Recognition of Named Entities in Spanish Texts*

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

[1] Faculty of Sciences
UNAM Ciudad Universitaria Mexico City, Mexico
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. Proper name recognition is a subtask of Name Entity Recognition in Message Understanding Conference. For our corpus annotation proper name recognition is a crucial task since proper names appear approximately in more than 50% of total sentences of the electronic texts that we collected for such purpose. Our work is focused on composite proper names (names with coordinated constituents, names with several prepositional phrases, and names of songs, books, movies, etc.) We describe a method based on heterogeneous knowledge and simple resources, and the preliminary obtained results.

1 Introduction

A big corpus is being compiled by our research group. Since we defined its size in tenths of million words and its objective as unrestricted text analysis, the easiest and quickest manner to obtain texts was extracting electronic texts from Internet. We selected four Mexican newspapers daily published in the Web with a high proportion of their paper publication. We found that almost 50% of the total unknown words were proper names. This percentage shows the relevance of proper name recognition and it justifies a more wide analysis.

Proper names have been studied in the field of Information Extraction [15] for diverse uses. For example [5] employed proper names for an automatic newspaper article classification. Information Extraction requires the robust handle of proper names for successful performance in diverse tasks as pattern filling with correct entities that perform semantic roles [11]. The research fulfilled in the Message Understanding Conference (MUC) structure entity name task and it distinguishes three types: ENAMEX, TIMEX and NUMEX [4]. ENAMEX considers entities such as organizations (corporations names, government entities, and other type of organizations), persons (persons names, last names), and localities (localities names defined politically or geographically: cities, provinces, countries, mountains, etc.).

* 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.

R. Monroy et al. (Eds.): MICAI 2004, LNAI 2972, pp. 420–429, 2004.

In this paper, we are concerned with ENAMEX entity recognition but we focused our work on composite named entities: names with coordinated constituents, names with several prepositional phrases, and names of songs, books, movies, etc. We postponed name classification to future work.

NER works in MUC have been dedicated to English language, they considered complex tools or huge resources. For example, in [12] three modules were used for name recognition: List Lookup (consulting lists of likely names and name cues), Part of speech tagger, Name parsing (using a collection of specialized name entity grammars), and Name-matching (the names identified in the text are compared against all unidentified sequences of proper nouns produced by the part of speech tagger). The system of [14] recognized named entities by matching the input against pre-stored lists of named entities, other systems use gazetteers (lists of names, organizations, locations and other name entities) of very different sizes, from 110,000 names (MUC-7) to 25,000-9,000 names [6].

NER works in Language-Independent Named Entity Recognition, the shared task of Computational Natural Language Learning (CoNLL) covered Spanish in 2002 [13]. A wide variety of machine learning techniques were used and good results were obtained for name entity classification. However composite names were limited: named entities are non-recursive and non-overlapping, in case a named entity is embedded in another name entity only the top level entity was marked, and only one coordinated name appears in the training file.

Since named entities recognition is a difficult task our method is heterogeneous; it is based on local context, linguistic restrictions, statistical heuristics and the use of lists for disambiguation (very small external lists of proper names, one of similes and lists of non ambiguous entities taken from the corpus itself). In this article, we present the text analysis carried out to determine the occurrence of named entities, then we detailed our method and finally we present the obtained results.

2 Named Entities in Newspaper Texts

Two aspects should be considered in named entities recognition: known names recognizing and new names discovering. However, newspaper texts contain a great quantity of named entities; most of them are unknown names. Since named entities 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.

We analyzed Mexican newspaper texts that were compiled from the Web. They correspond to four different Mexican newspaper, between 1998 and 2001. From the analysis, we concluded that almost 50% of total words were unknown words[1]. We found 168,333 different words that were candidates to be named entities since they were initialized or totally fulfilled with capital letters. These capitalized words represent a low percentage from all different words but they appear at least in 50% of the sentences. We present such statistics in Table 1. From those numbers we note the importance of named entities for syntactic analysis of unrestricted texts since 50% to 60% of total sentences include named entities.

[1] They were not recognized by our resources: a dictionary with POS and a spelling checker.

