Automatic Classification of Geographical Named Entities

Daniel Ferrés[*], Marc Massot[†], Muntsa Padró[*],
Horacio Rodríguez[*], Jordi Turmo[*]

[*] TALP Research Center
Universitat Politècnica de Catalunya
C/ Jordi Girona 1-3
08034 Barcelona, Spain
{dferres, mpadro, horacio, turmo}@lsi.upc.es

[†] Dept. d'Informàtica i Matemàtica Aplicada
Universitat de Girona
Edifici P4, Campus Montilivi
17071 Girona, Spain
marc@ima.udg.es

Abstract

Performing accurate Named Entity (NE) classification (NEC) has recently become a central issue in many NLP applications, such as Information Extraction and Question Answering, among others. Most state-of-the-art NEC systems use coarse-grained MUC-style datasets for performing the NEC task reducing it to distinguish among LOCATION, PERSON, ORGANIZATION and so. There is, however, a growing interest on using finer-grained classification sets. This paper describes a methodology that applies Machine Learning techniques for a finer-grained classification of NEs that have been previously classified as locations by a NERC system.

1. Introduction

Performing accurate Named Entity (NE) recognition (NER), classification (NEC) and disambiguation (NED) has recently become a central issue in many basic NLP tasks, as co-reference resolution, document linking or topic detection, currently present in most NLP applications such as Automatic Summarization, Question Answering, Document Classification and Filtering, and Information Extraction among others. Most state-of-the-art NEC systems use coarse-grained MUC-style datasets for performing the classification task reducing it to distinguish among LOCATION, PERSON, ORGANIZATION and so. There is, however, currently, a growing interest on going beyond using finer-grained classification sets. (Sekine et al, 2002), for instance, use an extended NE hierarchy of 150 types, while (Manov et al, 2003) use 97 classes for the location sub-ontology. Although some of the efforts in this direction are devoted to the general classification problem, as in (Mann, 2002), or are devoted to the personal name disambiguation problem, as in (Mann, 2003), most work has been done developing tools and resources related to the management of geographical references. Among others, these efforts have been applied to:

- (semi) automatic building of large-scale geographical gazetteers from corpora, ontologies and other lexical resources, as the Alexandria Digital Library Gazetteer (ADLP) covering about 5 million of geographical terms, or the Metacarta GazDB (Axelrod, 2003).

- Finer grained forms of NEC, as the Perseus system (Smith et al, 2001) for geographical NEs.

- Grounding of geographical NEs, i.e. mapping a geographical NE to its appropriate physical (spatial) location (coordinates, area, etc.), as in (Leidner et al, 2003).

Many different techniques have been applied for such purposes, from knowledge intensive to empirical/statistical approaches, using strong or weak supervised learning or bootstrapping.

What is presented here is a system that applies Machine Learning techniques for a finer grained classification of NEs that have been previously classified as locations by a general purpose NERC system.

The core of our system is an Inductive Logic Programming (ILP) learner that learns, from a set of positive and negative examples, a ranked list of rules to obtain a binary classifier for each geographical class. Both natural geographical entities (Sea, Mountain, River, etc.) and political or organizational divisions (Country, State, Province, City, etc.) are considered. Our learner (we have used Quinlan's FOIL (Quinlan, 1990)) follows a supervised schema, so a training set has been collected and automatically tagged. What has to be learned is the dependence of the different types of location on the context of their occurrences.

After this introduction the paper is organized as follows. Section 2 describes the method applied to classify geographical NEs. Section 3 shows the obtained results using this method. Section 4 states some conclusions and further work.

