Extraction of geographic entities from biological textual sources

Moisés A. Acuña-Chaves
Escuela de Computación
Instituto Tecnológico de Costa Rica
Cartago, Costa Rica
Email: moises@acuna-chaves.com

José E. Araya-Monge
Centro de Investigaciones en Computación
Instituto Tecnológico de Costa Rica
Cartago, Costa Rica
Email: jaraya@itcr.ac.cr

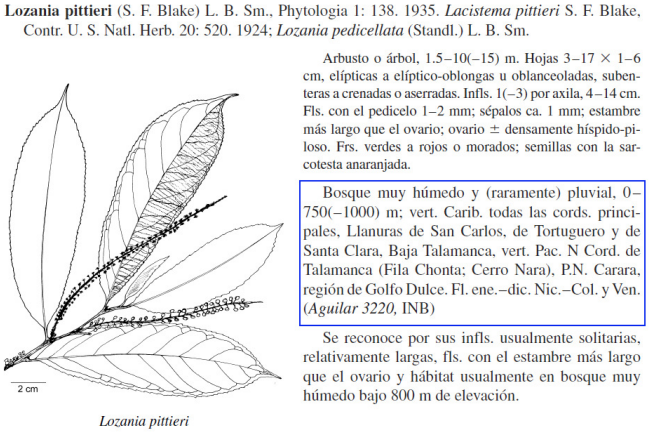
Abstract*— This work is focused on the exploration and application of entities extraction techniques for the codification and identification of geographical locations present in the geographic distribution section within botanic documents, such as the plant species manual of Costa Rica. Several technologies must be combined to achieve such objective, among them is Natural Language Processing (NLP) that helps in the extraction of entities with the usage of gazetteers. Another technology is the usage of rules (regular expressions, Deterministic Automata, context-free grammars). Additional to the identification and codification, an algorithm to bind the place names extracted to authorized sources such as gazetteer is presented. This algorithm identifies and enriches the entry text with extra information, extracted from the paragraphs where the distribution is defined in a semi unstructured text. The values of interest for this work are: world and Costa Rica distribution. After those values are identified, the information can be processed and become useful for diverse applications, such as geographic information systems. Other research projects might be interested in the results of this project. The evaluation consists in manually judging randomly selected sample of the results to establish if the algorithm yields useful data. The judgment features the evaluation of the world and Costa Rica distribution using the source context, given 3 possible values: GOOD, BAD, UNKNOWN. The ideal is to have the least BAD percentage. The algorithm is relatively good to geo-code and bind the world distribution. More work needs to be done for the Costa Rica distribution.

1. Introduction

The geographic distribution of a species is considered an important part of its biological description: in fact, several paragraphs are frequently devoted to detail it.

Scientists (botanists and ethnobotanists, among others) can use this geographic data as part of their research. Having these data available automatically or semi-automatically would save them time in collecting information. Examples of these investigations are bioprospecting of species in Costa Rican territory [1] and studies related to the biodiversity patterns in general [2] and [3]

Figure 1 highlights the geographical description for the species *Lozania pittieri* taken from the Manual de Plantas of


Lozania pittieri (S. F. Blake) L. B. Sm., Phytologia 1: 138. 1935. *Lacistema pittieri* S. F. Blake, Contr. U. S. Natl. Herb. 20: 520. 1924; *Lozania pedicellata* (Standl.) L. B. Sm.

Arbusto o árbol, 1.5–10(–15) m. Hojas 3–17 × 1–6 cm, elípticas a elíptico-oblongas u oblanceoladas, subenteras a crenadas o aserradas. Infls. 1(–3) por axila, 4–14 cm. Fls. con el pedicelo 1–2 mm; sépalos ca. 1 mm; estambre más largo que el ovario; ovario ± densamente híspido-piloso. Frs. verdes a rojos o morados; semillas con la sarcotesta anaranjada.

Bosque muy húmedo y (raramente) pluvial, 0–750(–1000) m; vert. Carib. todas las cords. principales, Llanuras de San Carlos, de Tortuguero y de Santa Clara, Baja Talamanca, vert. Pac. N Cord. de Talamanca (Fila Chonta; Cerro Nara), P.N. Carara, región de Golfo Dulce. Fl. ene.–dic. Nic.–Col. y Ven. (*Aguilar 3220,* INB)

Se reconoce por sus infls. usualmente solitarias, relativamente largas, fls. con el estambre más largo que el ovario y hábitat usualmente en bosque muy húmedo bajo 800 m de elevación.