Table 1. Statistics of newspaper texts

	Newspapers			
	# 1	**#2**	**#3**	**#4**
# Words	87,597,168	38,387,767	5,652,358	45,702,200
# Sentences	2,927,723	1,328,157	208,298	1,696,358
# Sentences w/ named entities	1,581,225	729,496	100,602	1,007,051

The initial step, for recognition of named entities was identification of context and style. We selected one Mexican newspaper, since we supposed that all newspapers present the named entities in similar manner. We analyzed newspaper #2 and we found that named entities are introduced or defined by means of syntactic-semantic characteristics and local context. The main characteristics observed were:

Conventions. Specific words could introduce names, for example: *coordinadora del programa Mundo Maya* (Mundo Maya program's coordinator), *subsecretario de Operación Energética de la Secretaría de Energía* (sub secretary of…), etc.

Redundancy. Information obtained from juxtaposition of named entities and acronyms, for example: *Asociación Rural de Interés Colectivo* **(***ARIC***)**, two names linked for the same entity by means of specific words: *alias*, (a), for example: ... *dinero de Amado Carrillo Fuentes* ***alias*** *El Señor de los Cielos*...

Prepositions usage. We consider two cases:

1. Prepositions link two different named entities. For example "a" indicates direction (*Salina Cruz* ***a*** *Juchitán*); "en" indicates a specific location (*Tratado sobre Armas Convencionales* ***en*** *Europa*), etc.
2. Prepositions are included in the named entities (*Instituto* ***para*** *la Protección* ***a****l Ahorro Bancario, Monumento* ***a*** *la Independencia, Centro de Investigaciones y Estudios Superiores* ***en*** *Antropología Social*).

Local context. Named entities are surrounded by local context signs. They could be used for identification of book, song and movie names. For example: *Marx escribió La ideología alemana* (Marx wrote...); ...*titulado La celebración de muertos en México* (titled...), etc. Some verbs (read, write, sing, etc.), some nouns (book, song, thesis, etc.), or proper names of authors often introduce or delimit such kind of names as those underlined.

Sets of names. Named entities could appear as sets of capitalized words. Punctuation (,;) is used to separate them, for example: *Bolivia, Brasil, Uruguay, Ecuador, Panamá*, or *... de Charles de Gaulle y Gagarin; de Juan Pablo II; de Eva Perón...*

Flexibility. Long named entities do not appear as fixed forms. For ex.: *Instituto para la Protección al Ahorro*, *Instituto para la Protección al Ahorro Bancario*, *Instituto para la Protección del Ahorro Bancario*, all of them correspond to the same entity. More variety exists for those names translated from foreign languages to Spanish.

Coordinated names. Some named entities include conjunctions (y, e, etc.) For example*: Luz* ***y*** *Fuerza del Centro, Margarita Diéguez* ***y*** *Armas, Ley de Armas de Fuego* ***y*** *Explosivos, Instituto Nacional de Antropología* ***e*** *Historia.*

Concept names. Some capitalized words represent abstract entities. In a strict sense they could not be considered as named entities and they should be tagged with a different semantic tag. For example: *Las violaciones a la Ley en que algunos …* (The violations to the Law in which some ...). Such kind of entities should be differentiate from those names representing an abbreviation of longer names (for example: *Ley del Seguro Social*) in a deep understanding level.

3 Named Entities Analysis

In order to analyze how named entities could be recognized by means of linguistics and context rules or heuristics we separate newspaper #2 sentences in two groups:

1. sentences with only one initial capitalized word, and
2. sentences with more than one capitalized word.

Group 1 could not contain named entities since the first word in each sentence could be name entity or one common word. [8] proposed an approach to disambiguate capitalized words when they appear in the positions where capitalization is expected. Their method utilize information of entire documents to dynamically infer the disambiguation clues. Since we have a big collection of texts we could apply the same idea.

We concentrated our analysis in group 2. We built a Perl program that extracts groups of words that we call "compounds", they really are the contexts when named entities 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 named entities (coordinated names and names with several prepositional phrases). The compounds are left and right limited by a punctuation mark and a word if they exist.

