Web-Based Sources for an Annotated Corpus Building and Composite Proper Name Identification

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. Nowadays, collections of texts with annotations on several levels are useful resources. Huge efforts are required to develop this resource for languages like Spanish. In this work, we present the initial step, lexical level annotation, for the compilation of an annotated Mexican corpus using Web-based sources. We also describe a method based on heterogeneous knowledge and simple Web-based sources for the proper name identification required in such annotation. We focused our work on composite entities (names with coordinated constituents, names with several prepositional phrases, and names of songs, books, movies, etc.). The preliminary obtained results are presented.

1. Introduction

Obtaining usage information of language from collections of texts, i.e. corpus, has been a common practice in lexicography. In natural language processing, the use of very large corpora is also a common practice for linguistic phenomena research that mainly appears in unrestricted materials. Researches recognize the potential of very large corpus usage for problem solving in the lexical, syntactic and semantic levels of analysis. However, the useful corpora required in natural language processing need annotations, i.e., lexical, syntactic and semantic marks.

Training methods to resolve distinct natural language processing tasks have been the main usage of annotated corpus. Manual work is the most employed method and the more accurate for corpus compilation and annotation. To reduce economic and manual efforts we intend to make by automatic processes most of the work to compile and annotate such type of corpus.

There is a wide range of possible usages of the WEB for corpus construction, from a specific retrieval of contexts [7] to a whole use, since it has been even proposed to use the WEB itself as a very huge corpus [8]. We use the Web as the main source of appropriate Mexican texts for our purposes. We selected four Mexican newspapers daily published in the Web with a high proportion of their paper publication.

J. Favela et al. (Eds.): AWIC 2004, LNAI 3034, pp. 115-124, 2004.

After text processing of Web-based sources for adequate format assignment we found that almost 50% of the total unknown words were proper names. Proper names have been mainly studied in the field of Information Extraction that 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 [5]. In this paper, we are concerned with ENAMEX entity (organizations, persons and localities) identification but we focused our work on names with coordinated constituents, names with several prepositional phrases, and titles.

Name entity recognition (NER) works in MUC have been dedicated to English. NER works in Language-Independent NER, the shared task of Computational Natural Language Learning (CoNLL) covered Spanish in 2002 [14]. A wide variety of machine learning techniques was used with good results for name entity (NE) classification. However composite names were limited: NE are non-recursive and non-overlapping, in case a NE is embedded in another one only the top level entity was marked, and test files do not include proper names with coordinated constituents.

In this work, we present some observations and results about text collection compilation, then we describe the development of the lexical annotation for our corpus. Then we present the heterogeneous method for proper name identification, it is based on local context, rules, statistics, heuristics and the use of Web-based lists. We present the definition of composite proper names, the method to identify them and finally the obtained results.

2. Characteristics of Annotated Corpus

Corpus compilation implies different problem solutions of the self texts. Ideally, it is desirable to obtain a large and representative sample of general language, it could be expected a larger quantity of words as longer is the corpus. A big quantity of words should imply bigger dictionary language coverage and it mainly should imply greater evidence of the diverse linguistic phenomena required. To be representative supposes several cultural language levels, several themes and genres. However these qualities do not imply each other, in some cases instead they are contrary. One contraposition that must be considered is that between quality and quantity. A big corpus does not guarantee to posses the expected quality.

The corpus should be balanced among those qualities. However, it seems to be not possible to balance appropriately a corpus, not without huge effort. In addition the sampling methods are quite expensive, for example those for quality selection. To reduce time and costs we must assume the obvious problems related to work with unbalanced data.

Because of the huge efforts to compile a corpus with all desired qualities, we limited the corpus qualities to those that are relevant for our goals: required information and size. The main goal for the Mexican corpus compilation is the syntactic analysis of unrestricted texts, similar to those of newspaper texts. One of the main problems in the syntactic analysis is the correct attachment of noun and prepositional phrases.

Therefore, it is important that the corpus contains extensive use of prepositional phrases and predicates.

About information required in a corpus, for example, [2] notes different use of prepositional phrases according to the text genre. [13] found significant differences of subcategorization frequencies in different corpus. Discourse and semantic influence were identified as the sources for those differences. The former is caused by the changes of language form used in different types of discourse. The semantic influence is based on semantic context of discourse. Therefore a corpus with different genres should be quite adequate for our goal.

