Grounding Toponyms in an Italian Local News Corpus

Davide Buscaldi
NLE Lab, ELiRF Group
DSIC, Universidad Politécnica de Valencia
Camino de Vera, s/n
Valencia, Spain
dbuscaldi@dsic.upv.es

Bernardo Magnini
Fondazione Bruno Kessler, FBK-IRST
Via Sommarive 18
Povo (Trento), Italy
magnini@fbk.eu

ABSTRACT

In this paper we present a study carried out over toponyms contained in an Italian news collection, in order to determine the degree of ambiguity of toponyms and how difficult could be to resolve such ambiguities. The results show that frequent toponyms are usually less ambiguous than rare toponyms. The resolution of ambiguities on a sample of $1,042$ toponyms with different features confirms that ambiguous toponyms are spatially autocorrelated.

Categories and Subject Descriptors

I.2.7 [**Artificial Intelligence**]: Natural Language Processing—*text analysis, language parsing and understanding*

General Terms

Algorithms, Measurement, Performance

Keywords

Toponym Resolution, Geographic Information Retrieval

1. INTRODUCTION

Toponyms, or place names, are very important in Natural Language Processing (NLP). In unstructured texts, they represent an important source of geographic information that can be detected and used, together with postal codes and phone numbers. This information can be used, for instance, to index news with the countries and cities they mention and to automatically visualise this information on geographical maps, such as in applications like News Explorer[1] [9]. The identification of toponyms is carried out usually with the help of gazetteers. Gazetteers are lists of geographic entities, usually enriched with additional information, such as their geographic coordinates, class (city, river, country, etc.), and size. Two examples of well-known gazetteers are

[1] http://emm.newsexplorer.eu

the Getty thesaurus of geographical names[2] and the Geonames[3] gazetteer. However, it is common to find toponyms that are ambiguous with other class of names (geo/non-geo ambiguity), such as *Java*, which may be a name used to indicate a software or an island, or with other toponyms (geo/geo ambiguity). The geo/geo ambiguity can be between places of different class (e.g. *Alabama* may be a river or a state), or between places from the same class (for instance, *Cambridge* may be used to identify a place in the U.S.A. or in the United Kingdom). Usually, in NLP, the resolution of geo/non-geo ambiguities is carried out by means of Named Entity Recognition (NER) tools, most of which are based on sequence learning methods, such as Hidden Markov Models (HMM) or Conditional Random Fields (CRF).

The resolution of geo/geo ambiguities is commonly referred to as toponym *grounding*, or toponym disambiguation. Investigation on toponym grounding is currently an open field: there are many methods that have been developed for toponym grounding, but no one has emerged over others. Some methods rely on an external knowledge source, which describes the topology of the geographic entities [12, 3] or ad-hoc defined rules [1]; in these cases the method are referred to as *knowledge-based* methods. Other methods (*geometric* methods) use only the information about coordinates and distances between places [10]. Methods based on machine learning, which usually depend on hand-labelled data and contextual information have also been used for this task [6]. A comparative study between a geometric method and a knowledge-based method [4] showed that the second one could obtain better results. However, both the test collection and the knowledge source used in the comparison were not specifically aimed to the study of toponyms, but they were an adaptation of resources commonly used in Word Sense Disambiguation (WSD).

In this work we do not aim to determine which method is better than another, but we try to discover the characteristics of toponyms that are contained in a large, localised Italian news collection, and determine how difficult it could be to resolve toponym ambiguities in such a collection, and what is the importance of features such as toponym frequency, or the distance of a toponym referent from the "home" of the news in order to resolve the ambiguities.

