
Bootstrapping Multilingual Geographical Gazetteers from Corpora

MARIEKE VAN ERP

ILK/Language and Information Science, Tilburg University

M.G.J.vanErp@uvt.nl

ABSTRACT. In this paper an approach to automatically generating multilingual geographical name gazetteers via two bootstrapping loops on different corpora is presented. First, a small seed-list of geographical names is matched to an unannotated dataset in one language, and training data for a memory-based classifier is generated. Memory-based learning is applied to extend the gazetteer. Then a cross-over to a different language is made by matching this extended gazetteer to a corpus in a different language. Again, training data for a classifier is generated and the bootstrapping process is repeated in order to extend the gazetteer further. This process is quite similar to co-training, in which information from other sources is introduced to enhance classification. To estimate the difference between the initial seed-list and the final gazetteer and thereby to evaluate the performance of the algorithm, they were matched to three datasets with manually annotated geographical entities.

1 Introduction

In corpus-based natural language processing one hardly ever finds clean data. A problem with real data can be that it is made up of data in several languages; this occurs for instance on webpages. Multilingual data may pose problems to natural language processing tools if one wants to extract information from these sources. One of the first tasks in information extraction is Named Entity Recognition (NER), the branch of information extraction that is concerned with identifying and classifying expressions that refer to named entities, such as people, organisations and companies, and geographical locations. It discerns the who, where, and what elements in a text, which can be seen as one of the first step towards further natural language processing tasks such as text summarisation, question answering and ultimately machine translation. At the seventh Message Understanding Conference (MUC) in 1998 a special track was devoted to NER. Since then much work has been done in the field of NER, demonstrating a wide variety of approaches. For comparisons of NER systems several competitions at conferences and workshops such as LREC¹ and CoNLL (Tjong Kim Sang 2002; Tjong Kim Sang and De Meulder 2003) have been organised.

One approach that is fairly simple and seems to be particularly well-suited for monolingual geographical NER is the use of gazetteers (Mikheev et al. 1999). One can start

¹<http://www.lrec-conf.org/>

off with a small seed-list of geographical names and extend that by applying machine learning techniques to increase recall. Bootstrapping gazetteers is fairly common practice nowadays, see for instance: Jones et al. (1999), Niu et al. (2003), Pratim Talukdar et al. (2006), Riloff and Jones (1999), and Uryupina (2003). However, NER systems are generally language-dependent and thus not suitable for the abundance of multilingual data nowadays found on, for instance, the World Wide Web. The ability to deal with multilingualism is needed for tasks such as cross-lingual information extraction (Riloff et al. 2002) or for Natural Language Processing (NLP) tasks on multilingual textual databases.² In this paper a language-independent approach to the problem of recognising geographical entities is proposed. To the author's knowledge, bootstrapping gazetteers has not been done cross-lingually in order to create a gazetteer that can deal with multilingual text. The aim of this work is to investigate the possibility of inducing a multilingual gazetteer by a bootstrapping process from unannotated data in English, Dutch and German, starting with a small English seed-list. Apart from providing a possible approach to dealing with named entities from multilingual sources, bootstrapping from multilingual source data may be beneficial for monolingual NER as well, as a corpus in a different language may contain useful information that is not present in the corpus in the first language. This assumption has been found useful for various NLP tasks such as automatic verb classification (Merlo et al. 2002), machine translation (Callison-Burch and Osborne 2003) and word sense disambiguation (Diab and Resnik 2002). Using different languages to aid classification can also be compared to co-training, as proposed by Blum and Mitchell (1998), who classified webpages using the content and the hyperlinks pointing to that page as different input feature spaces for the same classification task. In the present paper, the information different information sources are the different languages.

The choice for recognition of geographical named entities was made because gazetteers were found to be particularly useful for this NE-class (Mikheev et al. 1999).