About the big size of the corpus, the current corpora have a range from millions of words to hundred of millions of words depending on its type, i.e. plain text or diverse type of annotations. [1] argue that a corpus must be big enough to avoid sparse data and reflect natural use of language in order to obtain a good probability approximation. They use the one million of words Wall Street Journal corpus. Other authors, on the contrary, do not use the whole corpus for their research but a subcorpus with specific characteristics [12] which requires huge manual efforts. Therefore, we consider the size of tenths of millions of words and Web-based sources.

2.1. Corpus Annotation

There are several levels for corpus annotation: lexical, syntactic, semantic, etc. and even more levels of annotation could exist inside each level. Lemma and part of speech assignments are considered in lexical annotation. There is a wide variety of annotation schemes. For example, the Penn Tree-bank [9] uses 36 marks for part of speech and 12 for punctuation and other symbols. The Brown Corpus [6] uses 87 simple annotations but it permits composed annotations.

The sentence structure in the syntactic level is generally showed grouping words by parenthesis, and labeling those groups additionally. Since a complete structure requires more learning time of the scheme by annotators and more time for sentence annotation, there are different grades of sentence hierarchic structure realization. For example, in the Penn Tree-bank development the distinction of sentence arguments and adjuncts was ignored in a first stage. Nevertheless, the argument annotation is crucial for the semantic interpretation of verbs. In the semantic level, it has been considered the signification annotation and a type of concept.

The first step in the compilation of a Mexican Spanish corpus is the lexical level; other stages will consider the syntactic and semantic annotation.

3. Corpus Compilation

3.1. Plain Text and Word Compound

The WEB organization of the four Mexican newspapers daily published permitted us an automatic extraction for monthly and yearly periods in a quite short time. The texts correspond to diverse sections: economy, politics, culture, sport, etc. from 1998 to 2002.

The size of the original texts was 2540 MB from which we obtain 1592 MB in plain text with some marks. These texts were obtained by the following steps:

1. HTML labels deleting. Since there is no consistent use of HTML formats even inside each newspaper, several programs were developed to obtain plain texts.
2. Article structure assignment. The texts were automatically labeled with marks of title, subtitle, text body and paragraph. The paragraph assignment was respected as it was defined in the Internet publication (not exactly as its paper publication). The paragraphs were automatically split in sentences by means of punctuation-based heuristics.
3. "Wrong" and "correct" word assignment. We obtained all different words from the texts. The "correct" words were automatically annotated using the orthographic tool of a word processor and two Spanish dictionaries. Therefore, "correct" words were those recognized by such resources. Initially, from 747,970 total different words, 60% were marked as wrong words, among them we encountered the following three cases:

a) Some specific Mexican words. A manual non exhaustive work let us identify words used in Mexico that does not appear in DRAE[1], neither in María Moliner dictionary, like *ámpula*, but it appears in DEUM[2]. Other "wrong" words were words of Indian origin (náhuatl, maya, otomí, etc.). For example: *zotehuela, xochimilca, xoconostle*, etc.
To override this problem we obtained lists of this type of words from the WEB and from specialized manuals in order to compare text words against that of Web-based sources. We obtained a new group of "correct" words of Indian origin.

b) Word forms not contained in dictionaries. Some heuristics based on suffixes were used to detect diminutives, plurals and other regular word forms.

c) Composed words by hyphens. Groups of words connected by hyphens are quite common in English. On the contrary, stylistic Spanish manuals suggest a minimal use of hyphens. However, newspaper texts show the use of hyphens for diverse purposes:

- Emphasize a group of words. Ex: *única-y-mejor-ruta* (the only and best route)
- Evocate slogans. Ex: *sí-se-puede* (it is possible)
- Indicate slow pronunciation of one word. Ex: *é-x-i-t-o* (successful).
- Give different attributes. Ex: *étnico-nacionalista* (ethnical, nationalist)
- Show links among corporations or real names. Ex: *ABB-Alsthom*, *ADM-Dreyfus-Novartis-Maseca*
- Give a route. Ex: *Durango-Zacatecas*
- For some English words. Ex: *e-mail*, *e-business*, *zig-zag*, etc.
- Indicate football matches or other kind of sport games. Ex. *Pachuca-Santos*.

There were found some mistakes introduced by copying texts of the newspaper edition containing the natural use of hyphens for syllable separation. Ex: *De-rechos* (ri-ghts)

All the previous cases were treated first as a compound of words if the words individually could be considered as "correct" words. If such heuristic was wrong a second one treats them as a one word, joining the elements divided by hyphens, if the whole element could be found as a "correct" word. In general, a tag of noun or adjective can

---

[1] Spanish language dictionary of the Real Spanish Academy. Espasa, Calpe, 21 ed. 1995
[2] Usual Spanish in Mexico Dictionary. Ed. Colegio de México. México, 1996.

be assigned. Future work should consider some semantic mark-up for all detected uses of hyphens.

