Automatic Building Gazetteers of Co-referring Named Entities

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

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

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

Abstract

Noun phrase (NP) co-reference resolution is a problem involved in many Natural Language areas, such as Dialog, Information Extraction, Summarization and Question Answering, among others. Especially important issues regarding this problem are the detection of aliases and the detection and expansion of acronyms. In this sense, terminological and general gazetteers of Named Entities (NEs) being aliases and of pairs acronym-expansion can be helpful. This paper describes a methodology to acquire semi-automatically these gazetteers.

1. Introduction

Noun phrase (NP) coreference resolution is a problem involved in many Natural Language areas, such as Dialog, Information Extraction and Retrieval, Summarization and Question Answering, among others. Especially important issues regarding this problem are the detection of aliases and the detection and expansion of acronyms. In this sense, terminological and general gazetteers of Named Entities (NEs) including aliases and of pairs acronym-expansion can be helpful.

Different approaches have been applied in order to build these resources. Manually building these collections implies a hard time/human cost in keeping the collection updated, with acceptable coverage of new possible items. Automatic or semi-automatic approaches try to deal with such a drawback. Commonly, they are based on mining document collections (closed or not) by means of the use of techniques such as those involved in Information Retrieval, Information Extraction or Clustering.

Here we present a methodology that combines Information Extraction, Clustering and Machine Learning techniques for semi-automatically obtaining these gazetteers. Preliminary resources have been built from the AQUAINT[1] corpus. Section 2 describes the method used to extract aliases from a given document collection. Section 3 focuses on specific kind of aliases, acronyms or abbreviations. Section 4 states some conclusions and further work.

2. Acquiring Co-referent Named Entities

A NE is alias of another one if it is more general and can be used to refer the same world entity (e.g. *Valdés* is an alias of *Juan Valdés* and *Albert Valdés*). The acquisition of NEs that are aliases from corpora implies the following problems:

- Recognizing NEs.
- Classifying NEs.
- Deciding whether two NEs are aliases.

Recognizing consists on locating a sequence of one or more contiguous words that can be considered candidate to be a NE and deciding if it is an actual one. Classifying implies assigning a class from a closed dataset to the NE. Most Named Entity Classification (NEC) systems reduce this set to the basic MUC classes: LOCATION, PERSON, etc., while finer grained classification has been faced recently in extended NEC (Sekine et al, 2002). NE recognition and classification tasks can be carried out in sequence or merged into a unique task (NERC). In our work we use *Abionet*, a NERC system (Carreras et al, 2002). This system can deal with documents written in several languages, such as English, Spanish, Catalan and German (Carreras et al., 2003), and is easily portable to new languages and document types.

We do not face in this paper the related problem of NE disambiguation (i.e. mapping a recognized and classified NE into its real world referent), although disposing of a gazetteer of aliases could undoubtly help in this task.

In order to decide whether the alias relation between pairs of NEs holds, a measure of NE similarity is used. The global measure computes the maximum of four simple ones: the fact of being prefix, suffix or infix one of another, and the fact of having a very low number of ortographic differences, probably orthographic errors, such as having one different letter or having a permutation of two consecutive

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

{*Fort_Monroe*, *Monroe*, *Monroe_Avenue*, *Monroe_County*, *Monroe_Drive*, *Monroe_High_School*, *Monroe_St.*, *Monroe_Street*, *Monroe_Township*, *West_Monroe*}

{*AT&T_Network_Systems_Group*, *Comverse_Network_Systems_Ltd.*, *Coyote_Network_Systems*, *Coyote_Network_Systems_Inc.*, *Geocast_Network_Systems*, *Geocast_Network_Systems_Inc.*, *Hughes_Network_Systems*, *International_Network_Systems*, *Network_Systems*, *Shenzhen_Liming_Network_Systems*, *Tasmania_Network_Systems*, *Triton_Network_Systems*}

{*L.J.*, *L.J._Dragovic*, *L.J._Shelton*, *L.J._Smith*}

Figure 1: Clusters examples.

letters, among others. This latter feature has been computed in terms of the ratio of trigrams (in characters) shared by both NEs.

This measure is used to build clusters of NEs having a common referent. For instance, a proper name such as *Valdés* can refer to different persons, such as those named as *Juan Valdés* and *Albert Valdés*. This means that we can build the cluster {*Juan Valdés*, *Albert Valdés*, *Valdés*} of NEs having *Valdés* as possible referent. This referent is taken as the centroid of the cluster. Other possible clusters are {*Juan Valdés*, *Juan*} and {*Albert Valdés*, *Albert*} having *Juan* and *Albert* as centroids, respectively.

Note, however, that using such a definition of cluster, a NE can belong to more than one cluster at the same time. Clustering methods commonly used (Everitt, 1993) are not suitable for this task. In general, these methods build a partition of the elements, meaning that an element cannot belong to more than one cluster.

In order to deal with our particular problem, we have used a Clustering approach in which non-empty intersections between clusters are allowed. Initially, a cluster is built for each NE from the input set of them. This NE is taken as the centroid and the unique component of the cluster. At each step, one of these NEs is selected to be included in other clusters, those which centroid is similar enough to the particular NE. This is decided by considering a threshold value of the similarity metric described before between both NEs (0.85 by default). In this sense, the centroid is a common alias of the rest of clustered NEs. After this process, those clusters consisting of only one element are removed because they are not productive.

