Identification of Composite Named Entities in a Spanish Textual Database*

Sofía N. Galicia-Haro[1], Alexander Gelbukh[2,3], and Igor A. Bolshakov[2]

[1] Faculty of Sciences
UNAM Ciudad Universitaria México, D. F.
sngh@fciencias.unam.mx
[2] Center for Computing Research
National Polytechnic Institute, Mexico City, Mexico
{gelbukh,igor}@cic.ipn.mx; www.Gelbukh.com
[3] Department of Computer Science and Engineering, Chung-Ang University,
221 Huksuk-Dong, DongJak-Ku, Seoul, 156-756, Korea

Abstract. Named entities (NE) mentioned in textual databases constitute an important part of their semantics. Lists of those NE are an important knowledge source for diverse tasks. We present a method for NE identification focused on composite proper names (names with coordinated constituents and names with several prepositional phrases.) We describe a method based on heterogeneous knowledge and simple resources, and the preliminary obtained results.

1 Introduction

Textual databases have been moved from desks to computers and also to the Web for many reasons: to save tons of paper, to allow people to have remote access, to provide much better access to texts in an electronic format, etc. Searching through this huge material for information of interest is a high time consuming task.

Named entities (NE) mentioned in textual databases constitute an important part of their semantics and of their lexical content. From a collection of political electronic texts we found that almost 50% of the total sentences contains at least one NE. This percentage shows the relevance of NE identification and its property to be used to index and retrieve documents. In [5] authors employed proper names for an automatic newspaper article classification. The quantity of proper names and their informative quality in such type of texts make them relevant to improve the clustering thanks to a measure of similarity that highlights them with regard to the other words in a text.

The research fulfilled in the Message Understanding Conference (MUC) [8] structure entity name task and it distinguishes three types: ENAMEX, TIMEX and NUMEX [4]. In this work we are concerned with ENAMEX that considers entities such as organizations, persons, and localities. Name entity recognition (NER) works in

* Work done under partial support of Mexican Government (CONACyT, SNI, COFAA-IPN), Korean Government (KIPA Professorship for Visiting Faculty Positions in Korea), and ITRI of CAU. The second author is currently on Sabbatical leave at Chung-Ang University.

F. Meziane and E. Métais (Eds.): NLDB 2004, LNCS 3136, pp. 395–400, 2004.

Language-Independent NER, the shared task of CoNLL-2002 [10] covered Spanish for name entity classification. However, composite names were limited.

In this paper, we are not concerned with classification but with identification of NE focusing our work on composite NE: names with coordinated constituents and names with several prepositional phrases. Since NE recognition is a difficult task our method is heterogeneous; it is based on local context, linguistic restrictions, heuristics and lists for disambiguation (two very small lists of proper names, one of similes, and a list of non ambiguous entities taken from the textual database itself). In this article, we present the text analysis carried out to determine the occurrence of NE, then we detailed our method and finally we present the obtained results.

2 Named Entities in Textual Databases

Textual databases could contain a great quantity of NE; most of them are unknown names. Since NE belong to open class of words, entities, as commercial companies are being created daily, unknown names are becoming important when the entities they referred to became topical or fashioned.

Our textual database is a collection of political Mexican texts that were compiled from the Web. They correspond to two different Mexican newspapers (1998 to 2001). We called them: collec#1 (442,719 total sentences, 243,165 with NE) and collec#2 (208,298 total sentences, 100,602 with NE). Although NE represent at most 10% of total words of our textual database they appear at least in 50% of the sentences.