Since the orthography of words of Indian origin is not quite well known by Mexican speakers, there are many spelling mistakes of this type. A future work will consider some kind of error correction, since typographic error, spelling check errors and words in capital letters without accents were considered as wrong words.

4. Linking of composed prepositions. There are many composed prepositions in Spanish besides simple prepositions. Words groups as *al cabo de, con respecto a*, require a manipulation as a set (*con_respecto_a, a_fin_de, al_cabo_de*).
5. Foreign words. The Web-based sources contain foreign words. We use the orthographic tool of a word processor to detect some English and French "correct" words.
6. Proper names annotation. There were found 168,333 different words with capital letters either initializing each word or totally filling them. They are repeated in the texts counting 1804,959 occurrences. Initially, manual work was considered for 160 words having more than 1000 occurrences but they count only 420964 of total words (23% of total occurrences). Therefore, some kind of proper name identification is necessary.

- Acronyms. Ex: PRD, PT, ONU
- Proper names including prepositions and coordination. Ex: *Convergencia por la Democracia, Ley del Impuesto sobre la Renta, Luz y Fuerza del Centro*.
- Names. Ex: *San Lázaro, Benito Juárez, El séptimo sello.*

The corpus building involves tasks of different nature as we already described in the above subsections. Among them the detection of correct words and the proper name identification are required. The former needs complex tools to detect and correct wrong spelling. The second is being developed and it is described in sections 4 and 5.

3.2. Lexical Level Annotation

There are Spanish corpora with lexical level annotations, for example the LEXESP[3] corpus. As LEXESP corpus has been used for research purposes at our laboratory, we decided to use the same 275 different lexical labels for our corpus. The main reason for such quantity of labels in Spanish is agreement (gender, person and number).

The LEXESP corpus has the PAROLE[4] categories that considers the following POS classification:

- Adjective (A). Example: *frágiles* <AQ0CP00>
- Adverb (R). Example: *no* <RG000>
- Article (T). Example: *la* <TDFS0>
- Determinant (D), see figure 1. Example: *tal* <DD0CS00>
- Noun (N). Example: *señora* <NCFS000>
- Verb (V). Example: *acabó* <VMIS3S0>

[3] The LEXESP corpus was kindly provided by H. Rodríguez from Universidad Politécnica de Cataluña, Barcelona, Spain.
[4] http://www.ub.es/gilcub/castellano/proyectos/europeos/parole.html

Table 1. Characteristics of determinants

Type		Person	Gender		Number		Case	Posses
Value	Key		Value	Key	Value	Key		
Demonstrative	D	1	Feminine	F	Singular	S	0	0
Possessive	P	2	Masculine	M	Plural	P		
Interrogative	T	3	Common	C	Invariable	N		
Exclamatory	E							
Undefined	I							

- Pronoun (P). Example: *ella* <PP3FS000>
- Conjunctions (C). Example: *y* <CC00>
- Numerals (M). Example: *cinco* <MCCP00>
- Prepositions (SPS00). Example: *ante* <SPS00>
- Numbers (Z). Example: *5000* <Z>
- Interjections (I). Example: *oh* <I>
- Abbreviations (Y). Example: *etc*. <Y>
- Punctuation (F). All punctuation signs (.,:;-¡!'¿?"%). Example: "." <Fp>
- Residuals (X). The words that do not fit in previous categories. Ex.: *sine* <X> (Latin)

Example of the POS for the ambiguous word *bajo*:

```
bajo  bajar(verb) <VMIP1S0>   bajo(preposition) <SPS00>
      bajo(adverb) <RG000>    bajo(noun) <NCMS000>
      bajo(adjective) <AQ0MS00>
```

We only detail the complete key in PAROLE for determinants (see Table 1), showing the considered features. The "common" value in gender is employed for both feminine and masculine, for example *alegre* (cheerful). The "invariable" value in number is used for both singular and plural, for example: *se* (personal pronoun).

The POS annotation was realized with the MACO [4] program developed by the Natural Language Processing group of the Artificial Intelligence section of Software Department in Polytechnic University of Catalonia in collaboration with the Computational Linguistic Laboratory of the Barcelona University. In addition, some modifications have been included in the labels to consider marks for clitics inside verbal forms.

4. Composite Proper Names in Newspaper Texts

We selected one Mexican newspaper (MNP#2) in order to analyze how composite proper names could be identified and delimited. A Perl program extracted groups of words that we called "compounds" containing capitalized words joined with no more than three functional words. They were left and right limited by a punctuation mark and a non-capitalized word if they exist. We manually analyzed the compounds of approximately 2000 sentences randomly selected.

