University of Alicante at GeoCLEF 2005

O. Ferrández, Z. Kozareva, A. Toral, E. Noguera, A. Montoyo, R. Muñoz and Fernando Llopis
Grupo de Investigación en Procesamiento del Lenguaje y Sistemas de Información
Departamento de Lenguajes y Sistemas Informáticos
University of Alicante, Spain
{ofe,zkozareva,atoral,elisa,montoyo,rafael,llopis}@dlsi.ua.es

Abstract

For the participation of the University of Alicante in the first cross-language Geographic Information Retrieval, we have developed a system made up of three modules. One of them is an Information Retrieval module and the others are Named Entity Recognition modules based on machine learning and based on knowledge. We have carried out several runs with different combinations of these modules for resolving the monolingual and bilingual tasks. The system obtained better result in monolingual task achieving an improvement between 48% and 69% above the average. The results are shown and discussed in the paper.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.3 Information Search and Retrieval

General Terms

Algorithms, Experimentation, Measurement, Performance

Keywords

Information Retrieval, Geographic Information Retrieval, Named Entity Recognition

1 Introduction

The aim of GeoClef 2005 monolingual and bilingual tasks is to retrieve relevant documents from a monolingual collection. This documents are retrieved by using geographic tags like geographic places, geographic events and so on. Nowadays, the fast development of Geographic Information Systems (GIS) involve the need of Geographic Information Retrieval system (GIR) that help these system to obtain documents with relevant geographic information.

The cross-language Geographic Information Retrieval (GIR) system developed at the University of Alicante has been designed to retrieve relevant documents that contain geographic tags. For this reason, our system consist of several modules for the recognition of geographic entities also developed in the University of Alicante. We consider that an information retrieval module plus a named entity recognition (NER) modules will be better to identify relevant documents about specific geographic items.

This paper is organized as follows: next section describes the whole system and each module of the system in detail. Then, in the section of the results and discussion we describe the different runs carried out for the monolingual and bilingual tasks and we present the results obtained. Finally, the conclusions about our participation in GeoClef 2005 are expounded.

2 System description

Our Geographic Information Retrieval System is made up of three modules, which are detailed in the following subsections:

IR-n Information Retrieval module

NERUA Named Entity Recognition module based on machine learning

DRAMNERI Named Entity Recognition rule-based module. This module will allow, by means of rules, to obtain weak entities about geographic items. However, the rules are depending of the domain

For the resolution of the proposed tasks in GeoClef 2005, we have applied different combinations of these modules. These combinations are the different runs developed, and will be explained in the section 3.