2. DATA AND METHODOLOGY

The news collection is constituted by the articles of the "L'Adige" newspaper, from 2002 to 2006. The target au-

[2] http://www.getty.edu/
[3] http://www.geonames.org

dience of this newspaper is constituted mainly by the population of the city of Trento and its province, and in second place by the Italian-speaking community of the Südtirol province (Alto-Adige in Italian). The news stories are classified in 11 sections; some are thematically closed, such as "sport" or "international". Other sections are dedicated to important places in the province: "Riva del Garda", "Rovereto", for instance. This collection has been labeled with *TextPRO* [5], a suite of tools oriented towards a number of NLP tasks, such as Web page cleaning, tokenization, sentence splitting, morphological analysis, PoS-tagging, lemmatization, multiword recognition, chunking and NER. The NER function of the TextPRO suite is carried out by *EntityPRO*, a Support Vector Machine-based tool that obtained 82.1% in precision over Italian named entities [8].

The toponyms in the collection have been labelled by EntityPRO with the following labels: *GPE* (Geo-Political Entities) and *LOC* (LOCations). According to the ACE guidelines [7], "GPE entities are geographical regions defined by political and/or social groups. A GPE entity subsumes and does not distinguish between a nation, its region, its government, or its people. Location (LOC) entities are limited to geographical entities such as geographical areas and landmasses, bodies of water, and geological formations". The precision of EntityPRO over GPE and LOC entities has been estimated, respectively, in 84.8% and 77.8% in the EvalITA-2007[4] exercise. In the collection there are $70,025$ entities labelled as GPE or LOC, with a majority of them (58.9%) occurring only once. In the data, we observed that names of countries and cities have been labelled with GPE, whereas LOC has been used to label everything that can be considered a place, including street names.
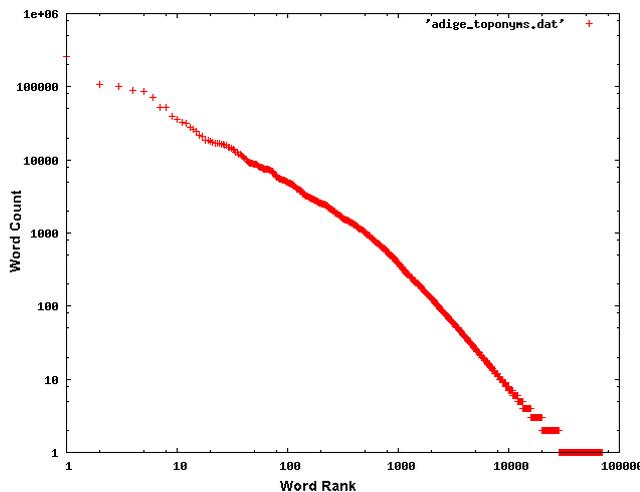

Figure 1: Toponyms frequency in the news collection, sorted by frequency rank. Log scale on both axes.

As can be seen in Figure 1, toponyms follow a zipfian distribution, independently from the section they belong to. This is not particularly surprising, since the toponyms in the collection represent a corpus of natural language, for which Zipf law holds ("in any large enough text, the frequency ranks of wordforms or lemmas are inversely proportional to

the corresponding frequencies" [13]). We can also observe that the set of most frequent toponyms change depending on the examined section (see Table 1). Only 4 of the most frequent toponyms in the "international" section are included in the 10 most frequent toponyms in the whole collection, and if we look just at the articles contained in the local "Riva del Garda" section, only 2 of the most frequent toponyms are also the most frequent in the whole collection. "Trento" is the only frequent toponym that appears in all lists.

In order to study the ambiguity of place names in "L'Adige", we had to use a resource providing a mapping from place names to their actual geographic coordinates. We planned to use the Geonames gazetteer, but this resource do not cover street names, which count for 9.26% of the total number of unique toponyms in the collection. Therefore, we had to build a repository of possible referents by integrating the data in the Geonames gazetteer with those obtained by querying the Google maps API geocoding service[5]. For instance, this service returned 9 places corresponding to the toponym "Piazza Dante", one in Trento and the other 8 in other cities in Italy (see Figure 2). A problem with this kind

Figure 2: Places corresponding to "Piazza Dante", according to the Google geocoding service (retrieved Nov. 26 2009).

of toponyms is that they are particularly ambiguous, especially if the name of the street indicates the city pointed by the axis of the road: for instance, there is a "via Brescia" both in Mantova and Cremona, in both cases pointing towards the city of Brescia. Another common problem occurs when a street crosses different municipalities while keeping the same name.