We classified the proper names in two categories: non-ambiguous and ambiguous identification.

Non-ambiguous identification. Most of compounds without functional words are proper names with non-ambiguous identification, mainly those with high frequency scores. Other characteristics defining non-ambiguous identification are: 1) Redundancy**:** information obtained from juxtaposition of proper names and acronyms, for example: *Asociación Rural de Interés Colectivo* **(***ARIC***),** and 2) Flexibility: long proper names do not appear as fixed forms. For example: *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 names translated from foreign languages but the differences are functional words.

Ambiguous identification. Three main syntactic causes of ambiguous identification are: coordination, prepositional phrase attachment, and embedded sentences. The last one corresponds to names composed of several words where only the first one is capitalized, titles of songs, books, etc. As far as we observed, the titles found in MNP#2 are no delimited by punctuation marks. This use is rather different to that considered in CoNLL-2002 files, where the titles are delimited by quotation marks.

- *Coordination*. In the simpler and more usual case, two names with the conjunction cover the last embedded substructure. For example: *Ley de Armas de Fuego **y** Explosivos, Instituto Nacional de Antropología **e** Historia.* However, there are other cases where the coordinated pair is an internal substructure of the entire name, for example: *Mesa de* [*Cultura y Derechos*] *Indígenas*. Other compounds are ambiguous since one coordinated name could be also coordinated, for example: *Comisión Federal de Electricidad **y** Luz y Fuerza del Centro, Margarita Diéguez y Armas **y** Carlos Virgilio*.
- *Prepositional phrases*. We consider a diverse criterion than that considered in CoNLL: in case a NE is embedded in another one or in case a NE is composed of several entities all should be identified since syntactic analysis should find their relations for deep understanding. For example: *Teatro y Danza de la UNAM* (UNAM's Theater and Dance), *Comandancia General del Ejército Zapatista de Liberación Nacional* (General command of..) A specific grammar for composite proper name identification should cope with the already known prepositional phrase attachment problem. Therefore, diverse knowledge was considered to decide on splitting or joining prepositional phrases.

5. Method

Identification of composite proper names in our texts collection was mainly based on their syntactic-semantic context, on discourse factors and on their specific construction. The heterogeneous knowledge contributions required are:

- *Local context*. It has been considered in different tasks. We use it for title identification. Two words preceding the capitalized word were defined as left limit; one of them could be a cue of the manually compiled list of 26 items plus synonyms and variants (gender, number). For example, in the following sentences the cue is underlined: *En <u>su libro</u> La razón de mi vida* (*Editorial Pax*)... (In his <u>book</u> The reason of my life (Pax publisher), *...comencé <u>a releer</u> La edad de la discreción de Simone de Beauvoir,…* (I began to <u>reread</u> Simone de Beauvoir's The age of discretion),

...en su programa Una ciudad para todos que... (in his program A city for all that …)

The right limit consider all posterior words until a specific word or punctuation sign is found, they could be: 1) proper name, 2) sign of punctuation (period [10], comma, semicolon, etc.) and/or 3) conjunction. In the above examples: "(", "*Simone de Beauvoir*" and the conjunction "que" delimit the names.

– *Linguistic knowledge*. It is settled by linguistic restrictions, they mainly correspond to preposition use, discourse structure and punctuation rules. For example: 1) lists of groups of capitalized words are similar entities, and then the last one should be a different coordinated entity. For example: *Corea del Sur, Taiwan, Checoslovaquia y Sudáfrica*. 2) preposition use, considering the localization meaning, direction meaning, etc. For example: two proper names joined by destination preposition ("a", "hasta") should be separated if they are preceded by a preposition denoting an origin position ("de", "desde"). For example: ***de** Salina Cruz **a** Juchitán*.
– *Lists*. Many NER systems use lists of names. We included two Web-based sources (list of personal names: 697 items, list of main Mexican cities: 910 items) and one manually compiled list of similes [3] to disambiguate identification of coordinated proper names.
– *Heuristics*. One example is: a personal name should not be coordinated in a single proper name. For example: *Margarita Diéguez y Armas y Carlos Virgilio*, where Carlos is an item of personal names list.
– *Statistics*. Statistics of groups of capitalized words were obtained from MNP#2 (from one word to three contiguous words, capitalized words related to acronyms). The top statistics for such groups were used to split composite names, for example: *Comandancia General del Ejército Zapatista de Liberación Nacional* could be separated in: *Comandancia General* and *Ejército Zapatista de Liberación Nacional.*

Knowledge application

A Perl program process sentences in two steps to disambiguate identification of composite proper names: 1) compounds extraction using a dictionary with part of speech, 2) decision on splitting, delimiting or leaving as is each compound. The order of decisions in the second step is: 1) look up the compound in the acronym list, 2) decide on coordinated groups ambiguity using the list of similes, rules, and statistics, 3) decide on prepositional phrase ambiguity using rules, lists, heuristics and statistics, 4) delimit possible titles using context cues, rules, and statistics, and 5) decide on the rest of groups of capitalized words using heuristics and statistics.

