A Comparison of Methods for the Automatic Identification of Locations in Wikipedia[*]

Davide Buscaldi
Dpto. Sistemas Informáticos y Computación
Universidad Politécnica de Valencia
Valencia, Spain
dbuscaldi@dsic.upv.es

Paolo Rosso
Dpto. Sistemas Informáticos y Computación
Universidad Politécnica de Valencia
Valencia, Spain
prosso@dsic.upv.es

ABSTRACT

In this paper we compare two methods for the automatic identification of geographical articles in encyclopedic resources such as Wikipedia. The methods are a WordNet-based method that uses a set of keywords related to geographical places, and a multinomial Naïve Bayes classificator, trained over a randomly selected subset of the English Wikipedia. This task may be included into the broader task of Named Entity classification, a well-known problem in the field of Natural Language Processing. The experiments were carried out considering both the full text of the articles and only the definition of the entity being described in the article. The obtained results show that the information contained in the page templates and the category labels is more useful than the text of the articles.

Categories and Subject Descriptors

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

General Terms

Algorithms, Measurement, Performance

1. INTRODUCTION

Recently, Geographical Information Retrieval (GIR) has captured the attention of many researchers that work in the field of Natural Language Processing. Geographical information is spread over most of the web. It has been discovered that a significant proportion of internet searches contains at least a geographical term [12]. One of the issues encountered in GIR is the identification of place names in text, that is crucial in order to perform some GIR-related tasks such as the determination of the geographical scope of web pages [7], and expansion of queries expressed in natural language.

The identification of place names can be considered a specialization of the broader task of Named Entity Classification (NEC). NEC is a well known task in the field of Natural Language Processsing (NLP). It consists in assigning to an expression, previously identified as a Named Entity (NE), one of a set of possible categories. For instance, we may classify "Lisbon" as a *city*, "Sporting Lisbon" as a *football club* and "Lisbon story" as a *movie*. When we are interested in the identification of place names we are carrying out a binary classification of the named entities into 'geographical' and not. In this example, it is important to tell that the "Lisbon" in "Lisbon story" is not a place.

Geographical-oriented ontologies [5, 6], or even general ontologies adapted to this task [2] are often used in order to address GIR issues; in the former case, the drawback is constituted by the huge amount of work needed in order to create the ontology. In the latter one, the quantity of geographical information included in a general ontology is usually too small in order to be used as an effective geographical resource.

One of the last development is the use of encyclopedias such as Wikipedia[1], alone or in combination with the resources named above [1]. Recently, [3] has proposed a method based on Wikipedia for the NEC task. Due to the lack of standardization that can be observed in the pages of Wikipedia, because of the 'open' nature of the Wiki project, the automatic identification of a Wikipedia page as one referring to a geographical location may prove to be a difficult task.

Overell and Rüger developed a rule-based system that exploit the information contained in page templates and the category label [9]. They also created a collection of manually tagged Wikipedia articles that was used as test set. In this paper we present a method based on the similarity of the article to a set of keywords extracted from WordNet [8]. We compared it with the results obtained by Overell and Rüger and a multinomial Bayesian classifier trained over a portion of the Ludovic Denoyer's Wikipedia xml corpus [4].

2. OUR WORDNET-BASED METHOD

We extracted from WordNet a set of geographical keywords using the *holonymy* (part-of) relationship and its inverse, *meronymy*. We retrieved iteratively all the meronyms that can be reached from two root synsets: *northern_hemisphere* and *southern_hemisphere*. The result is the list of all the geographical synset included in WordNet. The words

[*] We would like to thank the TIN2006-15265-C06-04 research project for partially supporting this work.

GIR'07, November 9, 2007, Lisbon, Portugal.
ACM 978-1-59593-828-2/07/0011.

[1] http://www.wikipedia.org

contained in these synsets and in the definition of each of them (the *gloss*) were added to the set of keywords, with the exception of stop-words. For instance, consider the following synset and its gloss:

```
 Lisbon, Lisboa, capital of Portugal - (capital
and largest city and economic and cultural cen-
ter of Portugal; a major port in western Portu-
gal on Tagus River where it broadens and empties
into the Atlantic)
```

The terms added to the set of keywords in this case are: *capital, largest, city, economic, cultural, center, Portugal, major, port, western, Tagus, river, broadens, empties, Atlantic.*

Considering that the 10 most frequent words selected in this way are: *city, state, population, area, world, km, country, new, north, river*, we can assert that the extracted keywords are quite representative of the geographical domain.