2. Approach

In the system presented here, the following approach has been applied:

- Firstly, an initial set of sources of highly confident classified resources has been selected. We have used the MUC6 Reference Gazetteer complemented with location names extracted from five different web sites (see Table 1). The information extracted includes not only the basic terminological information (i.e. lists of tagged NEs) but also some spatial relations (e.g. states in a country, islands in a sea, etc.). Up to 133,744 geographical names classified into 18 classes have been extracted in this way (with a very irregular distribution, from 117,598 cities to only 3 forests). Table 2 shows the number of names per class.

http://www.world-gazetteer.com/
http://people.depauw.edu/djp/
http://www.worldatlas.com/
http://en.wikipedia.org/
http://www.gazeteer.com/

Table 1: Web sites used to extract the gazetteer.

Classes	Number
Airport	729
City	117,598
Country	303
Country-zone	220
Desert	43
Forest	3
Gulf	22
Island	917
Island-sea	698
Lake	47
Mountains	27
Peak	2,218
Port	4,641
Province	5,331
River	333
Sea	45
State	530
Volcano	39
Total	133,744

Table 2: Number of geographical names per class.

- From this initial set we have removed all the NEs belonging to more than one class in order to reduce, as much as possible, the use of contexts corresponding to ambiguous NEs.
- We have merged the classes with few members and semantically related (e.g. port and airport, mountain and peak), dropped out poorly represented classes and selected a maximum of 500 names per class. In addition, we have performed a shallow manual revision. A total of 11 classes remained after this step: mountain or peak, river, sea, lake, island, desert, port or airport, city, country, state and province.
- We have looked for the first 500 occurrences of the members of these lists in the AQUAINT [1] corpus. A previous preprocess was carried out including POS tagging (Brants, 2000) and NERC (Carreras et al., 2003). Restricting the number of names per class and the number of examples per name is needed for getting the resulting set as balanced as possible.
- From these corpus we have extracted the context, up to 10 tokens on each side, of each occurrence as well as the needed morphological information. This procedure resulted in a total of 110,576 examples (see last column in Table 4).

With this material we have fed FOIL (Quinlan, 1990) to learn one classifier for each class. FOIL is a relational learning system aimed at inductively learning first-order rules (in prolog format) from positive and negative examples. By default, FOIL considers the *close-world assumption* to automatically generate the set of negative examples, meaning that all non-positive elements are negative ones. We have used, however, the examples corresponding to each particular class as positive examples for learning this class and the examples related to the rest of classes as negative ones. This experimental setting has proved to provide better results than the multiclass approach with *close-world assumption*. With respect to the background knowledge used to learn, FOIL requires each of the examples, positive and negative ones, to be represented as a set of predicates. For our particular learning problem we have used the features presented in table 3 from wich the following set of propositional predicates has been designed:

- Context predicates:
  - diw_x: the ith word in the direction d (right or left) is x.
  - dip_x: the ith POS-tag in the direction d is x.
  - dis_x: x is the NE class (LOC, ORG, PER and MISC) about ith word in the direction d.
- Internal NE predicates:
  - zi_x: the ith token of the NE is x (i could be 1 or 0, for the two last tokens of the NE)

In all cases, with the exception of case three, i can be omitted, it means that the information appears in any position in the direction d.

Feature type	Features
lexical information	- Bag of words of positions -5 to +5 (NE not included). - Words in position from -1 to -3. - Words in position from +1 to +2. - Two last Tokens included in the NE.
morphological information	- Bag of POS of positions -5 to +5 (NE not included). - POS in position from -1 to -3. - POS in position from +1 to +2.
semantic information	- NE class of positions -1 to -3. - NE class of positions +1 to +2

Table 3: Features used by FOIL.