Due to the usage limitations of the Google maps geocoding service, we had to limit the size of the sense repository in order to obtain enough coverage in a reasonable time. Therefore, we decided to include only the toponyms that appeared at least 2 times in the news collection. The result was a repository containing $13,324$ unique toponyms and $62,408$ possible referents. This corresponds to 4.68 referents per toponym, a degree of ambiguity considerably higher if compared to other resources used in the toponym disambiguation task, as can be seen in Table 2. The higher de-

---

[4]http://evalita.fbk.eu/2007/index.html

[5]http://maps.google.com/maps/geo

Table 1: Frequencies of the 10 most frequent toponyms, calculated in the whole collection ("all") and in two sections of the collection ("international" and "Riva del Garda").

all		international		Riva del Garda	
toponym	frequency	toponym	frequency	toponym	frequency
Trento	260, 863	Roma	32, 547	Arco	25, 256
provincia	109, 212	Italia	19, 923	Riva	21, 031
Trentino	99, 555	Milano	9, 978	provincia	6, 899
Rovereto	88, 995	Iraq	9, 010	Dro	6, 265
Italia	86, 468	USA	8, 833	Trento	6, 251
Roma	70, 843	Trento	8, 269	comune	5, 733
Bolzano	52, 652	Europa	7, 616	Riva del Garda	5, 448
comune	52, 015	Israele	4, 908	Rovereto	4, 241
Arco	39, 214	Stati Uniti	4, 667	Torbole	3, 873
Pergine	35, 961	Trentino	4, 643	Garda	3, 840

Table 2: Average ambiguity for resources typically used in the toponym disambiguation task.

Resource	Unique names	Referents	ambiguity
Wikipedia (Geo)	180, 086	264, 288	1.47
GeoNet	2, 954, 695	3, 988, 360	1.35
WordNet2.0	2, 069	2, 188	1.06

gree of ambiguity is due to the introduction of street names and "partial" toponyms such as "provincia" (province) or "comune" (community). Usually these names are used to avoid repetitions if the text previously contains another (complete) reference to the same place, such as in the case "provincia di Trento", or "comune di Arco", or when the context is not ambiguous.

We studied how ambiguity is distributed with respect to frequency. We first defined the probability of finding an ambiguous toponym at frequency F by means of Formula 1.

$$P(F) = \frac{|T_{amb_F}|}{|T_F|} \qquad (1)$$

Where $f(t)$ is the frequency of toponym t, T is the set of toponyms with frequency $\leq$ F: $T_F = \{t | f(t) \leq F\}$ and T_{amb_F} is the set of ambiguous toponyms with frequency $\leq$ F, i.e. $T_{amb_F} = \{t | f(t) \leq F \wedge s(t) > 1\}$, with $s(t)$ indicating the number of senses for toponym t.

In Figure 3 we plotted $P(F)$ for the toponyms in the collection, taking into account all the toponyms, only street names and all toponyms except street names. As can be seen from the figure, less frequent toponyms are particularly ambiguous: the probability of a toponym with frequency $f(t) \leq 100$ of being ambiguous is between 0.87 and 0.96 in all cases, while the probability of a toponym with frequency $1,000 < f(t) \leq 100,000$ of being ambiguous is between 0.69 and 0.61. It is notable that street names are more ambiguous than other terms: their overall probability of being ambiguous is 0.83, compared to 0.60 of all toponyms (including street names) and 0.58 of all toponyms except street names.