Composite NE are common in Spanish texts, for example, the political texts of January 3rd 2000 contain among other NE the following composite names[1]:

6	Comandancia General del Ejército Zapatista de Liberación Nacional
6	Comité Clandestino Revolucionario Indígena
2	Comité Clandestino Revolucionario Indígena de la Comandancia General del Ejército Zapatista de Liberación Nacional
8	Ejército Zapatista de Liberación Nacional

Since we could observe that *Ejército Zapatista de Liberación Nacional, Comandancia General*, and *Comité Clandestino Revolucionario Indígena* are embedded in composite NE the elementary names are more than previous ones:

8	Comandancia General
7	Comité Clandestino Revolucionario Indígena
16	Ejército Zapatista de Liberación Nacional

Obtaining the real quantities of elementary NE should improve different tasks as texts classification.

Characteristics for Named Entities Identification

The initial step for NE recognition was identification of linguistic and style characteristics. We analyzed collec#1 and we found that NE are introduced or defined by means of syntactic-semantic characteristics and local context. The main characteristics observed were:

[1] Where "Ejército" means Army, "Comandancia General" means General Command, "Comité Clandestino Revolucionario Indígena" means Revolutionary Secret committee Native, and "Ejército Zapatista de Liberación Nacional" means Army Zapatista of National Liberation

Linguistic. NE could include: a) conjunctions: "y", "e" (*Centro de Investigaciones **y** Seguridad Nacional* is a single NE), b) prepositions (*Comisión **para** la Regularización **de** la Tenencia **de** la Tierra*). NE could be separated by a) punctuation marks (*Misantla*, *Chicontepec*, *Veracruz*, *Salina),* b) prepositions (*Salina Cruz **a** Juchitán).*

Style. It considers: a) information obtained from juxtaposition of NE and acronyms, for ex: *Partido de la Revolución Democrática* (*PRD*), b) introducing NE by specific words, for ex: *dirigentes de la Central Nacional de Estudiantes Democráticos* (leaders of the …), and c) diverse forms, for ex: *Centro de Investigación y Estudios Superiores de Antropología Social*, *Centro de Investigación y Estudios Superiores en Antropología Social*, correspond to the same entity, and more variety exists for NE translated from foreign languages.

3 Named Entities Analysis

We built a Perl program that extracts groups of words that we call "compounds"; they really are the contexts when NE could appear. The compounds contain no more than three non-capitalized words between capitalized words. We supposed that they should correspond to functional words (prepositions, articles, conjunctions, etc.) in composite NE. The compounds are left and right limited by a punctuation mark and a word if they exist. For example, for the sentence: *Esa unidad será dirigida por un Consejo Técnico que presidirá Madrazo Cuéllar y en el que participará el recién nombrado subprocurador Ramos Rivera.* (That unit will be directed by a Technical Advice whom Madrazo Cuéllar will preside over and in which the just named assistant attorney general Ramos Rivera will participate.) We obtained the following compounds:

- *por un Consejo Técnico que presidirá Madrazo Cuéllar y*
- *subprocurador Ramos Rivera.*

From 243,165 sentences 472,087 compounds were obtained from collec#1. We analyzed 500 sentences randomly selected and we encountered the main problems that our method should cope with. They are described in the following sections.

Syntactic Ambiguity

- *Coordination*. Singular conjunction cases ("word conjunction word") cover most of coordinated NE. For ex. *Hacienda y Crédito Público* is a single NE. However, there are cases where the coordinated pair is a sub-structure of the entire name, for ex: *Mesa de* [*Cultura y Derechos*] *Indígenas* (Meeting of Culture and Right Natives). Coordination of coordinated NE introduces ambiguity to determine single NE, for example, the group *Comercio y Fomento Industrial y Hacienda y Crédito Público* contains two organization NE where the second one is underlined.
- *Prepositional phrase attachment* (PPA) is a difficult task in syntactic analysis. NE identification presents a similar problem. A specific grammar for NE is not a solution since it should cope with the already known PPA. We consider a diverse criterion than that considered in CoNLL: in case a named entity is embedded in another name entity or in case a named entity is composed of several entities all the components should be determined. For example: *Centro de Investigación y Estudios Avanzados del Politécnico* (Polytechnic's Center of …) where *Centro de*

Investigación y Estudios Avanzados is a research center of a superior entity (Polytechnic).

