1. Training the *geoparser* with 1,000/10,000 news;
2. Executing the *geoparser* process with a test set of 5,000 aleatory news for D1 and D2 cases;
3. Analysing (manually) the output consisting of coloured HTML files, to collect statistics about correct toponym detection, correct toponym rejection, false negatives and false positives;
4. Comparing the results between the two training cases (1,000 and 10,000).

5. RESULTS AND DISCUSSION

For our proposed research question Q, the experiments were executed in a paired design, and the collected data did not show a normal distribution. Due to these characteristics of data, we chose to use the Wilcoxon test in order to analyse a possible improvement in toponyms detections for the training dataset of 1,000 (D1) and 10,000 (D2), respectively. After this first experiment, we reached the following results in Table 2, with a total of 1,211 ITs found.

Table 2: Results of Wilcoxon test

Wilcoxon (α=0.05)	H != 0 (p-value)	H > 0 (p-value)	H < 0 (p-value)
Correct detections	0.00003848	1	0.00001924
Correct rejections	-	-	-
False negatives	0.00002601	0.000013	1
False positives	-	-	-

This result shows a significant difference in toponyms detection between the two dataset sizes. The difference can be shown firstly by the case where H != 0. Here, we rejected the hypothesis that D1 and D2 would have the same rate of correct detections (p-value = 0.00003848). Still with regard to correct detections, we have the H < 0 case, where p-value = 0.00001924, indicating that D2 detected a higher number of correct toponyms in relation to D1 dataset size.

Regarding false negatives, it can be noticed a difference between D1 and D2 in case H != 0(p-value = 0.00002601), allowing us to reject the hypothesis that D1 and D2 would have the same correct rejection rate. There was still a significant change related to H > 0 case (p-value = 0.000013), denoting that the system presented a higher false negative rate when trained with D1 (in comparison to the D2 training dataset).

With reference to correct rejections and false positives, the results obtained for both D1 and D2 were almost the same for most cases, what could result in a low precision reported from the Wilcoxon test, due to the fact that this particular test depends on pairwise differences between the samples. This characteristic led us to do not execute the test for this two analysed values.

An additional experiment case was conducted, aiming to execute two incremental training cycles with the same training dataset. This experiment consisted in training the geoparser initially with 20,000 different news and collecting the number of discovered ITs. After that, we executed the training system again with the same training dataset, and collected the final number of detected ITs. Our objective was to determine whether the system could detect more ITs even with the same training documents. The results are shown in Table 3.

Table 3: Results for the second experiment

Parameter	geoparser (20000)	geoparser (40000)
Nº of discovered IT	1806	3014

This results shows a considerable increase in the number of detected IT as the training dataset increases, denoting that the system learned ITs that could not be learned in the first training case (with 20000 news). After the proposed experiments, we can answer the question Q by rejecting the null hypothesis H-0, with p-value = 0.00001924, as this value indicates that the learned ITs increased the number of correct toponyms detected. The experiment also indicates a lower rate of false negatives associated with the new IT detected (p-value = 0.000013).

6. CONCLUSION AND FUTURE WORK

This paper presented an approach to automatic discovery of influential terms from text. The solution was implemented in a geographic search engine prototype called GeoSEn, with the objective of validating the proposed methodology in a geographically oriented system. The system's geoparser has been adapted to learn the influential terms from a training dataset and to report statistics of the execution of the parsing process, making possible to perform further analysis of the obtained results. Our results indicate that the proposed algorithm performed considerably well for automatically detecting ITs, as well as indicate that the geoparsing efficacy improves significantly when these new influential terms are used for detecting toponyms.

The methodology presented was validated with Brazilian news, written in portuguese. However, there are evidences that it can be extended for many other languages without further modifications of the parser's code, which is intended to be verified in future work.

We believe that other relevant ITs could be identified if the parser is executed for other types of texts, such as more informal texts from social networks. This is also planned to be checked in future work.

REFERENCES

AMITAY, E.; HAR'EL, N.; SILVAN, R.; SOFFER A. Web-a-where: Geotaggingweb content. In **Proceedings of SIGIR, Workshop on Geographical Information Retrieval**, pages 273–280, 2004

CAMPELO, C. E. C.; BAPTISTA, C. S. A model for geographic knowledge extraction on web documents. In: **Advances in Conceptual Modeling-Challenging Perspectives** (pp. 317-326). Springer Berlin Heidelberg, 2009.

DOMINGUES, C.; ESHKOL-TARAVELLA, I. Toponym recognition in custom-made map titles. In: **International Journal of Cartography**, v. 1, n. 1, p. 109-120, 2015.

DHAVASE, N.; BAGADE, A. M.. Location Identification for Crime &Disaster Events by Geoparsing Twitter., In: **Convergence of Technology (I2CT), International Conference**, pages 1 – 3, 2014.

GAN, Q., ATTENBERG, J.; MARKOWETZ, A.; SUEL T. Analysis of geographicqueries in a search engine log,In: **Proceedings of the first international workshop on Location and the web ACM.** (pp. 49-56), 2008.

GELERNTER, J.; BALAJI, S. An algorithm for local geoparsing of microtext. In: **Geoinformatica,** Volume 17, Issue 4, pp 635-667, 2013.

GUILLÉN, R. GeoParsing Web Queries. In: **Advances in Multilingual and Multimodal Information Retrieval Lecture Notes in Computer Science** Volume 5152, pp 781-785, 2008.

JERÔNIMO, C. L. M; CAMPELO, C. E. C.; BAPTISTA, C. S. Mining influential terms for

toponym recognition and resolution, In: XVI **Brazilian Symposium on Geoinformatics,** Campos do Jordão, 2015. p 143-154. 2015.

JONES, C. B.; PURVES, R. S. GIR 2014 workshop report: the 8thACM SIGSPATIAL International Workshop on Geographic Information Retrieval., In: **International Conference on Advances in Geographic Information Systems.** 6.3, 52-52, 2015.

KELLER, M., FREIFELD, C.C., BROWNSTEIN, J.S. Expanding a Gazetteer-basedApproach for Geo-Parsing Text from Media Reports on Global Disease Outbreaks. In: **Advances in Disease Surveillance**, Vol. 5, No. 3, 2008.

LEIDNER, J. L.; LIEBERMAN, M. D. Detecting geographical references in the form of place names andassociated spatial natural language. In: **International Conference on Advances in Geographic Information Systems**, 3(2):5–11, 2011.

LIEBERMAN, M. D.; SAMET, H. Multifaceted toponym recognition forstreaming news. In: **Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval**, pages 843–852, Beijing, China, 2011.

RAUCH, E.; BUKATIN, M.;BAKER K. A confidence-based framework for disambiguating geographic terms. In **Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies 2003 workshop on Analysis of geographic references**-Volume 1 (pp. 50-54).