In the case of common words, the opposite phenomenon is usually observed: the most frequent words (such as "have", "be") are also the most ambiguous ones. The reason of this behaviour is that the more a word is frequent, the more are the chances it could appear in different contexts. Toponyms are used somehow in a different way: frequent toponyms usually refer to well-known location and have a defi-
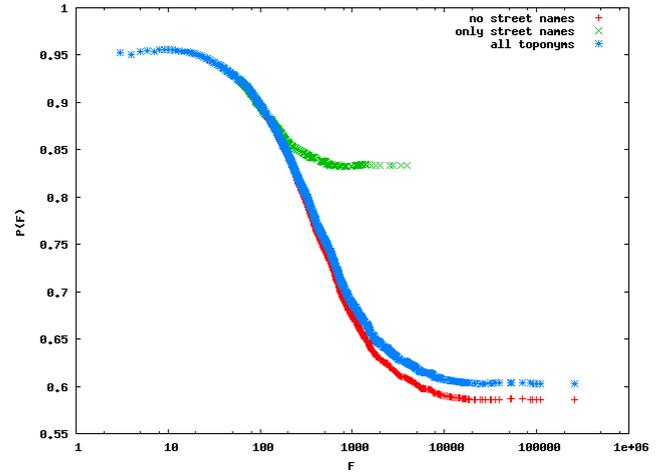

Figure 3: Correlation between toponym frequency and ambiguity, taking into account only street names (*no street names*), all toponyms, and all toponyms except street names. Log scale applied to x-axis.

nite meaning, although used in different contexts. However, this point deserves to be studied thoroughly in the future.

We studied also the spatial distribution of toponyms in the collection with respect to the "source" of the news collection. Since "L'Adige" is based in Trento, we counted how many toponyms are found within a certain range from the center of the city of Trento (see Figure 4).

These results confirm the findings by Brunner and Purves [2] about spatial autocorrelation of ambiguous toponyms in news collections. It can be observed that the majority of place names is used to reference places within 400 Kms. of distance from Trento.

We developed a disambiguation method based on geometric features. The reason of this choice was that both knowledge-based methods and machine learning methods were inapplicable. In the first case, it was not possible to discriminate places at an administrative level lower than province, since it is the lowest administrative level provided by the Geonames gazetteer. For instance, it is possible to distinguish "via Brescia" in Mantova from "via Brescia" in Cremona (they are in two different provinces), but it is not
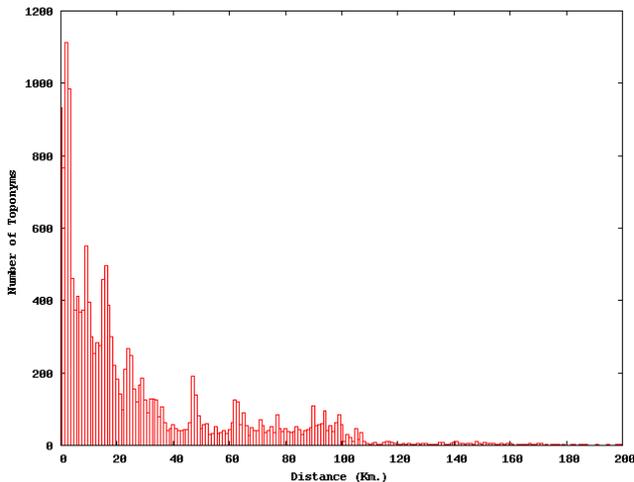
Figure 4: Number of toponyms found at different distances from Trento. Distances are expressed in Kms. divided by 10.

possible to distinguish "via Mantova" in Trento from "via Mantova" in Arco, because they are in the same province. Google does actually provide data at municipality level, but they were incompatible for merging them with those from the Geonames gazetteer. In the case of machine learning, we discarded this possibility because we had no availability of a large enough quantity of labelled data.