2 cm

Lozania pittieri

Figure 1: *Lozania pittieri* geographical distribution sample paragraph.

Costa Rica [4]. Costa Rica has developed, for many years, a culture of conservation that is recognized worldwide. It is a privileged nation with a high percentage of global biodiversity. In particular it has many species of the kingdom Plantae. So it is not suprising that many scientists, have studied for years the diversity of Costa Rican flora.

The knowledge generated by some public and non-governmental institutions is mostly found in paper: books, magazines, articles and gray literature. Another great part of the knowledge is found in unstructured electronic documents that do not make the relationships between the concepts of the subject area computationally explicit.

One component of this knowledge is geographical distribution of a species. Leveraging this knowledge on a large scale is difficult unless suitable structures are used for processing and analysis by experts. Paragraphs referring to geographic distributions contain data in a language that is unsuitable for automatic processing. Although it has a general structure, it remains a natural jargon for the biological sciences.

Different applications have taken different approaches to solve the problem of identifying geographical points in texts (geo-parsing). Leidner and Lieberman [5] explain three

types of methods for geo-parsing: gazetteers, rule-based and machine learning.

Other systems for extracting geographical points (both proprietary and open source) include the C & C tagger [6](Machine Learning), Apache OpenNLP [7], which is a Java API, OpenCalais from Thomson Reuters [8] and the ANNIE module that is part of the GATE framework [9].

The objective of this article is to present the results obtained when implementing geo-parsing and geo-coding techniques for extracting geographic entities in biological descriptions. Section 2 explains the general scheme of the algorithm, its parts and the experimental design. Sections 3 and 4 present the experimental results and the conclusions, respectively. Finally Section 5 presents future work.

2. General outline of the algorithm

2.1. Architecture Design

Figure 2 presents the overall architecture design of the process developed. The input consists of XML files containing geographical distribution texts about species, genus and families of interest. A first step consist in a module responsible for the analysis of paragraphs: tokenization, parsing and PoS labeling. The following module takes this analysis, divides the paragraphs and finds the possible geographical points, geo-parsing. The next module takes these possible geographical points and try to link them with entries in a gazetter, geocoding. To match the entries, the geocoding module uses a search engine (Apache Solr [10]) where the gazetteer was previously stored. The output is a new collection of XML files with geographical distribution data explicitly tagged. Finally, to store the gazetteer in the search engine, an application platform called Aspire Comunity Edition [11] was used. During the geo-parsing processing, ids were generated for the possible geographical points extracted. These ids were later used to collect random samples of the geographical points in order to evaluate the effectiveness of the entire process. The following sections describe the components of this process in further detail.

2.2. Input and Output

The input consists of XML files containing information about different taxa (species, genus nad families) in an XML element named snippet (see Code 1). The element text contains the geographical description of the taxon.

```
<snippet>
   <category>
      <id />
      <name>Distribucion</name>
   </category>
   <taxon>
      <id />
         <family>Myrtaceae</family>
         <genre>Eugenia</genre>
         <species>Eugenia austin-smithii</species>
   </taxon>
   <text>Bosque muy humedo, pluvial, nuboso y enano, 600-2100x metros; vertiente
         (del) Caribe, Cordillera de Guanacaste, ambas vertientes Cords. de
        Tilaran y Central, Cerros de La Carpintera, vertiente (del) Pacifico(a)
         N Cordillera de Talamanca, Tablazo. Fl. marzo-junio, octubre Costa
        Rica y Panama (Haber & Bello C. 2434, MO)</text>
</snippet>
```

Code 1: Input snippet sample.