For example, for the following sentence:

Un informe oficial aseguró que Cuba invierte anualmente cerca de 100 millones de dólares en tecnologías informáticas y que en el trabajo para enfrentar al error del milenio, el país participó intensamente en el Grupo Regional de México, Centroamérica y el Caribe, con apoyo del Centro de Cooperación Internacional Y2K que funciona en Washington y que fue creado por la Organización de Naciones Unidas.

We obtained the following compounds:

- *que Cuba invierte*
- *el Grupo Regional de México, Centroamérica y el Caribe,*
- *del Centro de Cooperación Internacional Y2K que funciona en Washington y*
- *la Organización de Naciones Unidas.*

From 723,589 sentences of newspaper #2, 1348,387 compounds were obtained. We analyzed randomly approximately 500 sentences and we encountered the main problems that our method should cope with. They are described in the following sections.

3.1 Paragraph Splitting

We observed two problems in paragraph splitting: 1) sentences that should be separated and 2) sentences wrong separated. The causes of such errors were:

1. Punctuation marks. Sentences ending with quotation marks and leaders. For ex.:
   "___ película personal." A pesar de ___
   It is a competence error since in Spanish the point appears before quotation marks when the whole sentence is wanted to be marked.
2. Abbreviations. For ex., in the following phrase "*Arq*." corresponds to "architect":
   ___ ante las cámaras de televisión, el Arq. Héctor E. Herrera León ___
   [9] consider several methods for determining English abbreviations in annotated corpus: combinations of guessing heuristics, lexical lookup and the document-centered approach. We only consider the first method to automatically obtain a list of abbreviations from newspaper #2. They were obtained with heuristics such as: abbreviations have length less than five characters, they appear after a capitalized word between commas, etc. They mainly correspond to professions and Mexican states.
3. Style. Some sentences show an unclear style. For example, the use of parenthesis
   ___ de nadie. (Y aunque muchos sabemos que los asaltos están también a la orden del día, precisamente en el día.) No hace mucho ___
   The traditional Spanish use of parenthesis is the isolation of a small sentence part.

3.2 Syntactic Ambiguity

We found three main syntactic ambiguities in compounds, introduced by coordination, prepositional phrase attachment, and named entities composed of several words where only the first word is capitalized.

The last one corresponds to titles of songs, movies, books, etc. For example: *Ya en El perro andaluz, su primer filme ...* (Already in The Andalusia dog, his first movie.) The titles appearing in the electronic texts begin with one capitalized word followed by several non capitalized words, and sometimes another name entity embedded. As far as we observed, there are no use of punctuation marks to defined them. This use is different to that considered in the CoNLL-2002 training file, where the included titles are delimited by quotation marks.

Recognition of named entities related to coordination and prepositional phrase attachment is crucial for our objective: unrestricted text analysis. For all singular conjunction cases, dependency grammars assign the following structure to coordinated structures in the surface level: (→) P1 → C → P2, where P1 is the sub tree root. In the simpler and more usual case, the components P1 and P2 with the conjunction cover named entities. For example *Luz y Fuerza.* However, there are other cases where the coordinated pair is a sub-structure of the entire name, for example: *Mesa de* [*Cultura y Derechos*] *Indígenas*.

The following compounds shows the ambiguity introduced by coordination (the second component is underlined):

- *Comisión Federal de Electricidad y Luz y Fuerza del Centro*, includes two organization names.
- *Margarita Diéguez y Armas y Carlos Virgilio*, includes two personal names.
- *Comunicaciones y Transportes y Hacienda* includes two organization names.

Some compound examples of single names containing coordinated words:

- *Comisión Nacional Bancaria y de Valores*
- *Subsecretario de Planeación y Finanzas*
- *Teatro y Danza de la UNAM*

Prepositional phrase attachment is a difficult task in syntactic analysis. Named entities present similar problem. 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 since syntactic analysis should find their relations for deep understanding in higher levels of analysis. For example:

1. *Teatro y Danza de la UNAM* (UNAM's Theater and Dance)

Teatro y Danza is a cultural department of a superior entity.

2. *Comandancia General del Ejército Zapatista de Liberación Nacional*

Comandancia General (General command) is the command of an entity (army).