Discourse Structures

Discourse structures could be another source for knowledge acquisition. Entities could be extracted from the analysis of particular sequences of texts. We consider:

- *Enumeration* can be easily localized by the presence of similar entities, separated by connectors (commas, subordinating conjunction, etc). For example, *Cocotitlán, Tenango del Aire, Temamatla, Tlalmanalco, Ecatzingo y Atlautla*
- *Emphasizing* words or phrases by means of quotation marks. For ex: "*Roberto Madrazo es el Cuello*", "*Gusano Gracias*", are parodies of well known names.

4 Method

We conclude on our analysis that a method to identify NE in the textual database should be based mainly on the typical structure of Spanish NE themselves, on their syntactic-semantic context, on discourse factors and on knowledge of specific composite NE. Then, our method consists of heterogeneous knowledge contributions.

We avoid complex methods and big resources. For example, in [9] three modules were used for name recognition: List lookup for names and cues, POS tagger and Name parsing, and Name-matching (against all unidentified sequences of proper nouns). Other systems use lists of names of very different sizes, from 110,000 names (MUC-7) to 25,000-9,000 names [6]. [7] experimented with different types of lists.

The lists of names used by NER systems have not generally been derived directly from text but have been gathered from a variety of sources. For example, [2] used several name lists gathered from web sites. We also included lists from Internet and a hand made list of similes [1] (stable coordinated pairs) for example: *comentarios y sugerencias*, *noche y día*, *tarde o temprano*, (comments and suggestions, night and day, late or early). This list of similes was introduced to disambiguate some coordinated groups of capitalized words. The lists obtained from Internet were: 1) a list of personal names (697 items), 2) a list of the main Mexican cities (910 items) included in the list of telephone codes.

Linguistic knowledge. It considers preposition use, POS of words linking groups of capitalized words, and punctuation rules. The linguistic knowledge is settled in linguistic restrictions. For example:

1. Lists of groups of capitalized words are similar entities. Then an unknown name of such lists has similar category and the last one should be a different entity coordinated by conjunction. For example: *Santo Domingo*, *Granada*, *Guatemala y Haití.*
2. Preposition *"por"* followed by an undetermined article cannot link groups of person names. The compound: *Cuauhtémoc Cárdenas por la Alianza por la Ciudad de México* must be divided in *Cuauhtémoc Cárdenas* and *Alianza por la Ciudad de México.* Therefore, the last compound could correspond to a single name.

Table 1. Results in a testing set of sentences

	NUMBER OF:		
	COORDINATED GROUPS	**PREPOSITIONAL. PHRASE GROUPS**	**ALL**
Precision	54	69	89
Recall	48	67	87

Heuristics. Some heuristics were considered to separate compounds. For example:

1. Two capitalized words belonging to different lists must be separated. For example: "*...en Chetumal Mario Rendón dijo ...*", where *Chetumal* is an item of main cities list and *Mario* is an item of personal names list.
2. One personal name should not be coordinated in a single NE. For ex: *José Ortega y Gasset y Manuel García Morente*, where *Manuel* is an item of personal name list.

Statistics. From collec#1 we obtained the statistics of groups of capitalized words, from one single word to three contiguous words, and groups of capitalized words related to acronyms. The top statistics were used to disambiguate compounds joined by

- *Functional words*. For ex. the compound *Estados Unidos sobre México* could be separated in *Estados Unidos* (2-word with high score) and *México.*
- *Embedded names*. For example: *Consejo General del Instituto Federal Electoral* could be separated in: *Consejo General* and *Instituto Federal Electoral*.

Application of the Method

Perl programs were built for the following steps to delimit NE:

First step: All composite capital words with functional words are grouped in one compound. We use a dictionary with part of speech to detect functional words.