2 Approach

2.1 Memory-based learning

Memory-based Learning (MBL) is an approach that is based on the idea that the direct use of examples is a better method to learn a solution to certain tasks than learning from rules deduced from examples. In the first phase labelled training examples are presented to the classifier. This set is treated as a collection of points in a multi-dimensional feature space, which is stored in the memory as an instance base. In the second phase, unseen and unlabelled test examples are classified by matching them to every instance in the instance base, calculating the matching distance between the new instance and every instance in the memory using a distance function. A class label is then assigned to the new instance according to the distance (Daelemans and Van den Bosch 2005). The approach in this work

²Such databases are found at the Dutch National Museum for Natural History and undoubtedly at many other places.

uses a k -Nearest Neighbour classifier (k -NN) (Dasarathy 1991). A class label is assigned by selecting the k examples with the smallest distance to an instance. In this experiment the nearness is calculated via the overlap metric (equations 1.1 and 1.2) but various metrics are applicable depending on the nature of the data. $\Delta(X, Y)$ is the distance between instances X and Y , both represented by n features, with δ the distance per feature. The distance between two instances is the sum of the differences between the features.

$$\Delta(X, Y) = \sum_{i=1}^n \delta(x_i, y_i) \quad (1.1)$$

where:

$$\delta(x_i, y_i) = \begin{cases} abs & \text{if numeric, else} \\ 0 & \text{if } x_i = y_i \\ 1 & \text{if } x_i \neq y_i \end{cases} \quad (1.2)$$

The implementation used in this paper is the TiMBL package (Daelemans et al. 2004).

2.2 Data

Three corpora of closely related languages were used for this experiment: an English, a Dutch, and a German corpus. In order to keep the experiments as language-independent as possible no preprocessing steps other than tokenising the data were undertaken. In the remainder of this section the characteristics of each corpus are described.

The Reuters Corpus Volume 1 (RCV1) For English we used RCV1³ which contains approximately 810,000 English news articles from the Reuters press agency. It contains newswire stories from August 20, 1996 until August 19, 1997, covering a wide variety of topics such as corporate or industrial news, economics, war, and sports.

The ILK corpus The Dutch data used in the work described in this paper comes from the ILK Corpus.⁴ This corpus was gathered by the Tilburg Induction of Linguistic Knowledge research group from various southern Dutch regional newspapers between 1985 and 1998. It consists of about 230,000 articles that together contain approximately 120 million words. The corpus has been partly annotated with prosody markers, named entities and it is NP-chunked, although these annotation layers were not included in this experiment because the other corpora were not annotated with this information.

The Frankfurter Rundschau corpus The last corpus used for this work is the part of the Frankfurter Rundschau Corpus that was made available for the Elsnets European Corpus Initiative.⁵ It consists of German newspaper texts from the Frankfurter Rundschau from July 1992 until March 1993. It contains about 34 million words.

³<http://about.reuters.com/~researchandstandards/corpus/statistics/>

⁴<http://ilk.uvt.nl/ilkcorpus/>

⁵<http://www.elsnet.org/resources/~eciCorpus.html>

2.3 Bootstrapping Gazetteers

The approach undertaken in this work consists of two parts: a language internal bootstrapping loop and a cross-lingual bootstrapping loop. The language internal bootstrapping loop is no different from previous monolingual geographical entity recognition bootstrapping work: a classifier is trained on data automatically labelled with the help of a small gazetteer. The harvest set, containing unseen and unlabelled instances, is then presented to the classifier for which the classifier needs to predict class labels. The items classified as geographical entities are added to the gazetteer and the whole process can be rerun to further expand the gazetteer.