The input is processed grammatically and semantically in order to extract smaller data elements which contains data for the distribution of a species in Costa Rica and in the World. The output of this process is an enriched XML file. Code 2 shows the result of processing the snippet shown in Code 1.

```
<snippet>
   <category>
      <name>Distribucion</name>
   </category>
   <distribucion-cr>
      <codigo>DESC</codigo>
      <id>6582</id>
      <latitud>0.0</latitud>
      <longitud>0.0</longitud>
      <nombre>vertiente de el Caribe</nombre>
   </distribucion-cr>
   <distribucion-cr>
      <codigo>DESC</codigo>
      <id>6583</id>
      <latitud>0.0</latitud>
      <longitud>0.0</longitud>
      <nombre></nombre>
   </distribucion-cr>
   <distribucion-cr>
      <codigo>DESC</codigo>
      <id>6584</id>
      <latitud>0.0</latitud>
      <longitud>0.0</longitud>
      <modificador>cerca</modificador>
      <nombre>de la Division Continental</nombre>
   </distribucion-cr>
   <distribucion-cr>
      <codigo>3621368</codigo>
      <id>6585</id>
      <latitud>9.5</latitud>
      <longitud>-83.66667</longitud>
      <modificador>E</modificador>
      <nombre>Cordillera de Talamanca</nombre>
      <original>Cordillera de Talamanca.</original>
      <tipo>MTS</tipo>
   </distribucion-cr>
   <ditribucion-mundo>
      <codigo>3624060</codigo>
      <id>3023</id>
      <latitud>10.0</latitud>
      <longitud>-84.0</longitud>
      <nombre>Republic of Costa Rica</nombre>
      <original>Costa Rica</original>
      <tipo>PCLK</tipo>
   </ditribucion-mundo>
   <ditribucion-mundo>
      <codigo>3703430</codigo>
      <id>3024</id>
      <latitud>9.0</latitud>
      <longitud>-80.0</longitud>
      <modificador>O</modificador>
      <nombre>Republic of Panama</nombre>
      <original>Panama.</original>
      <tipo>PCLK</tipo>
   </ditribucion-mundo>
   <elevacion>2300-2750metros</elevacion>
   <especimen>(Davidse et al. 29046, INB)</especimen>
   <floracion>agosto</floracion>
   <floracion>setiembre</floracion>
   <taxon>
      <family>Alstroemeriaceae</family>
      <genre>Bomarea</genre>
      <species>Bomarea suberecta</species>
   </taxon>
    <text>Bosque de roble, 2300-2750 metros; vertiente&lt;pp&gt;. (del) Caribe
          , y cerca de la Division Continental, E Cordillera de Talamanca.
Flor agosto, set. Costa Rica y O Panama&lt;pp&gt;. (Davidse et al. 29046, INB)<
      /text>
   <zonas-holdridge>
      <codigo>DESC</codigo>
      <id>3353</id>
      <nombre>Bosque de roble</nombre>
   </zonas-holdridge>
</snippet>
```

Code 2: Sample of algorithm output snippet

2.3. Geo-parsing

Geo-parsing is defined as the process in which words are recognized as places, according to Kimler [12]. In his thesis, he uses several of the heuristics proposed by Pouliquen et al [13], for example, gazetteers for names in capital letters.

The Algorithm 1 shows how the distribution in Costa Rica is extracted using Freeling [14] tags to identify important sections. The algorithm has a condition for elements

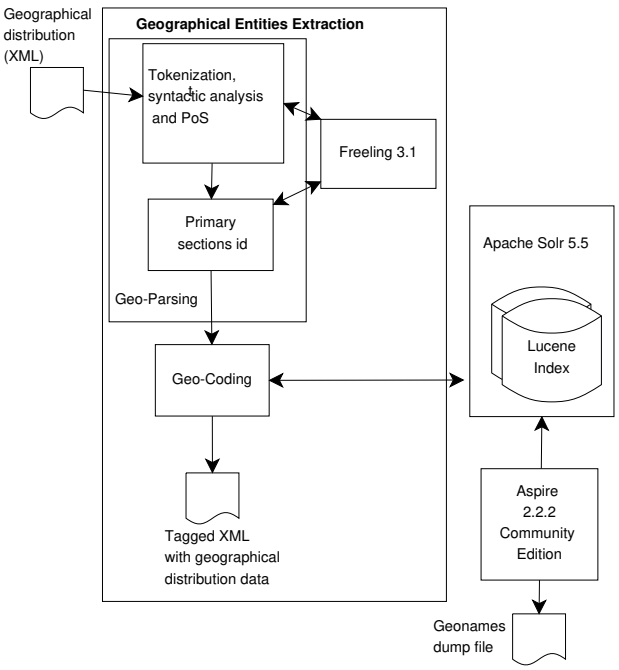


Figure 2: General software architecture that implements the geo-parsing and geo-coding algorithms.

with the labels of PoS point(Fp), comma (Fc) or semicolon (Fx). The same algorithm was also applied to identify sections of the distribution in the world. This process is described in Figure 2.

Algorithm 1 Processing and identification of the Costa Rica distribution input text.

$i \leftarrow ind$
for each word w in sentence0 **do**
  DistCRString $\leftarrow$ DistCRString + w
  if w.tag = 'Fp' OR 'Fc' OR 'Fx' **then**
    erase last in DistCRString
  end if
  $i \leftarrow i + 1$
end for
segments[distCR] $\leftarrow$ DistCRString

2.4. Geo-coding

Geo-Coding can be defined as the process of disambiguation and association of toponyms with actual locations.