[1] The corpus has been used for our participation in TREC-2003. More information about AQUAINT corpus can be obtained at http://www.ldc.upenn.edu/Catalog/docs/LDC2002T31

3. Experiments

We haved designed a set of three experiments to decide which predicates are the best to learn to classify geographical NEs. In these experiments we have only changed the features related to the NE (i.e. internal predicates as tokens *z1_Lake* and *z0_Garda* in the case of *Lake_Garda*), and we do not have modified context predicates. All the experiments have had the same set of context predicates. The following experiments have been done:

1. Experiment with all the predicates previously explained.

2. Experiment only with context predicates.

3. Experiment with all context predicates and using internal predicates only for NEs having more than 1 token (i.e. *Lake_Garda*):

FOIL has learned a set of binary classifiers for each class. We have used the k-Fold Cross-Validation measure to evaluate these classifiers. The k parameter means the number of sets to split the examples, k has been set to 5. We have balanced the number of examples used to learn the classifiers taking a treshold of 1200 in classes with many examples (see column 2 of Table 4).

Classes	#Examp. (5CV)	#Examp. (total)
Airport+Port	376	376
City	1,200	25,000
Country	1,200	25,000
Desert	517	517
Island	1,200	7,259
Lake	1,447	1,447
Mountains+Peak	1,186	1,186
Province	1,200	23,399
River	1,200	5,189
Sea	1,200	20,050
State	850	850
Total	11,576	110,273

Table 4: Number of examples used in 5-CV and total.

4. Results

The results of the three 5-fold cross-validation experiments are summarized in tables 5, 6 and 7. These tables contain the following average evaluation measures of 5-fold test sets for each class: precision, recall, F_1[2] and the variance of F_1. As shown in these tables, experiment 1 achieves the best overall performance with an 0.9553 average measure of F_1. It has produced 613 rules for all classes (an average of 55.72 rules per class). However, experiment 1 uses internal predicates that produce overfitting. These predicates are the last two tokens of the NE (i.e. *z0_York*, *z1_New*, *z_York* and *z_New*). Using these predicates can be useful to capture some rellevant features of the NE (i.e. capturing *z0_River* in *Colorado_River*), especially in these classes: airport+port (*z1_Airport*,*z0_Port*), desert (*z0_Desert*), lake (*z1_Lake*), mountains+peak (*z1_Mount*,*z0_Peak*), river (*z0_River*) and sea (*z0_Sea*, *z0_Ocean*). Besides, in the case of NEs having only one token it can affect negatively to the learning rules (i.e. capturing *z0_Sahara* in *Sahara*). Concluding, we cannot obtain robust rules using internal predicates with NEs having only one token.

Classes	Precision	Recall	F_1	varF_1
airport+port	0.7850	0.9632	0.8509	0.0086
city	0.8293	0.9475	0.8821	0.0006
country	0.8796	0.9017	0.8883	0.0014
desert	0.9980	0.9942	0.9961	0.0000
island	0.9908	0.9817	0.9862	0.0000
lake	1.0000	1.0000	1.0000	0.0000
mountains+peak	0.9873	0.9601	0.9729	0.0004
province	0.9341	0.9550	0.9440	0.0003
river	0.9992	0.9950	0.9971	0.0000
sea	1.0000	0.9983	0.9992	0.0000
state	0.9873	0.9953	0.9912	0.0000
Total Avg	0.9446	0.9720	0.9553	0.0011

Table 5: Results of 5-fold cross validation with internal predicates (Experiment 1).

Classes	Precision	Recall	F_1	varF_1
airport+port	0.7544	0.8222	0.7729	0.0249
city	0.6460	0.8183	0.7146	0.0007
country	0.6657	0.8833	0.7557	0.0010
desert	0.6271	0.7851	0.6954	0.0009
island	0.7228	0.8600	0.7839	0.0005
lake	0.6107	0.8527	0.7044	0.0008
mountains+peak	0.6552	0.6655	0.6587	0.0118
province	0.6306	0.9075	0.7391	0.0021
river	0.7842	0.9108	0.8400	0.0005
sea	0.7294	0.8817	0.7959	0.0010
state	0.6311	0.8188	0.7108	0.0004
Total Avg	0.6779	0.8369	0.7429	0.0041

Table 6: Results of 5-fold cross validation with only context predicates (Experiment 2).