This initial gazetteer, or seed-list, contains 25 items which have been selected manually on the basis of their perceived frequency in the global news. It contains the following geographical names:

| | | | | |
|----------|--------------------------|-------------|-----------|---------------|
| New York | United States of America | U.S. | Mexico | Chile |
| Paris | Rio de Janeiro | Brazil | The Hague | Great Britain |
| Tokyo | China | Taiwan | Taipei | Beijing |
| Rome | Santiago de Compostela | Afghanistan | Barcelona | Iraq |
| Sydney | Los Angeles | Washington | Pakistan | Buenos Aires |

To decide whether a word is a geographical named entity every item in the gazetteer is checked against every word in the dataset. Upon a match the word is assigned the label ‘GN’ for ‘geographic name’, else ‘O’ for ‘other’. Apart from the word and its label, the feature vector for each word also contained contextual and orthographic information. The context was encoded in a 2-1-2 context window, meaning that for each instance the two words before and the two words after the word that is to be classified are given. The other feature that was included here is a marker for whether or not the focus word is capitalised. A small part of the generated training instances is shown in Example 1.1.

$$\begin{aligned}
 & _,_,Emerging,evidence,that,+,O \\
 & _,Emerging,evidence,that,Mexico,-,O \\
 & Emerging,evidence,that,Mexico,'s,-,O \\
 & evidence,that,Mexico,'s,economy,+,GN \\
 & that,Mexico,'s,economy,was,-,O \\
 & Mexico,'s,economy,was,back,-,O \\
 & 's,economy,was,back,on,-,O
 \end{aligned}
 \tag{1.1}$$

The capitalisation feature was ignored for German because for this language capitalisation is not an informative cue for the presence of a named entity, due to the fact that all nouns are capitalised. Other features, such as whether words surrounding the focus word were capitalised, did not aid classification and were therefore abandoned.

The training instances were generated from 85% of the data for all three languages, the harvest sets from the remaining 15%. To speed up the experiments and to reduce the influence of false negatives, the ratio positive to negative examples was set to 1:5. The false negatives in the labelled data occur because not all geographical entities are

recognised in the first labelling round due to the small gazetteer. From a phrase like “between representatives of England, Wales, Scotland and Ireland” no positive instances will be generated because the seed-list does not contain the items “England”, “Wales”, “Scotland” or “Ireland”.

As well as adjusting the proportion of positive and negative examples, only sentences that contained a geographical entity recognised through the gazetteer were included, as with these a sufficient amount of negative instances could already be generated. In order to not introduce too much noise in the form of false positives in the automatically extracted gazetteer, precision was valued higher than recall.

The classifier was trained on the automatically labelled data and then applied to unseen and unlabelled data. Terms which were labelled as geographical named entities by the classifier were considered candidates to be added to the gazetteer. Since the classifier is not perfect we sought means to extract the items that were classified as geographical names with a high degree of certainty. To this end 4 filters were developed. The first filter checked whether a token that was labelled as a geographical entity had also been labelled as a non-geographical entity, if this was the case the token was discarded. The word “City”, for instance, occurs quite frequently as suffix in for example “Atlanta City” or “New York City”. However, if this word occurs on its own it is not a geographical named entity and must therefore be discarded. The second filter performed the following check: if a capitalised token classified as a geographical entity also occurred non-capitalised in the unlabelled set it was discarded; this also implies that a geographical name needs to be capitalised. Then, completely capitalised items as well as items of three letters and shorter were excluded. Finally, a threshold was put up to exclude items that occurred fewer than 5 times in the harvest set.

For the cross-lingual bootstrapping loop, the gazetteer from the first language (English) is matched to data in another language, in this experiment Dutch or German, to label data to train a classifier on. This yields instances labelled as geographical names because many geographical names, such as ‘Amsterdam’ are the same across different languages.

Ultimately one would like to perform fuzzy matching of geographical names in the cross-lingual bootstrapping loop, e.g. exploiting measures such as Levenshtein distance (Levenshtein 1965) to match “England” to its Dutch form “Engeland”. However, this would also yield many false positives such as “France” and “Francs”. Hence for the present paper we restricted the cross-lingual bootstrapping loop to strict matching.