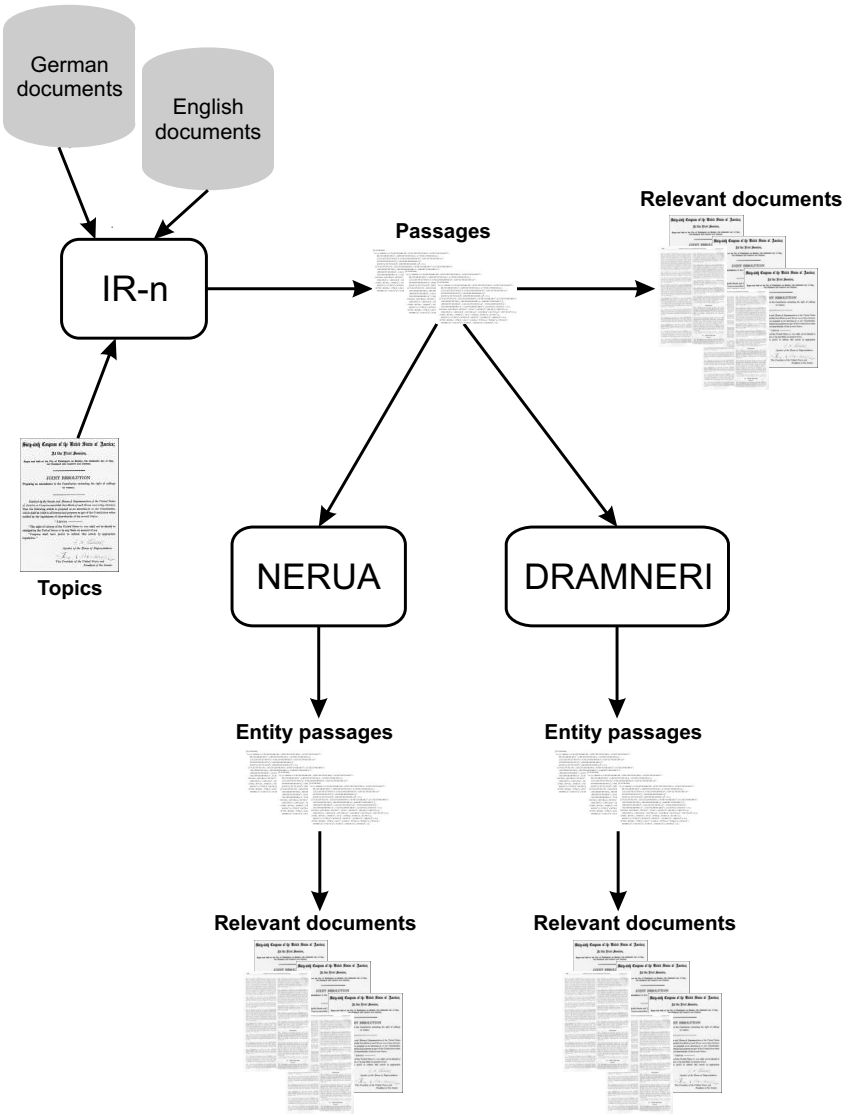

Figure 1: System architecture

An overview of our system is depicted in Figure 1. This shows also how the different modules interact among each other.

2.1 IR-n: Information Retrieval module

IR-n is a passage retrieval system (RP). RP systems [4] study the appearance of query terms in contiguous fragments of the documents (also called passages). One of the main advantages of these systems is that they allow us to determine not only if a document is relevant or not, but also the detection of the relevant part of the document.

The passages are usually composed for a fixed number of sentences. This number depends in a great measure of the collection used. To determinate this value, the system has been trained with the topics of year 2003 because the collections used in this task were used in the adhoc task CLEF 2003. The number of sentences that obtain the best results is 8 for both languages. Furthermore, IR-n system uses overlapping passages in order to avoid that some documents can be considered not relevant if words of the question appear in adjacent passages.

For every language, the resources used were provided by the organization of the clef[1]. These are stemmers and stopword lists (for English and German). Furthermore, we have used a splitter of compound nouns for German language.

IR-n system allows the use of distinct similarity measures, this involves an advantage, so that, in each task is used the best similarity measure. With this aim, it has been training the collections of the tasks which we have participated this year (English and German). For each collections the best similarity measure is Okapi [3].

According to others IR systems, IR-n system uses different techniques of the query expansion. Previous researches [1] have showed that the approaches get better results which are based on passages and in the complete document.

On the other hand, this year for the adhoc task has been implemented a technique called combined passages [6]. It applies fusion methods which are used in multilingual tasks to combine results with different size of passages.

2.2 NERUA: Named Entity Recognition module based on Machine Learning

NERUA [2] is a Named Entity Recognition system developed at the University of Alicante, build up of three diverse machine learning techniques: K-nearest neighbours, Maximum Entropy and Hidden Markov Models. The system consists of two passages, one for entity detection and another for entity classification of the already detected entities. Initially, the system was developed for Spanish language, using the train and test data sets of the CoNLL-2002. Compared to the systems participating in CoNLL-2002, our system reaches second place. For each one of the four categories, the achieved results are 78.46 % f-score for locations, 57.00% for miscellaneous entities such as names of sport events and movie titles, 78.93 % for organizations and 86.52% for person names.

The features behind the method are mainly lexical, contextual, gazetteers, trigger word lists, and morphological. However, the high performance of NERUA is due to the weighted voting strategy we incorporate during the classification task. Each classifier has weight depending on its performance for each one of the four categories. When two of the three or the three classifiers agree, the category of the entity is the one with the highest number of votes, when the classifiers disagree, the class from the classifier which weight is the highest is selected. Such weighted voting techniques are known to outperform the performance of a single classifier and improve the base model.