Therefore, we developed a disambiguation method loosely based on the method described by Smith and Crane [10]. We included knowledge related to the spatial autocorrelation of ambiguous toponyms, as suggested by [2]; this knowledge was included by adding to the context of the toponym to be resolved the place related to the news source: "Trento" for the general collection, "Riva del Garda" for the Riva section, "Rovereto" for the related section and so on. The base context for each toponym is composed by every other toponym that can be found in the same document. The size of this context window is not fixed: the number of toponyms in the context depends on the toponyms contained in the same document of the toponym to be disambiguated. We refined the disambiguation method by including also the information about the frequency of a context toponym: we supposed that place names with high frequency have a higher resolving power than place names with low frequency. For instance, terms that are frequently seen in news like "USA", "Europe", "Italy", are usually considered not ambiguous and they could be used to specify the position of ambiguous toponyms located nearby in the text. This assumption is based on the correlation between frequency and ambiguity observed in Figure 3. Finally, we considered that the word distance in text is key to solve ambiguities: usually, in text, people writes a disambiguating place just besides the ambiguous toponyms (e.g. Cambridge, Massachusetts). The resulting algorithm is the following one:

1. Identify the next ambiguous toponym t with senses $S = (s_1, \ldots, s_n)$;

2. Find all toponyms t_c in context;

3. Add to the context all senses $C = (c_1, \ldots, c_m)$ of the

toponyms in context (if a context toponym has been already disambiguated, add to C only that sense);

4. $\forall c_i \in C, \forall s_j \in S$ calculate the map distance $d_M(c_i, s_j)$ and text distance $d_T(c_i, s_j)$;

5. Combine frequency count $(F(c_i))$ with distances in order to calculate, for all s_j:
$F_i(s_j) = \sum_{c_i \in C} \frac{F(c_i)}{(d_M(c_i, s_j) \cdot d_T(c_i, s_j))^2}$;

6. Resolve t by assigning it the sense $s = \arg_{s_j \in S} \max F_i(s_j)$.

7. Move to next toponym; if there are no more toponyms: stop.

Text distance was calculated using the number of word separating the context toponym from t. Map distance is the great-circle distance calculated using the following formula:

$$d_M(c, s) = R \cdot \arccos(\sin(\phi_c)\sin(\phi_s) + \cos(\phi_c)\cos(\phi_s)\cos(\Delta\lambda)) \quad (2)$$

Where R is the Earth's radius, approximated to $6,371$ Km., ϕ_c and ϕ_s are the latitudes of points c and s, and λ is the difference of their longitudes.

If we take into account that TextPRO identified the toponyms and labelled them with their position in the document, greatly simplifying step 1,2 and the calculation of text distance, the complexity of the algorithm is in $O(n^2 \cdot m)$, where n is the number of toponyms and m the number of senses (or possible referents). Given that the most ambiguous toponym in the database has 32 senses, we can rewrite the complexity in terms only of the number of toponyms as $O(n^3)$. Therefore, we carried out the experiment only on a small test set and not on the entire document collection. We labelled $1,042$ entities of type GPE/LOC with the right referent, selected among the ones contained in the repository. This test collection was intended to be used to estimate the accuracy of the disambiguation method. In order to understand the relevance of the obtained results, we compared the results obtained with this method to the results obtained by assigning to the ambiguous toponyms the referent with minimum distance from the context toponyms (that is, without taking into account neither the frequency nor the text distance) and to the results obtained without adding the context toponyms related to the news source.

3. RESULTS

In Table 3 we show the result obtained using the proposed method, compared to the results obtained with the baseline method and a version of the proposed method that did not use text distance. Precision was calculated as the

Table 3: Results obtained over the set of $1,042$ ambiguous toponyms. *complete*: method including text distance, map distance, frequency and local context; $map + freq + local$: method that do not use text distance; $map + local$: method that uses only local context and map distance.

method	precision	recall
complete	88.43%	88.34%
map+freq+local	88.81%	88.73%
map+local	79.36%	79.28%
baseline (only map)	78.97%	78.90%

number of correctly resolved toponyms divided the number of resolved toponyms; recall was calculated as the number of correctly resolved toponyms divided the number of toponyms in the collection. The difference is due to the fact that the methods were able to deal with $1,038$ toponyms instead of the complete set of $1,042$ toponyms. It was not possible to disambiguate 4 toponyms because of the lack of context toponyms in the respective documents. The average context size was 6.96 toponyms per document, with a maximum and a minimum of 40 and 0 context toponyms in a document, respectively.