3 Experiments and Results

3.1 Monolingual Experiments

In the first series of experiments the original seed-list was applied to all three data sets in three separate experiments to get an idea of how well the small English oriented seed list works for bootstrapping on the three languages. It also serves as a baseline on which the cross-lingual bootstrapping should improve. Table 1.1 shows how many candidates for

| | run 1 | new | run 2 | new | run3 | new | total | TP (%) |
|----------------|-------|-----|-------|-----|------|-----|-------|-------------|
| Reuters | 47 | 47 | 47 | 2 | 47 | 0 | 74 | 53 (77.0%) |
| ILK | 117 | 116 | 147 | 51 | 166 | 24 | 216 | 156 (76.9%) |
| Frankfurter R. | 16 | 16 | 17 | 3 | 17 | 0 | 44 | 41 (93.2%) |

Table 1.1: Results of the monolingual bootstrapping runs

extending the gazetteer there were per run (“run 1”, “run 2”, “run 3”), as well as the number of unique and new items found in each run (“new”).

In the first run the classifier is trained on the initial seed, in the second run on the gazetteer that was created in the first run and in the third run the classifier is trained on the gazetteer that was created in the second run. The penultimate column in Table 1.1 shows the total number of items the gazetteer contains after the three runs, i.e., the initial seed-list plus the new items from the first, second and third run. The last column (“TP”) shows the percentage of true geographical entities or parts of geographical entities in the gazetteer. The items in the gazetteers were checked manually against atlases to determine whether they are true geographical names or not.

Although we attempted to include multiword entities this has only worked to some degree. Often parts like “River” in for instance “Yangtze River” have not been classified as geographical entities because they occur more often not as a named entity. The unlabelled sets (“harvest sets”) from which the new items were harvested consist of 1,000,000 instances for each language. As can be seen in Table 1.1, the number of new gazetteer items decreases sharply in each run, indicating that if no more data is added, language internal bootstrapping quickly leads to a dead end. Adding more instances to the harvest set could have been a solution here had it not been for the unacceptable number of false positives this yields in this experiment. This is due to the post-processing filter that removes items that have not been classified as geographical names often enough. Raising the threshold proportionally to the increase in instances yields similar results as the first internal bootstrapping experiment in which the harvest set has been kept the same throughout the runs.

3.2 Bilingual Experiments

A series of bilingual experiments was conducted to investigate the influence of one other language on the English gazetteer. To this purpose the gazetteer that was created in the first run of the language internal experiment on the English corpus was applied to Dutch, the labelled data was used to train a classifier which was then applied to the Dutch harvest set. After applying the same filters as in the language internal experiments the results were added to the gazetteer. This gazetteer was then applied to the English corpus and the training, classification and filtering were repeated. These three experiments were also carried out for English/German instead of English/Dutch. The results are shown in Table 1.2, where E stands for English, D for Dutch and G for German. The column “run 1” gives the results for the first parts of the experiments (E→D and E→G), “run

| | run 1 | new | run 2 | new | total | TP (%) |
|-------|-------|-----|-------|-----|-------|-------------|
| E→D→E | 124 | 122 | 60 | 22 | 216 | 143 (66.2%) |
| E→G→E | 17 | 17 | 47 | 2 | 91 | 73 (80.2%) |

Table 1.2: Results of the bilingual bootstrapping runs

| | run 1 | new | run 2 | new | run 3 | new | total | TP (%) |
|---------|-------|-----|-------|-----|-------|-----|-------|-------------|
| E→D→G→E | 124 | 122 | 9 | 9 | 48 | 3 | 206 | 149 (72.3%) |
| E→G→D→E | 17 | 17 | 129 | 127 | 49 | 3 | 219 | 162 (73.1%) |

Table 1.3: Results of the trilingual bootstrapping runs

2” gives the results for the second parts (D→E and G→E). In the column “total” the number of items after these bootstrapping loops is given, i.e. the gazetteer from the English monolingual bootstrapping experiment plus the new items from the bilingual bootstrapping runs. The column “TP” again gives the number of true geographical named entities per final gazetteer, measured manually. In both cases the number of items added to the English gazetteer is greater than in the monolingual experiments. However, more false positives, often person names such as “Adriaanse” and “Cathy”, are added to the gazetteer. Especially the English/Dutch gazetteer gets particularly corrupted although the number of true positives is still over twice as much as in the monolingual English gazetteer (143 against 53).