Once developed, NERUA was trained for Portuguese[2] and English languages. For English we used the CoNLL-2004 corpus provided for semantic role labelling competition. From this corpus, we considered only the words and the associated with them Named Entity tags. For English, the following set of characteristics has been used.

[1] http://www.unine.ch/info/clef

[2] http://poloxldb.linguateca.pt/harem.php

lexical

- **p**: position of w_0 (e.g. the word to be classified) in a sentence
- $wf[-3,+3]$: word forms of w_0 and the words in its window ± 3

orthographic

- **aC**: all letters of w_0 in capitals
- $iC[-3,+3]$: $w_{-3}, w_{-2}, w_{-1}, w_0, w_{+1}, w_{+2}, w_{+3}$ initiate in capitals

morphological information

- **aSubStr[1-5]**: ± 2, ± 3 and half substring of the word to be classified

Figure 2: The set of features used for NER

The advantage of NERUA is its ability to recognize entities using only the information coming from the corpus, rather than consulting and maintaining ample gazetteer lists. Since NERUA consists only of machine learning methods, its inconvenience is high computational cost and time performance.

2.3 DRAMNERI: Named Entity Recognition Rule-based module

DRAMNERI [7] (Dictionary, Rule-based and Multilingual Named Entity Recognition Implementation) is a system that identifies and classifies named entities. It is organized as a sequential set of modules.

One aim of this system is to be as customizable as possible. Thus, most of the actions it performs and the dictionaries and rules it uses are configurable by using parameter files. The main modules are briefly outlined in the following subsections.

2.3.1 Named Entity Identification

This task is applied on each sentence in the given text. Groups of tokens that match regular expressions jointed by prepositions are detected and identified as generic entities. The regular expressions and the maximum number of prepositions between matching tokens can be customized by the user.

For example, if we have 'of' and 'the' in the preposition list and the maximum number of prepositions between capitalised words is 1, then the string "in the University of Alicante" would be identified as "in the <ENTITY> University of Alicante </ENTITY>" but "Lilly of the Valley" would be identified as "<ENTITY> Lilly </ENTITY> of the <ENTITY> Valley </ENTITY>" instead of "<ENTITY> Lilly of the Valley </ENTITY>" because 1 is the maximum number of prepositions between capitalised words.

2.3.2 Named Entity Recognition

The goal of this phase is to assign a category to each of the entities detected in the previous step. For this to be accomplished, rules, dictionaries and triggers are used. The boundaries of the identified entities can be altered in this phase. This module is applied in two steps in a sequential manner:

Classification using triggers For trigger driven classification length-configurable left and right context of the identified entity are considered. Within these contexts front triggers and back triggers dictionaries are applied respectively. If any happens to be found then the entity is classified with the category of the dictionary that the matching trigger belongs to.

For example, if we have the string "Mr. <ENTITY>Smith</ENTITY>" and mr. is a person trigger, then Smith is classified as a person entity. The output string would be "<ENTITY type=PERSON>Mr. Smith</ENTITY>".

Classification using rules Dictionaries and rules are combined to perform entity classification. Rules follow the standard regular expression syntax and may contain elements that refer to dictionaries. Each rule is linked to an entity category. This way, if a rule matches a string of text then the category assigned is the one that is linked to the rule. An example follows:

rule: PER PREP PREP PER
entity: PER

This rule matches and entity that consists of a token which is in the Person dictionary (PER), followed by a token present in the preposition dictionary (PREP), etc. If a string of text matches then it is assigned the category PER. An example of string that would match is "Jorge de la Varga".

3 Results and discussion

In this section we present the results and analysis of our experimental runs. Our system has participated in the following tasks:

- Monolingual tasks:
  - English
  - German
- Bilingual tasks:
  - English-German
  - German-English
  - Spanish-English
  - Spanish-German
  - Portuguese-English
  - Portuguese-German

Several runs have been developed for each task. Each run is the result of combining the modules of our system, this combinations are explained in depth later.