Second step: Using the resources (statistics of collec#1 and lists), rules and heuristics above described the program decides on splitting, delimiting or leaving as is each compound. The process is 1) look up the compound in the acronym list, 2) decide first on coordinated groups, then on prepositional phrases, and finally decide on the rest of groups of capitalized words.

5 Results

Since collec#1 was used for training we use 500 sentences randomly selected of collec#2 to test the method. They were manually annotated and compared. The composite NE were split and each individual NE was annotated. The correct entities detected should have the same individual NE. The results are showed in Table 1 where:

Precision: # of correct entities detected / # of entities detected

Recall: # of correct entities detected / # of entities manually labeled (*eml*)

The table indicates the performance for coordinated names (63 *eml*), prepositional groups[2] (167 *eml*). The last column shows the overall performance (1496 *eml*) including the previous ones. The main causes of errors are: 1) foreign words, 2) personal names missing in the available list, and 3) names of cities.

[2] Where all prepositional phrases related to acronyms were not considered in this results.

The overall results obtained by [3] in Spanish texts for name entity recognition were 92.45% for precision and 90.88% for recall. But test file only includes one coordinated name and in case a named entity is embedded in another name entity only the top level entity was marked. In our work the last case was marked incorrect. The worst result was that of NE with coordinated words that should require enlargement of current sources. The 40% of coordinated correct entities detection was based on the list of similes that could be manually enlarged.

Conclusions

In this work, we present a method to identify and disambiguate groups of capitalized words. We are interested in minimum use of complex tools. Therefore, our method uses extremely small lists and a dictionary with POS. Since limited resources use cause robust and velocity of execution.

Our work is focused on composite NE (names with coordinated constituents and names with several prepositional phrases) to obtain elementary NE that are useful for different tasks like texts classification. The strategy of our method is the use of heterogeneous knowledge to decide on splitting or joining groups with capitalized words. The results were obtained on 500 sentences that correspond to different topics. The preliminary results show the possibilities of the method and the required information for better results.

References

1. Bolshakov, I. A., A. F. Gelbukh, and S. N. Galicia-Haro: Stable Coordinated Pairs in Text Processing. In Václav Matoušek and Pavel Mautner (Eds.). Text, Speech and Dialogue. Lecture Notes in Artificial Intelligence, N 2807, Springer-Verlag (2003) 27–35
2. Borthwick et al. Exploiting Diverse Knowledge Sources via Maximum Entropy in Named Entity Recognition Proceedings of the Sixth Workshop on Very Large Corpora (1998)
3. Carreras, X., L. Márques and L. Padró. Named Entity Extraction using AdaBoost In: Proceedings of CoNLL-2002, Taipei, Taiwan (2002) 167-170
4. Chinchor N.: MUC-7 Named Entity Task Definition (version 3.5). http://www.itl.nist.gov/iaui/894.02/related projects/muc/proceedings/muc 7 toc.html# appendices (1997)
5. Friburger, N. and D. Maurel.: Textual Similarity Based on Proper Names. Mathematical Formal Information Retrieval (MFIR'2002) 155–167
6. Krupka, G. and Kevin Hausman. Description of the NetOwl(TM) extractor system as used for MUC-7. In Sixth Message Understanding Conference MUC-7 (1998)
7. Mikheev A., Moens M., Grover C.: Named Entity Recognition without Gazetteers. In Proceedings of the EACL (1999)
8. MUC: Proceedings of the Sixth Message Understanding Conference. (MUC-6). Morgan Kaufmann (1995)
9. Stevenson, M. & Gaizauskas R.: Using Corpus-derived Name List for name Entity Recognition In: Proc. of ANLP, Seattle (2000) 290-295
10. Tjong Kim Sang, E. F.: Introduction to the CoNLL-2002 Shared Task: Language-Independent Named Entity Recognition. In: Proceedings of CoNLL-2002, Taipei, Taiwan (2002) 155-158