In order to determine whether a Wikipedia article is in the geographical domain or not, we need to measure its similarity to the set of geographical keywords. Let us name W_a the set of words in an article a of Wikipedia, T the set of keywords extracted from WordNet. We calculated the similarity score $S(a,T)$ between a and T is computed by means of the Dice formula:

$$S_{Dice}(a,T) = \frac{2|W_a \cap T|}{|W_a| + |T|} \quad (1)$$

With the Dice coefficient, similarity is determined only by the number of words that appear both in the document and in the set of WordNet keywords. A more precise measure of similarity is the cosine coefficient that takes into account also the number of times that words appear. Let $\bar{w_a}$ be the vector of the words contined in W_a and $\bar{t}$ the vector of the words in T. Then, the cosine scoefficient is calculated as:

$$S_{cosine}(a,T) = \frac{\bar{w_a} \cdot \bar{t}}{\sqrt{||\bar{w_a}|| * ||\bar{t}||}} \quad (2)$$

Here, $||\bar{w_a}||$ and $||\bar{t}||$ represent the Euclidean length respectively of the vectors $\bar{w_a}$ and $\bar{t}$, which is the square root of the dot product of the vector with itself.

3. EXPERIMENTS

We carried out our experiments using a snapshot of the English Wikipedia taken on June 19*th* 2006. Nine sets of article names were generated for both similarity formulae using thresholds between 0.02 and 0.18. That is, an article a was added to the list if $S(a,T) > \alpha$, with $\alpha \in \{0.02, 0.04, 0.06, 0.08, 0.10, 0.12, 0.14, 0.16, 0.18\}$.

The generated sets were compared to a set of Wikipedia articles paired to TGN[2] identifiers, created by Overell and Rüger, that contains 1,694 articles [9].

For each experiment we calculated *recall* and *precision* in the following way:

$$recall = \frac{|G \cap L|}{|G|} \quad (3)$$

$$precision = \frac{|G \cap L|}{|L|} \quad (4)$$

where G is the test set and L are the articles labeled by our method as locations.

We also compared our method to a multinomial Naive Bayes classifier, trained over 40,380 articles randomly extracted from the Wikipedia XML corpus [4]. Of these, 17,728 instances were labeled as "locations" in the corpus, and 22,652 as "organizations" or "persons". The articles already present in the Overell's test set were removed from the training set. The dimensionality of the feature space (originally 44,180 features) was reduced using the *Transition Point* (TP) technique as described in [10] with a neighbourhood of 12.5% around the TP, obtaining 2,903 features.

The TP technique is based on the assumption that terms of medium frequency are closely related to the conceptual content of the document. Therefore, terms closer to the TP can be used as indices of a document. The formula used to obtain this value is given in Formula 5.

$$TP = \frac{\sqrt{8 * I_1 + 1} - 1}{2} \quad (5)$$

where I_1 represents the number of words with frequency equal to 1.

4. RESULTS

We carried out some experiments in order to determine the best values for the α threshold for botg DIce and cosine similarity measures. We obtained the best F-measure (computed as: $2 * precision * recall/(precision + recall)$) with $\alpha = 0.08$ and $\alpha = 0.06$, respectively.

In Table 1 we compare the best results (in terms of F-measure) for the Dice and cosine formulae to the results obtained with the Naive Bayes classifier and the rule-based method by [9], which exploits meta-data such as templates and category labels.

Method	Recall	Precision	F
Dice ($\alpha = 0.08$)	36.0%	29.7%	.325
Cosine ($\alpha = 0.06$)	50.0%	56.5%	.530
TP-NB	62.9%	42.8%	.509
Overell (comb.)	80.2%	80.3%	.803

Table 1: Comparison of the best (on F-measure) results obtained with the Dice and cosine similarity mesaures with the Naive Bayes classifier (TP-NB) using TP index reduction and the combined rule-based method by Overell.

The WordNet based method can outperform the Naive Bayes approach, when using the cosine similarity measure; from Figures **??** and **??** it can be noted that it can also obtain a higher precision than the rule-based method, but at cost of an extremely low recall.

In Table 3 we report the results obtained by selecting only the definition part of the article. This has been done simply by considering only the first sentence. In this case the

Method	Recall	Precision	F
Cosine ($\alpha = 0.04$)	69.6%	60.7%	.648
TP-NB	78.9%	33.6%	.471

Table 2: Results obtained considering only the first sentence in the articles.