3.1 Monolingual tasks

There are two mandatory runs for each task, the first of them uses only the topic title and the topic description, this run is called *Mand1*, whereas the other (*Mand2*) uses both the topic title and description plus all the geographic tags. Neither of them use the topic narrative. For carrying out these runs we have applied only the Information Retrieval module, which obtains the top 1000 ranked documents of the provided collections from the topics title, description and geographic tags.

In addition to these mandatory runs we have developed other runs using the NER modules. The first run uses the NER module based on Machine Learning (NERUA). The application of this module has been focused on the recognition of locations in the text. Even though NERUA is built up of three machine learning techniques, because of the large computing time required by these algorithms, we only have used the K-nearest neighbours technique. In a nutshell, this run combines the Information Retrieval module and NERUA module, in such a way that for each

passage that IR-n returns, NERUA will be considered as relevant depending on the existence of a location entity in the passage. This run is called *IRn+Nerua*.

In order to improve the recognition of location entities, we apply DRAMNERI. This is a Rule-based named entity recognition module. We have tailored the configuration files to only recognise location entities. Moreover specific gazetteers of locations, countries, geographic items and so on, have been incorporated to achieve better results. The DRAMNERI module takes the relevant passages returned from IR-n and analyses them to find specific locations entities, if any entity is found then the passage will be considered relevant, *IRn+Dramneri*.

The last run we have developed consists of an expansion of the topics adding synonyms of the main nouns. This run has only been carried out for English topics and we have used WordNet 1.5 in order to take the synonym words. We have denoted this run like *syn*.

Finally, all runs have been developed fully automatically and the results achieved are shown in TABLE 1.

Language	Run	AvgP	Dif.
English	CLEF Average	20.63	
	Mand1	32.53	
	Mand2	34.71	
	IRn+Nerua	34.95	+69.41%
	IRn+Dramneri	29.77	
	syn	33.28	
German	CLEF Average	8.28	
	Mand1	11.89	
	Mand2	12.27	+48.19%
	IRn+Nerua	12.14	
	IRn+Dramneri	12.02	

Table 1: *GeoClef 2005 officials results for Monolingual tasks*

The results achieved in the monolingual tasks are significantly different if the retrieved documents are from the English or German collections. All the runs that retrieved documents are from the English collections are considerable better than the others runs, the reason for this is that the different modules of our system were developed for the English language and, although we adjusted these modules for German, the result obtained haven't been as good as English results.

Moreover, we can observe that Nerua improves the result for English, but not for German language. This is due to the same reason; Nerua have been prepared for English and we adjusted Nerua for German, but the resources that Nerua needs for German Language were limited and inadequate.

The run with Dramneri doesn't obtain good results, we consider that Dramneri would need more resources or resources like specific gazetteers of locations, countries, geographic items and so on, more extensive.

3.2 Bilingual tasks

In order to resolve the bilingual task we have followed a similar strategy to the one used in [5]. This strategy consists of merging several translations built by several on-line translators. The motivation behind this idea is that the words that appear in different translations have more relevance that those that only appear in one translation. The translators used were: Freetranslation[3], Babel Fish[4] and InterTran[5]. Babel Fish and Freetranslation translators do not have direct translation from Spanish to German or from Portuguese to German, for this reason we have used English as the intermediate language for these translations.

[3]http://www.freetranslation.com
[4]http://world.altavista.com
[5]http://www.tranexp.com

Table 2 shows the scores achieved for bilingual tasks. For each couple of languages the same runs developed for monolingual tasks were carried out.

Language	Run	AvgP	Dif.
English-German	CLEF Average	10.28	
	Mand1	16.42	
	Mand2	15.67	
	IRn+Nerua	15.60	
	IRn+Dramneri	12.81	
	syn	17.52	+70.43%
German-English	CLEF Average	27.45	
	Mand1	30.83	
	Mand2	31.76	
	IRn+Nerua	31.78	+15.77%
	IRn+Dramneri	29.38	
Spanish-English	CLEF Average	27.45	
	Mand1	25.98	
	Mand2	25.97	
	IRn+Nerua	26.06	-5.06%
	IRn+Dramneri	23.65	
Spanish-German	CLEF Average	10.28	
	Mand1	9.61	-6.52%
	Mand2	9.51	
	IRn+Nerua	9.25	
	IRn+Dramneri	7.36	
Portuguese-English	CLEF Average	27.45	
	Mand1	26.09	
	Mand2	26.87	
	IRn+Nerua	26.91	
	IRn+Dramneri	27.00	-1.64%
Portuguese-German	CLEF Average	10.28	
	Mand1	8.71	
	Mand2	9.03	-12.16%
	IRn+Nerua	8.93	
	IRn+Dramneri	6.88	