Results

We test the method on 200 sentences of MNP#4. They were manually annotated and compared against the results of the 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 labelled

Table 2. Results in a testing set of sentences

	NUMBER OF:			
	COORD. GROUPS	PREP. PHRASE GROUPS	TITLES	ALL
Precision	61	70	55	88
Recall	51	68	55	85

Table 2 indicates the performance for coordinated names (53 items), prepositional groups (84 items), and titles (11 items). The last column shows the performance for total proper names (753 items) including the previous ones. The main causes of errors were: 1) over splitting of prepositional phrases, 2) foreign words (names, cities, prepositions), 3) names missing in the available lists, and 4) missing signs for delimitation of titles and punctuation marks inside titles.

Conclusions

We presented the development of a resource for the linguistic processing of Mexican Spanish texts: a collection of Web-based sources to build an annotated corpus. We detailed the things found and results on text collection, processed automatically. We explained the process used for part of speech annotation. The main advantage of our method is that most of the work was realized in an automatic form to reduce time and costs.

One of the main problems in lexical level annotation was proper name identification. We present a method for identification of composite proper names (names with coordinated constituents, names with several prepositional phrases, and names of songs, books, movies, etc.). 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.

The strategy of our method is the use of heterogeneous knowledge to decide on splitting or joining groups of capitalized words. Results were obtained on 200 sentences corresponding to different topics. The preliminary results show the possibilities of the method.

References

1. Berthouzoz, C. and Merlo, P. Statistical ambiguity resolution for principle-based parsing. In Proceedings of the Recent Advances in Natural Language Processing (1997) 179–186
2. Biber, D. Using Register. Diversified Corpora for general Language Studies. Computational Linguistics Vol. 19–2 (1993) 219—241
3. 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

4. Carmona, J., S. Cervell, L. Màrquez, M.A. Martí, L. Padró, R. Placer, H. Rodríguez, M. Taulé & J. Turmo. An Environment for Morphosyntactic Processing of Unrestricted Spanish Text. First International Conference on Language Resources and Evaluation. Granada, Spain (1998)
5. Chinchor N.: MUC-7 Named Entity Task Definition http://www.itl.nist.gov/iaui/894.02/ relatedprojects/muc/proceedings/muc7toc.html#appendices (1997)
6. Francis, W. N. and Henry Kučera. Frequency Análisis of English Usage: Lexicon and Grammar. Houghton Mifflin (1982)
7. Gelbukh, A., G. Sidorov, and L. Chanona-Hernández. Compilation of a Spanish representative corpus. Proc. CICLing-2002, 3rd International Conference on Intelligent Text Processing and Computational Linguistics, Mexico City. Lecture Notes in Computer Science N 2276, Springer-Verlag (2002) 285–288
8. Kilgariff, A. *Web as corpus*. In: Proc. of Corpus Linguistics Conference, Lancaster University (2001) 342–344
9. Marcus, M., Santorini, B. and Marcinkiewicz, M. Building a large annotated corpus of English The Penn Treebank. Computational Linguistics Vol.19–2 (1993)
10. Mikheev A.: Periods, Capitalized Words, etc. http://www.ltg.ed.ac.uk/~mikheev/ papers.html
11. MUC: Proceedings of the Sixth Message Understanding Conference. (MUC-6). Morgan Kaufmann (1995) http://www.itl.nist.gov/iaui/894.02/related_projects/tipster/muc.htm
12. Ratnaparkhi, A. Statistical Models for Unsupervised Prepositional Phrase Attachment. In Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics. Montreal, Quebec, Canada (1998) http://xxx.lanl.gov/ps/cmp-lg/9807011
13. Roland. D. and D. Jurafsky. How Verb Subcategorization Frequencies are Effected by Corpus Choice. In Proc. International Conference COLING-ACL'98. Quebec, Canada (1998) 1122–1128
14. Tjong Kim Sang, Erik F. Introduction to the CoNLL-2002 Shared Task: Language-Independent Named Entity Recognition. http://lcg-www.uia.ac.be/~erikt/papers/