4. CONCLUSIONS

The results show that the improvement by adding the local context to the method is small: this confirms the results obtained by [2] about the autocorrelation of ambiguous toponyms. The greatest improvement ($\sim 9\%$) was obtained by taking into account the importance of places, measured as frequency in the collection. Text distance did not improve the results over the test set, although we cannot expect that these results generalize due to the language (Italian) and locality of the news collection. In the future, we plan to use the labelled data to set up a supervised classifier, with the help of bootstrapping as proposed by [11], in order to find the optimal parameters for combining the available features.

5. ACKNOWLEDGMENTS

6. REFERENCES

[1] G. Andogah, G. Bouma, J. Nerbonne, and E. Koster. Placename ambiguity resolution. In *LREC 2008 workshop on Methodologies and Resources for Processing Spatial Language*, 2008.

[2] T. J. Brunner and R. S. Purves. Spatial autocorrelation and toponym ambiguity. In *GIR '08: Proceeding of the 2nd international workshop on Geographic information retrieval*, pages 25–26, New York, NY, USA, 2008. ACM.

[3] D. Buscaldi and P. Rosso. A conceptual density-based approach for the disambiguation of toponyms. *International Journal of Geographical Information Systems*, 22(3):301–313, 2008.

[4] D. Buscaldi and P. Rosso. Map-based vs. knowledge-based toponym disambiguation. In *GIR '08: Proceeding of the 2nd international workshop on Geographic information retrieval*, pages 19–22, New York, NY, USA, 2008. ACM.

[5] C. G. Emanuele Pianta and R. Zanoli. The TextPRO Tool Suite. In N. C. et al., editor, *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, may 2008. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2008/.

[6] E. Garbin and I. Mani. Disambiguating toponyms in news. In *conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT05)*, pages 363–370, Morristown, NJ, USA, 2005. Association for Computational Linguistics.

[7] Linguistic Data Consortium. *ACE English Annotation Guidelines for Entities*, 2008. http://projects.ldc.upenn.edu/ace/docs/English-Entities-Guidelines_v6.6.pdf.

[8] E. Pianta and R. Zanoli. Exploiting SVM for Italian Named Entity Recognition. *Intelligenza Artificiale, Special issue on NLP Tools for Italian*, IV(2), 2007. In Italian.

[9] B. Pouliquen, M. Kimler, R. Steinberger, C. Ignat, T. Oellinger, K. Blackler, F. Fuart, W. Zaghouani, A. Widiger, A.-C. Forslund, and C. Best. Geocoding Multilingual Texts: Recognition, Disambiguation and Visualisation. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC-2006)*, pages 53–58, Genova, Italy, 2006.

[10] D. A. Smith and G. Crane. Disambiguating geographic names in a historical digital library. In *Research and Advanced Technology for Digital Libraries*, volume 2163 of *Lecture Notes in Computer Science*, pages 127–137. Springer, Berlin, 2001.

[11] D. A. Smith and G. S. Mann. Bootstrapping toponym classifiers. In *Proceedings of the HLT-NAACL 2003 workshop on Analysis of geographic references*, pages 45–49, Morristown, NJ, USA, 2003. Association for Computational Linguistics.

[12] R. Volz, J. Kleb, and W. Mueller. Towards ontology-based disambiguation of geographical identifiers. In *I3 Workshop held at the 16th International World Wide Web Conference (WWW2007)*, Banff, Alberta, Canada, 2007.

[13] G. K. Zipf. *Human Behavior and the Principle of Least Effort.* Addison-Wesley (Reading MA), 1949.