Table 2: *GeoClef 2005 officials results for Bilingual tasks*

The results achieved for the bilingual runs achieved have a similar problem that the result for monolingual tasks; when the retrieved documents are from the English collections our system obtains better results than when the retrieved documents are from the German collections.

The Spanish-English and Portuguese-English scores are very similar, whereas the result achieved for German-English task are better. We have use the same translators for all runs, but maybe these translators work better with German than Spanish or Portuguese language. Finally, the result against the German documents collections have been quite low.

4 Conclusion

For our participation of the first cross-language Geographic Information Retrieval, we have developed a system made up of three modules: an Information Retrieval (IR) module and two modules of Named Entity Recognition (NER), one of them based on machine learning an the other based on knowledge. The NER modules are given the job of detection and classification of entities regarding locations, places and geographic items. For this reason, an appropriate combination of these modules could obtain the relevant documents with the specific location entities.

We have carried out several runs for each monolingual and bilingual tasks, each run combining the modules of our system in different ways. Our experiments achieve better results when the retrieved documents is from the English collections, this situation is due to the fact that our system has been prepared for English language and, even though we have adjusted the system for German, the lack of resources for this language makes that our system doesn't obtain good results for this language.

Regarding knowledge based NER, the problem we have encountered is that of absence of adequate resources. Hence, a structured knowledge resource with information about location and geographic items, which would have relationships of each item with its geographic location would be very useful.

Finally, as future work we intend to adjust our system in other languages. Regarding NER modules we have to obtain resources more adequate for each language, resources like annotated corpus for training, gazetteers of locations more complete, would improve these modules.

Acknowledgements

This research has been partially funded by the Spanish Government under project CICyT number TIC2003-07158-C04-01 and by the Valencia Government under project numbers GV04B-276 y GV04B-268

References

[1] Aitao Chen and Fredric C. Gey. Combining query translation and document translation in cross-language retrieval. In Carol Peters, Julio Gonzalo, Martin Braschler, and et al., editors, *4th Workshop of the Cross-Language Evaluation Forum, CLEF 2003*, Lecture notes in Computer Science, pages 108–121, Trondheim, Norway, 2003. Springer-Verlag.

[2] Óscar Ferrández, Zornitsa Kozareva, Andrés Montoyo, and Rafael Muñoz. Nerua: sistema de detección y clasificación de entidades utilizando aprendizaje automático. In *Accepted for Sociedad Española del Procesamiento del Lenguaje Natural*, volume 31, 2005.

[3] Savoy J. Fusion of probabilistic models for effective monolingual retrieval. In Carol Peters, Julio Gonzalo, Martin Braschler, and et al., editors, *4th Workshop of the Cross-Language Evaluation Forum, CLEF 2003*, Lecture notes in Computer Science, Trondheim, Norway, 2003. Springer-Verlag.

[4] M. Kaskziel and J. Zobel. Passage retrieval revisited. In *Proceedings of the 20th annual International ACM Philadelphia SIGIR*, pages 178–185, 1997.

[5] Fernando Llopis, Rafael Muñoz, Rafael M. Terol, and Elisa Noguera. Ir-n r2 : Using normailized passages. In *Lecture Notes in Computer Science*, volume 3491, 2004.

[6] Noguera E. Llopis F. Combining passages in monolingual experiments with ir-n system. In *Workshop of Cross-Language Evaluation Forum (CLEF 2005)*, In this volume, Vienna, Austria, 2005.

[7] Antonio Toral. Dramneri: a free knowledge based tool to named entity recognition. In *Proceedings of the 1st Free Software Technologies Conference*, pages 27–31. A Coruña, Spain, July 2005.