3.3 Trilingual Experiments

The third series of experiments concern a trilingual bootstrapping loop. Two different loops have been investigated: English→Dutch→German→English and English→German→Dutch→English. The setup of the experiments is similar to the bilingual experiments, with the addition of an extra experiment on a third language. The results are shown in Table 1.3. Compared to the English/Dutch gazetteer the precision has increased, which is probably due to the conservative behaviour of the German classifier. Also the number of true positives has increased to 162 for the English-to-German-to-Dutch loop, indicating that conservativeness especially in the early runs seems to pay off.

3.4 Recall

In order to estimate the recall of lookup with the different gazetteers, the seed list and gazetteers from the experiments were matched to the Dutch CoNLL shared task 2002 test set and the English and German CoNLL 2003 test sets (Tjong Kim Sang 2002; Tjong Kim Sang and De Meulder 2003). The results are presented in Table 1.4. The first number in each column is the recall, the second (in brackets) the precision. The numbers in brackets behind the names of the test sets is the number of geographical names present in that test set. As can be expected with the still small gazetteers, also after bootstrapping, it comes

| | English (1660) | Dutch (772) | German (1285) |
|-----------|----------------|--------------|---------------|
| seed list | 9.4% (100%) | 2.8% (100%) | 0.5% (100%) |
| English | 9.8% (100%) | 3.1% (100%) | 0.5% (100%) |
| Dutch | 11.3% (99.5%) | 12.7% (100%) | 2.7% (65.7%) |
| German | 9.5% (100%) | 3.1% (100%) | 0.5% (100%) |
| E→D→E | 12.3% (99.5%) | 6.9% (100%) | 2.3% (80.0%) |
| E→G→E | 11.3% (99.5%) | 6.6% (100%) | 2.3% (79.3%) |
| E→D→G→E | 11.7% (99.5%) | 6.8% (100%) | 2.3% (79.3%) |
| E→G→D→E | 11.3% (100%) | 8.4% (100%) | 2.3% (100%) |

Table 1.4: Recall and precision (in brackets) on CoNLL shared task test sets

as no surprise that recall on the CoNLL datasets is very low. The German test set proves to contain most unknown geographical entities, although some of its geographical named entities such as “Anne-Frank-Schule” (“Anne Frank School”) do not occur in atlases and thus do not fall within our notion of geographical named entities. The goal to focus on precision rather than recall has been reached as in most cases no non-geographical names were flagged, indicating that the false positives in the gazetteers are not words that occur very frequently.

4 Conclusions and Future Work

Previous work has shown that bootstrapping is a suitable technique to label unseen data, when an iterative labelling scheme is used that feeds back to a classifier. The results in this work have shown that bootstrapping geographical entities with a memory-based learner can also be used in a cross-linguistic setting. Where precision decreases in monolingual bootstrapping if too much data is added and where the bootstrapping reaches a dead end if no more data is added, cross-lingual bootstrapping provides a way out, with the additional property of being portable to another language. Although the gazetteer does not only contain English geographical entities after the cross-lingual bootstrapping loop such as the German form “Genf” for “Geneva”, it shows only a minor increase in recall when applied to an English text – but with a high precision. Moreover, the multilingual gazetteers also seem to work for Dutch and German texts. To make these gazetteers more useful for for instance cross-lingual information extraction a means to link the different names for entities in different languages, such as “Vienna” to “Wenen” and “Wien” needs to be found. For some entity pairs like “Ireland”, “Ierland” and “Irland”, this might be relatively easy as the words are very similar but for other pairs such as “Geneva” and the German form “Genf”, or for “Germany” and its native name “Deutschland”, contextual information is needed.

Since the focus of this work has been on precision one aim for future work is to explore this technique further to come to better results on recall. In order to do this we will further experiment with different filters and different cross-overs. As more sophisticated filters are

applied it might also be possible to experiment with fuzzy matching, which may also help link up the entities in different languages. Another interesting avenue of research is to investigate how portable this approach is to other languages. Intuitively this approach should also work for other sets of related languages, such as Spanish, Italian and French, or Norwegian, Swedish and Danish. It is also interesting to find out what effect adding or substituting one language with a slightly more distant language such as French would have on this approach.