Classes	Precision	Recall	F_1	varF_1
airport+port	0.8161	0.9082	0.8413	0.0087
city	0.6975	0.8367	0.7606	0.0003
country	0.7465	0.8408	0.7894	0.0012
desert	0.8201	0.7935	0.7981	0.0055
island	0.7290	0.8917	0.8004	0.0006
lake	1.0000	1.0000	1.0000	0.0000
mountains+peak	0.9791	0.9516	0.9644	0.0008
province	0.6431	0.7225	0.6803	0.1162
river	0.9950	0.9908	0.9929	0.0000
sea	0.9224	0.9442	0.9311	0.0007
state	0.7556	0.9059	0.8223	0.0009
Total Avg	0.8277	0.8896	0.8528	0.0123

Table 7: Results of 5-fold cross validation with reduced internal predicates (Experiment 3).

More and most robust rules have been learned with experiments 2 and 3. These rules have been reported the following measures of F_1 in average 0.7429 and 0.8528 respectively. These latter experiments have produced 4695

[2] F_β is the harmonic mean of recall (ρ) and precision (π) (van Rijsbergen, 1979). The F_β function formula is: $F_\beta = \frac{(\beta^2+1)\pi\rho}{\beta^2\pi+\rho}$.

and 2139 rules, respectively, with an average of 426.62 and 194.45 rules per class, respectively. Examples of the best ranked rules obtained for desert class can be seen in figures 1 and 2.

One of the advantage of using an ILP system as FOIL is the readibility of learned rules. This property allows to easily analyze these rules and modify or remove those which are considered irrelevant ones. This is the case of rules:

desert(A) :- l1w_the(A), rp_RB(A), not(lp_NN(A)).

desert(A) :- lw_villages(A).

The following rules can be considered relevant rules:

desert(A) :- l1w_the(A), not(z0_River(A)), not(z0_Sea(A)), not(z0_Mountains(A)), not(rs_NNP(A)), not(ls_VB(A)), not(z0_State(A)), rs_IN(A), not(r1p_IN(A)).

desert(A) :- rw_desert(A).

desert(A) :- rw_desert(A).
desert(A) :- l1w_the(A), rp_RB(A), not(lp_NN(A)).
desert(A) :- l1w_the(A), not(rp_NNP(A)), lp_VBN(A), not(lp_NN(A)), not(rp_VB(A)).
desert(A) :- l1w_the(A), not(rp_NNP(A)), not(rp_JJ(A)), rp_VBZ(A), not(lp_JJ(A)), not(r2p_VBZ(A)).
desert(A) :- lw_the(A), not(lp_NNS(A)), l3p_RB(A), not(lp_VBN(A)).
desert(A) :- not(rp_NNP(A)), l2w_in(A), not(r1p_IN(A)), rp_NN(A), not(rp_,(A)), not(r1p_NN(A)).
desert(A) :- lw_the(A), rp_,(A), rp_DT(A), not(r2w_the(A)), not(l2p_IN(A)), not(lw_in(A)).
desert(A) :- lw_the(A), rp_,(A), rp_RB(A), not(l3p_NN(A)), not(rp_VBD(A)).
desert(A) :- lw_desert(A), not(rw_of(A)).
desert(A) :- l2p_IN(A), rw_in(A), not(l3p_NNS(A)), not(l3p_NN(A)).

Figure 1: First rules obtained with only context predicates (Experiment 2).