A specific grammar for named entities should cope with the already known prepositional phrase attachment problem. Therefore diverse knowledge described in the following section must be included to decide the splitting or joining of named entities with prepositional phrases.

3.3 Discourse Structures

Discourse structures could be another source for knowledge acquisition. Entities could be extracted from the analysis of particular sequences of texts. We are particularly interested in

1. Enumeration that can be easily localized by the presence of similar entities, separated by connectors (commas, subordinating conjunction, etc). For example, in the following sequence:
   La Paz, Santa Cruz y Cochabamba
   José Arellano, Marco A. Meda, Ana Palencia García y Viola Delgado
2. Emphasizing words or phrases by means of quotation marks and capitalized words.
   For example: "*Emilio Chichifet*", "*Roberto Madrazo es el Cuello*", "*Gusano Gracias*", are parodies of well known names and the sentence author denote it by quotation marks.
3. Author's intension. A specific intension could be denoted by capitalized words since author chose the relation in the structure. For example, "Convent" in :
   ___*una visita al antiguo Convento de la Encarnación, ubicado en*___ ()
   ___*y así surgió el convento de Nuestra Señora de Balvanera*. ()

   The first one shows the author's intension to denote the whole name of the building covering its old purpose. The author's intension in the second one is to make evident to whom is dedicated the convent.

4 Method

We conclude on our analysis that a method to identify named entities in our electronic texts collection should be based mainly on the typical structure of Spanish named entities themselves, on their syntactic-semantic context, on discourse factors and on

knowledge of specific composite named entities. Then, our method consists of heterogeneous knowledge contributions.

Local context. Local context has been considered in different tasks. For example, [7] used it for semantic attribute identification in new names. We consider local context to identify names of songs, books, movies, etc. For such purpose a window of two words preceding the capitalized word was defined. In such window a word appearing in a manually compiled list of 26 items plus synonyms and variants (feminine, masculine, plural) was considered a cue that introduce a name of song, book, etc. For example:

- *En su libro La razón de mi vida* (*Editorial Pax*)... (In his book *The reason of my life* (Pax publisher)
- *...comencé a releer La edad de la discreción de Simone de Beauvoir*,… (I began to reread Simone de Beauvoir's *The age of discretion*,)
- *...en su programa Una ciudad para todos que*... (in his program *A city for all* that)

Some heuristics were determined to obtain the complete name: all posterior words are linked until a specific word or punctuation sign is found. Such word or punctuation sign could be: 1) a name entity, 2) any sign of punctuation in texts (period, comma, semicolon, etc.) and 3) a conjunction.

In the above examples the signs: "(", "*Simone de Beauvoir*" and the conjunction "que" delimit the names. For more complex cases, statistics are included.

The phrases delimited by quotation marks preceded by a cue were also considered as names of songs, books, movies, etc.

Linguistic knowledge. We mainly consider the preposition use, part of speech 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 have similar category and the last one should be a different entity coordinated by conjunction. For example: *Corea del Sur, Taiwan, Checoslovaquia y Sudáfrica*.
2. Preposition use, considering the meaning of prepositions for localization, direction, etc. For example:

Preposition *"por"* followed by a undetermined article cannot link groups of person names. For example the compound: *Juan Ramón de la Fuente por la Federación de Colegios de Personal Académico* must be divided in *Juan Ramón de la Fuente* and *Federación de Colegios de Personal Académico.* Therefore, the compound *Alianza por la Ciudad de México* could correspond to a single name.

Two named entities joined by preposition "a" should be separated if they are preceded by preposition indicating an origin position ("de", "desde"). For example: *de Oaxaca a Salina Cruz.*

Heuristics. Some heuristics were considered to separate compounds.

1. Two capitalized words belonging to different list 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 name entity. For example: *Margarita Diéguez y Armas y Carlos Virgilio*, where Carlos is an item of personal name list.

3. A group of capitalized words with functional words followed by an acronym should be defined a single name if most of initial letters are in the acronym. For example: *FIFA Federación Internacional de la Asociación de Fútbol*; *OAA Administración Americana para la Vejez.*
4. All capitalized words grouped by quotation marks, without punctuation marks, are considered one name entity. For example: "*Adolfo López Mateos*".