Acknowledgements

This research is funded by NWO (Netherlands Organisation for Scientific Research) and carried out at the Naturalis Research Labs in Leiden. The author would like to thank Antal van den Bosch and Caroline Sporleder for discussions and advice, and the anonymous reviewers for their helpful comments.

Bibliography

- Blum, A. and T. Mitchell (1998). Combining labeled and unlabeled data with co-training. In *Proceedings of the 11th Annual Conference on Computational Learning Theory*, pp. 92–100.
- Callison-Burch, C. and M. Osborne (2003). Co-training for statistical machine translation. In *Proceedings of the 6th Annual CLUK Research Colloquium*, Edinburgh, Scotland.
- Daelemans, W. and A. Van den Bosch (2005). *Memory-based language processing*. Cambridge: Cambridge University Press.
- Daelemans, W., J. Zavrel, K. Van der Sloot, and A. Van den Bosch (2004). TiMBL: Tilburg Memory Based Learner version 5.1, reference guide. Technical Report ILK Technical Report Series 04-02, ILK/Tilburg University, Tilburg, The Netherlands.
- Dasarathy, B. V. (1991). *Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques*. Los Alamitos, CA: IEEE Computer Society Press.
- Diab, M. and P. Resnik (2002). An unsupervised method for word sense tagging using parallel corpora. In *Proceedings of the 40th Annual Meeting of the Association of Computational Linguistics (ACL'02)*, Philadelphia, PA, pp. 255–262.
- Jones, R., A. McCallum, K. Nigam, and E. Riloff (1999). Bootstrapping for text learning tasks. In *IJCAI-99 Workshop on Text Mining: Foundations, Techniques and Applications*, pp. 52–63.
- Levenshtein, V. I. (1965). Binary codes capable of correcting deletions, insertions, and reversals. *Cybernetics and Control Theory* 10(8), 707–710. Originally in: *Doklady Nauk SSSR* 163(4): 845-848 (1965).
- Merlo, P., S. Stevenson, V. Tsang, and G. Allaria (2002). A multilingual paradigm for automatic verb classification. In *Proceedings of the 40th Annual Meeting of the Association of Computational Linguistics (ACL'02)*, Philadelphia, PA, pp. 207–214.
- Mikheev, A., M. Moens, and C. Grover (1999). Named entity recognition without gazetteers. In *Proceedings of EACL*, pp. 1–8.
- Niu, C., W. Li, J. Ding, and R. K. Srihari (2003). A bootstrapping approach to named entity classification using successive learners. In *Proceedings of the Association for Computational Linguistics (ACL)*, pp. 335–342.

- Pratim Talukdar, P., T. Brants, M. Liberman, and F. Pereira (2006). A context pattern induction method for named entity extraction. In *Proceedings of the Tenth Conference on Computational Natural Language Learning (CoNLL-X)*, New York, NY.
- Riloff, E. and R. Jones (1999). Learning dictionaries for information extraction by multi-level bootstrapping. In *Proceedings of the Sixteenth National Conference on Artificial Intelligence (AAAI-99)*, pp. 474–479.
- Riloff, E., C. Schafer, and D. Yarowsky (2002). Inducing information extraction systems for new languages via cross-language projection. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING'02)*, Taipei, Taiwan.
- Tjong Kim Sang, E. F. (2002). Introduction to the CoNLL-2002 shared task: Language-independent named entity recognition. In *Proceedings of CoNLL-2002*, pp. 155–158.
- Tjong Kim Sang, E. F. and F. De Meulder (2003). Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of CoNLL-2003*, pp. 142–147.
- Uryupina, O. (2003). Semi-supervised learning of geographical gazetteers from the internet. In *Proceedings of the HLT-NAACL Workshop on the Analysis of Geographic References*, pp. 18–25.