The geo-parsing module may extract false toponyms. The list of names produced by the geo-Parsing stage contains only potential location names.

Once the possible toponyms are stored in a list of names, it is possible to consult the information of the chosen gazetteer to try to bind those names with entries in that gazetteer.

This procedure uses the index created by Apache Solr to search the gazetteer. Apache Solr provides an API that allows querying this index from programs using the Java language.

The queries take into account whether the search is about the distribution in Costa Rica or the distribution in the world, since modifications must be made to the query in order to increase the possibility of success.

Among the values that are obtained from the gazetteer are: ***geonamesid***, ***id***, ***lat***, ***lng***, ***nombre***. Where ***geonamesid*** is the identifier that links the value in the text with an entry in the gazetteer. The ***id*** field is generated sequentially for later use in tests. The fields ***lat*** and ***lng*** are the coordinates of the centroid of the entry in the gazetteer.

```
fq=+(alternate_country_code:CR OR
    country_code:CR) -feature_code:HTL -
    feature_code:EST
qf=name^10 asciiname^3 alternatenames^3
mm = 100 %
```

Code 3: Costa Rica distribution search parameters.

```
fq=+feature_code:(PCL* OR ADM* OR PRSH
    OR TERR OR ZN OR ZNB) +
    feature_class:A
qf=name^10 asciiname^3 alternatenames^3
bq=feature_code:ADM1^6
bq=feature_code:ADMD^6
bq=feature_code:PCLI^40
bq=feature_code:PCLD^30
bq=feature_code:PCLF^20
bq=feature_code:PCLH^20
bq=feature_code:ADM2^4
bq=feature_code:ADM3^2
bq=feature_code:ADM4
mm = 75 %
```

Code 4: Search parameters for worldwide distribution.

Code 3 shows the values of the search parameters for possible distributions in Costa Rica. The parameter ***fq*** defines the filters to be used, without affecting the relevance of the search. In this case GeoNames entries that have `country_code`=CR and `alternate_country_code`=CR are accepted, but entries for hotels and other locations are excluded. The qf parameter establishes the relative importance of the ***name***, ***asciiname*** and ***alternatenames*** fields in searches. As can be seen, much more importance is given to documents that *match* the ***name*** field. The ***alternatenames*** field of the gazetteer contains alternative names of a place and also names in other languages mainly for the original language. The `mm`=100% means that all the terms in the query must match the gazetteer entry.

Similarly as the query in Costa Rica above, the query for matches worldwide (shown in Code 4) uses the query fields (***qf***) parameter but the filtered query (***fq***) is different in the sense that it's forcing the results to start with ***PCL*** or ***ADM*** which means they're are Independent or Dependent Political

Entities or Administrative zones of one of them, ***PRSH*** (parrishes), territories or zones. Also the results must be of `feature_class`=A which means they're administrative boundary features, which means the entries are filtered to be countries or regions. Several boosting values are given to different `feature_codes`, with the most importance set to ***PCLI*** (Independent Political Entity). The `mm`=75% means that three out of four terms coming in the query must match in the resulting document, which makes the query more flexible than the query for the Costa Rica distributions.

2.5. Gazetteer

The gazetteer selected for this work must include names of places of Costa Rica and also names of regions and cities, geographical locations such as rivers, mountains, for both Costa Rica and the world. It should also contain additional information, such as latitude and longitude, or different administrative units (country, province, canton, district).

Two gazetteers with these characteristics were analyzed:

- NGA GEOnet Names Server (GNS) [15], contains in total about 10 million names from around the world. It has the drawback that it has no names for the United States or Antarctica. It has almost 6 thousand names that refer to Costa Rica.
- GeoNames [16], contains a number of names similar to GNS, but also contains names for the United States. It has almost the same number of names for Costa Rica.

After analyzing the completeness of the content of each Gazetteer, GeoNames was chosen, since it has in general more content in the archives (including the fact that it does contain names for the United States, important for the labeling of distribution in the world).

2.6. Feeding and data processing

The gazetteer file is processed using Aspire Community Edition [11], which is an application platform for processing and enrich structured and unstructured text from virtually any container. In this case the tab separated file from the GeoNames [16] site is processed using Aspire.