Statistics. From newspaper #2 we obtained the statistics of groups of capitalized words, one single word to three contiguous words, and groups of capitalized words related to acronyms. The top statistics for such groups were used to disambiguate compounds joined by

- Functional words. For example, the compound *Estados Unidos sobre México* could be separated in *Estados Unidos* (a 2-word group with high score) and *México* (a 1-word with high score). In the same manner the compound *BP Amoco Plc* is kept as is and *ACNUR Kris Janowski* is separated in *ACNUR* and *Kris Janowski.*
- Prepositions. For example: *Comandancia General del Ejército Zapatista de Liberación Nacional* could be separated in : *Comandancia General* and *Ejército Zapatista de Liberación Nacional*.

Many NER systems use lists of names, for example [6] made extensive use of name lists in their system. They found that reducing their size by more than 90% had little effect on performance, conversely adding just 42 entries led to improved results. [10] experimented with different types of lists in an NER system entered for MUC7. They concluded that small lists of carefully selected names are as effective as more complete lists.

The lists of names used by named entity 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 containing lists of people first names, companies and locations. 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 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) considered in the list of telephone codes.

Application of the method. Perl programs were built for the following steps that have been taken for delimiting named entities:

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 previous resources (statistics of newspaper #2 and lists) and the 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 on coordinated groups using the list of similes, rules (based on enumeration and statistics), 3) decide on prepositional phrases using rules, heuristics and statistics, 4) delimit possible titles using context cues and rules, and 5) decide on the rest of groups of capitalized words using heuristics and statistics.

Table 2. Results in a testing set of sentences

	NUMBER OF:			
	COORD. GROUPS	**PREP. PHRASE GROUPS**	**TITLES**	**ALL**
Precision	56	70	55	90
Recall	49	67	32	88

5 Results

We test the results of our method in 400 sentences of newspaper#4. They were manually annotated and compared against the results obtained with our method. The results are presented in Table 2 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 (55 eml), prepositional groups[2] (137 eml) and titles (19 eml). The last column shows the overall performance (1279 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.

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 title recognition since 60% of them were not introduced by a cue. Recognition of titles and named entities with coordinated words 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.

6 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 use extremely small lists and a dictionary with part of speech. Since limited resources use cause robust and velocity of execution, important characteristics for processing huge quantity of texts.

Our work is focused on composite named entities (names with coordinated constituents, names with several prepositional phrases, and names of songs, books, movies, etc.) The strategy of our method is the use of heterogeneous knowledge to decide on splitting or joining groups with capitalized words. We confirmed that conventions are very similar in different newspapers then heuristics are applicable in the four newspapers selected.

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

The results were obtained from 400 sentences that correspond to different topics. The preliminary results shows 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, pp. 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. Mani I., McMillian R., Luperfoy S., Lusher E. & Laskowski S.: Identifying unknown proper names in newswire text. In Pustejovsky J. & Boguraev B. (eds.) Corpus processing for lexical acquisition. MIT Press, Cambridge, MA. (1996)
8. Mikheev A.: A Knowledge-free Method for Capitalized Word Disambiguation. In Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (1999) 159–166
9. Mikheev A.: Periods, Capitalized Words, etc. Computational Linguistics Vol. 28-3 (2002) 289–318
10. Mikheev A., Moens M., Grover C.: Named Entity Recognition without Gazetteers. In Proceedings of the EACL (1999)
11. MUC: Proceedings of the Sixth Message Understanding Conference. (MUC-6). Morgan Kaufmann (1995)
12. Stevenson, M. & Gaizauskas R.: Using Corpus-derived Name List for name Entity Recognition In: Proc. of ANLP, Seattle (2000) 290-295
13. 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
14. Wakao, T., R. Gaizauskas & Y. Wilks.: *Evaluation of an Algorithm for the Recognition and Classification of Proper Names*. In Proceedings of the 16th International Conference on Computational Linguistics (COLING96), Copenhagen (1996) 418–423
15. Wilks Y. Information Extraction as a core language technology. In M. T. Pazienza (ed.), Information Extraction, Springer-Verlag, Berlin (1997)