desert(A) :- l1w_the(A), not(z0_River(A)), not(z0_Sea(A)), not(z0_Mountains(A)), not(rp_NNP(A)), not(lp_VB(A)), not(z0_State(A)), rp_IN(A), not(r1p_IN(A)).
desert(A) :- z0_Valley(A).
desert(A) :- rw_desert(A).
desert(A) :- l1w_the(A), not(z0_River(A)), not(z0_Sea(A)), not(rs_IN(A)), lw_of(A), not(r2p_RB(A)).
desert(A) :- not(z0_River(A)), not(z0_Sea(A)), l1w_the(A), not(z0_Mountains(A)), not(r1p_NN(A)), not(l3p_NN(A)), not(z0_State(A)), not(r2p_NNP(A)), not(l3s_LOC(A)), not(rw_an(A)).
desert(A) :- lw_the(A), not(z0_River(A)), not(z0_Sea(A)), not(rw_of(A)), l2p_IN(A), rs_,(A), not(rs_.(A)), not(r2p_NN(A)).
desert(A) :- lw_south(A), not(l3w_south(A)).
desert(A) :- l1w_western(A).
desert(A) :- rw_Israel(A).
desert(A) :- lw_villages(A).

Figure 2: First rules obtained with internal NE predicates, but only NEs having more than 1 token (Experiment 3).

5. Conclusions and Further Work

A system that uses ILP for grained classification of geographical NEs has been presented. The system has been applied for learning eleven binary classifiers corresponding to a set of subclasses of geographical NEs. The experimental set-up consisted of three different experiments that have been evaluated with 5-fold cross-validation. Although no direct evaluation on a test corpus has been performed, we can guess that small variance of F_1 measure resulting in most of the classes and experiments is a clear indicator that we can obtain similar results by training with the whole training set and testing with a test corpus.

Future work includes:

- Evaluating the system on a test corpus.
- Examining the possibility of applying different feature sets for each class and other types of NEs such as persons or organizations.
- Applying the same methodology for other languages.
- Studying different types of combination of the set of binary classifiers in order to generate a multiclass classifier.

6. Acknowledgments

This work has been partially supported by the European Comission (CHIL, IST-2004-506909) and the Spanish Research dept. (ALIADO, TIC2002-04447-C02). Our research group, TALP Research Center, is recognized as a Quality Research Group (2001 SGR 00254) by DURSI, the Research Department of the Catalan Government.

7. References

ADLP. Alexandria Digital Library Project, http://www.alexandria.ucsb.edu

Axelrod, A. 2003. On building a high performance gazetteer database. *Proceedings of HLT-NAACL Workshop of Geographic Reference*, Edmonton, Canada.

Brants, T. 2000. TnT, A Statistical Part-of-Speech Tagger, *Proceedings of the 6th ANLP-NAACL*, Seattle, USA.

Carreras, X., Márquez L., Padró, L. 2003. A Simple Named Entity Extractor Using AdaBoost. *Proceedings of CoNLL-2003*, Edmonton, Canada.

Leidner, J., Sinclair, G., Webber, B. 2003. Grounding spatial named entities for information extraction and question answering. *Proceedings of HLT-NAACL Workshop of Geographic Reference*, Edmonton, Canada.

Mann, G.S. 2002. Fine-Grained Proper Noun Ontologies for Question Answering. *Proceedings of SemaNet'02: Building and Using Semantic Networks*, Taipei, Taiwan.

Mann, G.S., Yarowsky, D. 2003. Unsupervised Personal Name Disambiguation. *Proceedings of the CoNLL*, Edmonton, Canada.

Manov, D., Kiryakov, A., Popov, B., Bontcheva, K., Maynard, D., Cunningham, H. 2003. Experiments with geographic knowledge for information extraction. *Proceedings of HLT-NAACL Workshop of Geographic Reference*, Edmonton, Canada.

Quinlan R. 1990. Learning Logical Definitions from Relations. *Machine Learning*, 5(3):239–266.

Sekine, S., Sudo, K., Nobata, C. 2002. Extended Named Entity Hierarchy. *Proceedings of Thirth International Conference on Language Resources and Evaluation (LREC-2002)*, Las Palmas, Spain.

Smith, D. A., Crane, G. 2001. Disambiguating geographic names in a historical digital library. *Proceedings of ECDL*, pages 127–136, Darmstadt, Germany.

van Rijsbergen, C. J. 1979. *Information Retrieval.* London: Butterworths.