Figure 3 shows the design diagram of the feed and processing of each entry in the GeoNames file as it is indexed in an Apache Solr core. The input is a CSV file which path is set into the Filesystem Reader that then passes the Aspire Object of the element to the Tabular Subjob Extractor. The latter stage generates a subjob per row in the original CSV, and an Aspire Object attached to it. The Aspire Object contains the row data. Each subjob is then passed to a Subjob pipeline where the alternate names are expanded and then each row is sent as a document to Solr using the PostHTTP component.

2.7. Search Engine: Apache Solr

In Solr the GeoNames [16] entries are represented as documents of an index or *core*. For each index on the server

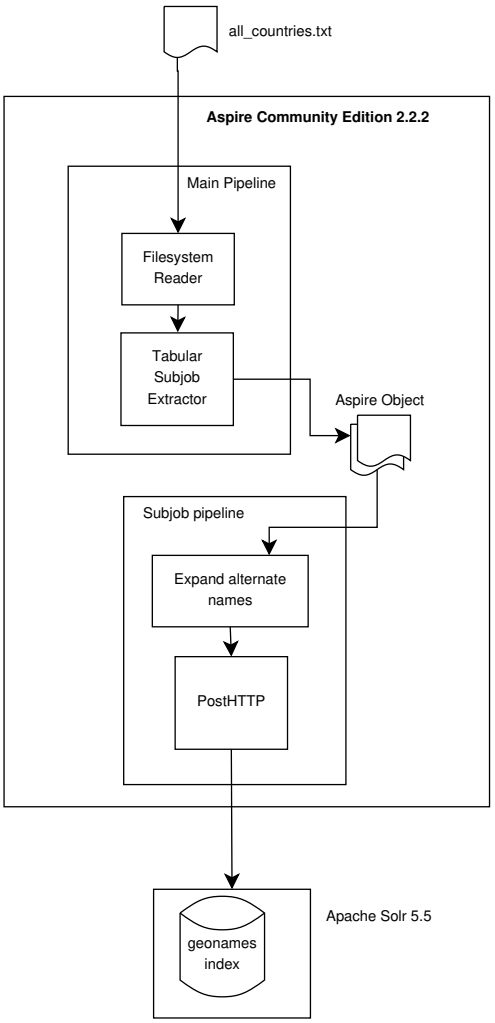


Figure 3: Aspire Community feed application design.

a schema specifies the fields, the types of those fields, and what linguistic analysis is applied to the text in each of the fields.

2.8. Experimental Design

This section presents how the performance of the algorithm for extracting geographic entities from biological textual sources was evaluated, specifically for the *Manual de Plantas de Costa Rica Vol. VI.* [4].

To evaluate the accuracy of the geo-coding and geo-parsing algorithms a manual approach was used. That is, a subset of the extracted toponyms is chosen at random and it is determined if they have been correctly extracted and if they have been correctly located in the gazetteer. A subset with 5% of the distribution toponyms of Costa Rica and the world was randomly selected. The random selection uses the identifiers of each distribution element in Costa Rica and worldwide generated during the different phases of the algorithms.

TABLE 1: Total number of clauses generated and number of clauses selected for distribution in Costa Rica and the world for samples of 5 %.

Type of distribution	Total number of clauses	Number of sample clauses (5%)
Costa Rica Distribution	6638	331
Worldwide Distribution	3052	152

Volume VI of the *Manual de Plantas de Costa Rica* could not be used for evaluation since it was used to develop and adjust the algorithm. For this reason, the volume V [17] of the *Manual de Plantas de Costa Rica* was used for evaluation tests.

The test subset was evaluated manually and each toponym in the subset was given one of three possible ratings: GOOD, BAD, and UNKNOWN (UNK).

To evaluate the effectiveness of the algorithms the percentages obtained for the three grades are calculated.

When an input file is processed, in addition to the XML output file with the tagged geographical distribution data, one CSV files is generated for the 5% sampling. This file has two columns, one with the randomly chosen identifiers and another column to be filled with the results of the manual evaluation.

3. Results and Analysis

The paragraphs of *Vol. VI.* were extracted semiautomatically using a tool developed as part of a research project to extract knowledge from biological literature. This project was financed by Instituto Tecnológico de Costa Rica. This project allows identifying and marking the different parts of biological descriptions: morphological descriptions, diagnostic descriptions, geographic distributions and more. As a result of this process, it is possible to generate XML documents in which the snippet element contains the text that corresponds to one of the previous categories.