[2] The Getty Thesaurus of Geographical Names, http://www.getty.edu/research/conducting_research /vocabularies/tgn/

WordNet-based method obtains an improvement in all measures, whereas the Bayesian classifier improves only in recall. This is compatible with the fact that the keywords extracted from WordNet are part of definitions.

On the other hand, the results obtained with the Bayesian classifier seem to indicate that the vocabulary used in Wikipedia for the geographical articles is not particularly different from the one used for other type of articles. In order to confirm this hypothesis, we calculated the perplexities of the 3-gram Language Model (LM) generated from the geographical and non-geographical sections of the training set with respect to the geographical part of the test set. The results are resumed in Table 3.

Language Model	Perplexity	Entropy
Training Set (Geo)	346.91	8.44
Training Set ($\neg$ Geo)	353.15	8.46

Table 3: Perplexity and Entropy of the LMs generated from the training set with respect to the geographical part of the test set.

The perplexity of a LM depends on the domain of discourse. Taking into account that, according to [11], typical perplexity values for narrow-domain text collections are smaller than 105, and the perplexity of general English has been measured to be 247, we can assert that there is not a great difference between the language used for geographical pages with respect to the other pages of Wikipedia.

The comparison with the results previously obtained by Overell and Rüger confirms that the information contained in the article's metadata is undoubtly more valuable than the text of the article itself.

5. CONCLUSIONS

We presented a method which is based on WordNet-extracted keywords related to the geographical domain. We evaluated the method using the Dice and the cosine similarity measures, with the second one resulting the best one. For some values of the α parameter, the cosine-based method outperforms a multinomial Naive Bayes classifier. However, the rule-based method by Overell and Rüger, based on page templates and category labels, prove to be more reliable than the other methods. This is due to the fact that it does not take into account the textual information in the pages, that in the majority of the cases is not particularly indicative of the geographical nature of the articles' contents.

6. REFERENCES

[1] D. Buscaldi, P. Rosso, and P. Peris. Inferring geographical ontologies from multiple resources for geographical information retrieval. In C. Jones and R. Purves, editors, *Proceedings of 3rd SIGIR Workshop on Geographical Information Retrieval*, August 2006.

[2] D. Buscaldi, P. Rosso, and E. Sanchis. Wordnet as a geographical information resource. In *Proceedings of the 3rd Global WordNet Association (GWA06)*, 2006.

[3] S. Cucerzan. Large scale named entity disambiguation based on wikipedia data. In *The EMNLP-CoNLL Joint Conference*, 2007.

[4] L. Denoyer and P. Gallinari. The Wikipedia XML Corpus. *SIGIR Forum*, 2006.

[5] G. Fu, C. B. Jones, and A. I. Abdelmoty. Bulding a geographical ontology for intelligent spatial search on the web. In *Proceedings of the IASTED International Conference on Databases and Applications*, 2005.

[6] C. B. Jones, R. Purves, A. Ruas, M. Sanderson, M. Sester, M. J. van Kreveld, and R. Weibel. Spatial information retrieval and geographical ontologies: An overview of the spirit project. In *Proceedings of the 25th Annual International ACM SIGIR Conference*, 2002.

[7] B. Martins, M. Chaves, and M. J. Silva. Assigning geographical scopes to web pages. In *Advances in Information Retrieval*, volume 3408 of *Lecture Notes in Computer Science*, pages 564–567. Springer, Berlin, 2005.

[8] G. A. Miller. Wordnet: A lexical database for english. In *Communications of the ACM*, volume 38, pages 39–41, 1995.

[9] S. Overell and S. Rüger. Identifying and grounding descriptions of places. In C. Jones and R. Purves, editors, *Proceedings of the 3rd SIGIR Workshop on Geographic Information Retrieval*, pages 14–16, August 2006.

[10] D. Pinto, H. Jiménez-Salazar, P. Rosso, and E. Sanchis. Buap-upv tpirs: A system for document indexing reduction at webclef. In S. Verlag, editor, *Accessing Multilingual Information Repositories, Revised Selected Papers CLEF05*, volume 4022, pages 873–879, 2006.

[11] S. Roukos. Language representation. *Cambridge Studies In Natural Language Processing Series*, pages 30–36, 1997.

[12] M. Sanderson and J. Kohler. Analyzing geographic queries. In *Proceedings of the 1st SIGIR Workshop on Geographic Information Retrieval*, 2004.