The evaluation presented in this section was performed manually on a set of random samples taken on the following information items: distributions in Costa Rica and distributions in the world, which include items with subdistributions that are taken into account in the random sample.

The details of the selection are presented in the Table 1.

3.1. Evaluation criteria

The evaluation ratings follow certain criteria that are detailed below:

- **GOOD**
  All the statutes listed below must be fulfilled:
  - A term was correctly identified as a possible geographical point in Costa Rica or the world.
  - The term was found in the gazetteer or taxonomy and matches the context of the place or zone.
- **BAD**
  At least one of the following statutes must be fulfilled:
  - The term found does not correspond to a possible geographical point in Costa Rica or the world.
  - The possible term is found in the gazetteer or taxonomy, but does not match the geographical point to which the context of the distribution paragraph refers.
  - The possible term is not found but is contained by the gazetteer or taxonomy.
- **UNKNOWN**
  All of the following statutes must be complied with:
  - The term corresponds to a geographical point of Costa Rica or the world.
  - The term is not found but is not contained in the gazetteer or taxonomy.

TABLE 2: Results of manual evaluation

Type of Clause	Good (frequency / %)	Bad (frequency / %)	Unknown (frequency / %)
Distribution in Costa Rica	137 / 41.39	88 / 26.59	106 / 32.02
Distribution in the world	133 / 87.5	16 / 10.53	3 / 1.97

3.2. Classification of errors

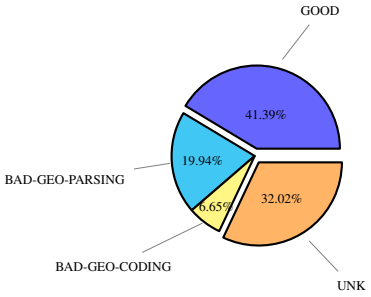
The error type frequency for the clauses classified as BAD corresponding to the sample of 5 % of the total of the processed clauses is shown in Table 3. The two types of possible errors are:

- **GEO-PARSING**
  When the term found does not correspond to a possible geographical point in Costa Rica, or in the world (depending on the case).
- **GEO-CODING**
  When the possible term is found in the gazetteer or taxonomy, but does not match the geographical point to which the context of the distribution paragraph refers. It can also be classified as a GEO-CODING error when the term is not found but is contained in the gazetteer or taxonomy.

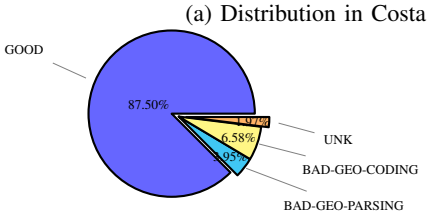
Figure 4 shows how the two types of BAD are distributed in the total of the results. Thus, in the Figure 5 shows the distribution of the values qualified as GOOD plus those that were qualified with UNKNOWN, using the values in formulas (2) and (3) as described in general in formula (1).

TABLE 3: Type of error within clauses that were classified with BAD in the sample of 5 %

	Distribution in Costa Rica (frequency / %)	**Distribution in the world (frequency / %)**
GEO-PARSING	66 / 75.00	6 / 37.50
GEO-CODING	22 / 25.00	10 / 62.50
Total	88	16



(a) Distribution in Costa Rica



(b) Distribution in the world

Figure 4: Results of manual evaluation of distribution clauses in Costa Rica and the world with a sample of 5%.

The evaluation results presented show that in some cases it is possible to implement algorithms to extract geographical points and associate them with gazetteers that performs with high effectiveness. Although the main algorithm had a 41.39 % yield for distribution in Costa Rica, it obtained a much higher value, 87.5 % for distribution in the world.

These data imply a low performance mainly in distribution in Costa Rica. However under certain circumstances it is possible to extract geographical terms and associate them with gazetteers with good success (almost 90 % for distribution in the world).

The fact that the distribution in the world has better results is related to the complexity of the text and the size of the gazetteer. The sentences with the geographical distributions for Costa Rica are, grammatically more complex than the sentences with the distribution in the world. See the following example; for the distribution in Costa Rica: *cerca de la División Continental, Cordillera Central (Turrialba), N Cordillera de Talamanca (vecindad de El Empalme)*; for the distribution in the world: *Honduras-Panama, Venezuela.*.

As shown in Table 3, the toponyms that were evaluated with BAD basically correspond to two types of errors: GEO-PARSING and GEO-CODING, which agree with the two main steps in the algorithm presented in sections 2.3 and 2.4. The toponyms that were classified as UNKNOWN,

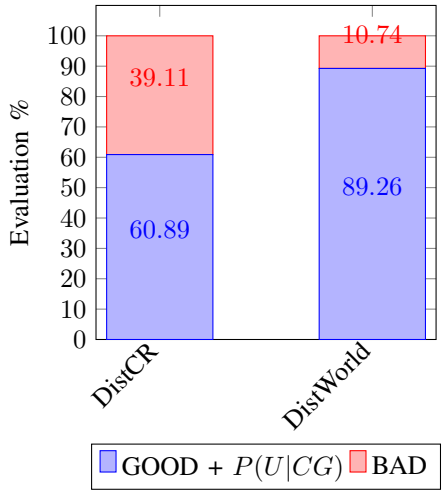


Figure 5: Comparison of results of GOOD + $P(U|CG)$ against BAD in the sample of 5 %

following the criterion of the section 3.1, have the problem that they simply do not appear in the gazetteer.

Let's assume $P(G|CG)$ as the probability that a term is classified as GOOD given a complete gazetteer. Let $P(G|IG)$, $P(U|IG)$ and $P(B|IG)$ be the probabilities that a term is classified as GOOD, UNK, and BAD (respectively) given an incomplete gazetteer.

The results could be extrapolated assuming that if the gazetteer were complete, then the UNKNOWN cases will be distributed as GOOD or BAD following the same distribution of the known cases.

$$P(G|CG) = P(G|IG) + P(U|IG)\frac{(P(G|IG)}{(P(G|IG) + P(B|IG))} \quad (1)$$

The following values are graphically described in Figure 5:

- Worldwide distribution :

$$P(G/CG) = 0.8750 + 0.0197\frac{0.8750}{0.8750 + 0.1053}$$

$$= 0.8926 \quad (2)$$

- Costa Rica distribution :

$$P(G/CG) = 0.4139 + 0.3202\frac{0.4139}{0.4139 + 0.2659}$$

$$= 0.6089 \quad (3)$$

Following the above distributions there is approximately a 90 % probability of obtaining a GOOD result by labeling a distribution term in the world and falling to 61 % for distribution in Costa Rica.

4. CONCLUSIONS

The results of the tests and their subsequent analysis give rise to the following conclusions:

1) An algorithm was designed and implemented to extract geographic entities of plant species from the *Manual de Plantas de Costa Rica. V: Dicotele-doneas (Clusiaceae-Gunneraceae)* [17] and *Manual de Plantas de Costa Rica. VI: Dicotiledoneas (Haloragaceae-Phytolaccaceae)* [4].
2) An algorithm was designed and implemented to associate the geographic entities of the species of plants of the previous point, with a gazetteer of Geonames.
3) The effectiveness of the implemented algorithms was evaluated through experimentation. Some of the conclusions related to the results are:
    a) The error rate is high for geographic distributions in Costa Rica (32 %). While for distribution in the world remains close to 10 %.
    b) The part of the algorithm that mainly fails in the tests on geographic distributions in Costa Rica is the geo-parsing (total percentages: 24 %). On the other hand, for the distribution in the world, the geo-parsing errors are rather few (4 % of the total of the sample.)
    c) The precision of the two types of geographic distribution: in Costa Rica and in the world is: 41.39 % and 87.50 % respectively.
    d) Adding the clauses with the qualification of UNKNOWN and assuming that the gazetteer has an increase in the geographical points; A projected precision of 60.89 % and 89.26 % can be estimated for the geographical distribution in Costa Rica and the world, respectively.

5. Future Work

Although the algorithm presented has a low overall performance in some cases, there is a lot of potential to improve performance. Some improvements can be listed below:

- Deepen the geo-parsing algorithm so that it can better recognize terms within enumerative sentences and with spatial, compositional, and possessive prepositions. Freeling provides valuable information on its PoS module.
- Develop a specific gazetteer for locations of interest in Costa Rica that were not included in the consulted gazetteer. Incorporate initiatives that other institutions such as the National Geographic Institute of Costa Rica have.
- Refine the geo-coding algorithm, specifically in the *search & match* search engine: Implement an Engine Scoring exercise with the terms of the gazetteer in Solr. So that you can evaluate each change in the search parameters. "Engine Scoring" is based on the premise of having a set of base searches and a set of results ordered by a user. Then an algorithm is executed and returns a s number which is called *score*, by itself has no meaning, but is very useful to be compared with a value of s, executed with a Different Solr settings.
- Investigate how to address the problem of locating the correct names in English in the gazetteer, when the term in the source text is in Spanish: extend the thesaurus entries, investigate other approaches for multi-language searches or use a translation service like Google Translator.
- Investigate similar applications as BioGeomancer ([18] and [19]) and compare the efficacy of the results with those presented in this paper.
- Implement an HTTP interface for the algorithm that allows exposing it to interaction with other software using the REST architecture.

References

[1] B. J. Doyle, "Medicinal plants from Costa Rica for the treatment of menopause: Pharmacognosy of pimenta dioica," *ProQuest Dissertations and Theses*, p. 142, 2008. [Online]. Available: http://ezproxy.itcr.ac.cr:2164/docview/304330427?accountid=27651

[2] A. C. Gilman, "Biodiversity patterns in tropical montane rainforest flora of Costa Rica," *ProQuest Dissertations and Theses*, p. 123, 2007.

[3] M. G. Gei, "Biological nitrogen fixation in tropical dry forests of Costa Rica: Patterns and controls," *ProQuest Dissertations and Theses*, p. 236, 2014. [Online]. Available: http://ezproxy.itcr.ac.cr:2164/docview/1563373715?accountid=27651

[4] B. Hammel, M. Grayum, C. Herrera, and N. Zamora, *Manual de Plantas de Costa Rica*. Missouri Botanical Garden Press, 2007, vol. VI: Dicotiledoneas (Haloragaceae-Phytolaccaceae).

[5] J. L. Leidner and M. D. Lieberman, "Detecting geographical references in the form of place names and associated spatial natural language," *SIGSPATIAL Special*, vol. 3, no. 2, pp. 5–11, Jul. 2011. [Online]. Available: http://doi.acm.org/10.1145/2047296.2047298

[6] J. R. Curran, S. Clark, and J. Bos, "Linguistically motivated large-scale nlp with c&c and boxer," in *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ser. ACL '07. Stroudsburg, PA, USA: Association for Computational Linguistics, 2007, pp. 33–36. [Online]. Available: http://dl.acm.org/citation.cfm?id=1557769.1557781

[7] (2016) OpenNLP. Apache Software Foundation. [Online]. Available: http://opennlp.apache.org/index.html

[8] (2016) OpenCalais. Thomson Reuters. [Online]. Available: http://www.opencalais.com

[9] H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan, "GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications," in *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02)*, 2002.

[10] (2016) Apache Solr. Apache Software Foundation. [Online]. Available: http://archive.apache.org/dist/lucene/solr/ref-guide/apache-solr-ref-guide-5.5.pdf

[11] (2016) Aspire Community Edition. Search Technologies Corporation. [Online]. Available: https://www.searchtechnologies.com/aspire

[12] M. Kimler and R. Göbel, "Geo-coding: Recognition of geographical references in unstructured text, and their visualisation," 2004.

[13] B. Pouliquen, R. Steinberger, C. Ignat, and T. De Groeve, "Geographical information recognition and visualization in texts written in various languages," in *Proceedings of the 2004 ACM Symposium on Applied Computing*, ser. SAC '04. New York, NY, USA: ACM, 2004, pp. 1051–1058. [Online]. Available: http://doi.acm.org/10.1145/967900.968115

[14] L. Padró and E. Stanilovsky, "Freeling 3.0: Towards wider multilinguality," in *Proceedings of the Language Resources and Evaluation Conference (LREC 2012)*. Istanbul, Turkey: ELRA, May 2012.

[15] (2016) NGA GEOnet Names Server. [Online]. Available: http://geonames.nga.mil/gns/html/

[16] M. Wick. (2016) GeoNames. [Online]. Available: http://www.geonames.org

[17] B. Hammel, M. Grayum, C. Herrera, and N. Zamora, *Manual de Plantas de Costa Rica*. Missouri Botanical Garden Press, 2010, vol. V: Dicoteledoneas (Clusiaceae-Gunneraceae).

[18] R. P. Guralnick, J. Wieczorek, R. Beaman, R. J. Hijmans, and the BioGeomancer Working Group, "Biogeomancer: Automated georeferencing to map the world's biodiversity data," *PLOS Biology*, vol. 4, no. 11, pp. 1–2, 11 2006. [Online]. Available: http://dx.doi.org/10.1371/journal.pbio.0040381

[19] A. Chapman and J. W. (eds). (2016) Guide to Best Practices for Georeferencing. Copenhagen: Global Biodiversity Information Facility. [Online]. Available: http://www.gbif.org/orc/?doc_id=1288