2.1. Execution and results

In order to obtain a set of clusters of NEs sharing a common alias, we have used the AQUAINT corpus. This corpus is a large collection of news in English (more than 3 Gbytes) extracted from the Associated Press Journal (APW), the New York Times (NYT) and the Xinhua English (XIE).

The whole corpus was firstly pre-processed with the *TnT* POS-tagger (Brants, 2000), then we obtained the lemmas using the lemmatizer included in *WordNet*[2] (version 1.7.1). Then we have used *Abionet* to extract the NEs from the AQUAINT pre-processed corpus. This system classifies NEs into 4 classes: person (PER), organization (ORG), location (LOC) and other. We have only considered the first three classes. In table 1, the number of extracted NEs is shown. The process of clustering explained above has been applied to the three sets of NEs extracted from the AQUAINT. The results of this process is a set of NE clusters sharing a common alias. Figure 1 shows some examples of generated clusters (the centroid is underlined).

NE type	Num. NEs AQUAINT
LOC	147,276
ORG	302,743
PER	577,971
Total	1,027,990

Table 1: Number of NEs extracted from AQUAINT corpus.

NE type	Num. Clusters	Avg.ClusterSize
LOC	31,346	5.86
ORG	83,126	9.54
PER	145,772	7.14
Total	260,244	7.75

Table 2: Results of clustering.

The results of the process are summarized in table 2. As shown in this table, the average number of NEs per cluster was 5.86, 9.54 and 7.14, for classes LOC, ORG and PER, respectively. These results are expected because names of persons and locations are generally shorter than names of organizations. In figure 2, we can see the distribution of size of the clusters.

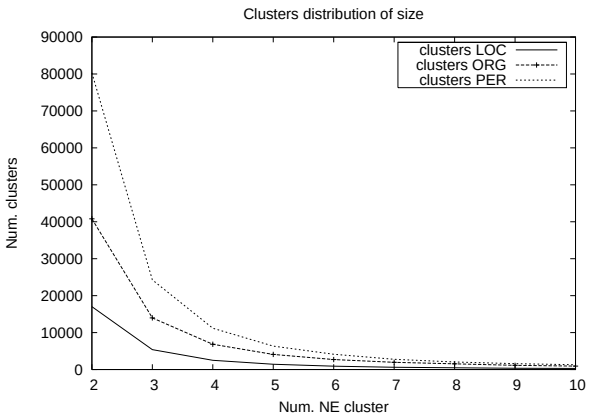

Figure 2: Clusters distribution of size.

[2]http://www.cogsci.princeton.edu/~wn/

2.2. Evaluation

In order to evaluate the results, we have randomly selected 200 clusters per NE type from the set of generated clusters, a total of 600 clusters. We have manually revised all the selected clusters to verify that the cluster centroid is an alias of the rest of clustered NEs. This revision has been done without taking into account recognition and misclassification errors. Because of the difficulty of getting all the possible NEs that could be classified as a member of a cluster, we have only computed the precision measure. The following measures have been used:

- Microaverage precision: obtained dividing the global sum of correct NEs by the global sum of NEs.
- Macroaverage precision: for each cluster we compute its precision and then we compute the mean of these results.

NE type	MaPr	MiPr
LOC	0.9377	0.9571
ORG	0.9478	0.9767
PER	0.9162	0.9498
Total	0.9339	0.9638

Table 3: Results of manual clusters evaluation.

The results of manual revision of the 600 clusters are shown in table 3. We observe that our methodology achieves good precision. This was as expected due to the use of the affix measures, that can't introduce errors in the cluster. Besides, the orthographic measure can capture some minimal orthographic differences or orthographic errors such as: {*Robert_Schwartzman*, *Robert_J._Schwartzman*}, but it also can produce errors as: {*Sanford_Stadium*, *Stanford_Stadium*}.

3. Acquiring acronym-expansion pairs

An important issue regarding co-referring NE is the detection and expansion of acronyms. Some authors make the distinction between acronyms and abbreviations, the former referring to multiword entities (as "USA" for "United States of America") while the latter to single word entities (as "Ltd." for "limited" or "Barna" for "Barcelona"). Although techniques for looking for abbreviation and acronym expansion could be slightly different, see (Toole J., 2000), we consider an acronym an abbreviate reference to a Named Entity so, for us, the first and third examples above can be considered acronyms while the second no (because its expansion is not a NE). An attempt to apply the same technique applied for aliases to acronyms (using, of course, a different distance measure) resulted in a serious degradation of the accuracy. So we used a different approach for obtaining lists of <acronym, expansion> pairs. As in the case of general NE, we do not face in this paper the related task of acronym disambiguation. For instance, two valid expansions for IBM can be "International Business Machines" and "Intercontinental Ballistic Missiles". Disposing of a gazetteer of this kind can, however, help in the task of acronym expansion.

Several problems have to be faced for building such gazetteers:

- Locating terms candidate to be classified as acronyms.
- Deciding whether a candidate (in a context) is really an acronym (and not, for instance, an abbreviation of a common name).
- Expanding the acronym, i.e. looking for a phrase (usually a NP) that can be considered as co-referent or expansion of the acronym.

For the first issue we have followed a very simple approach (loosely inspired in *Acrophile* system, (Larkey et al, 2000)). We have obtained an initial list of highly confident acronyms (from (WWWAAS)). From this list (containing about 8,000 items) we have extracted a set of regular patterns covering all its content. Applying this set to the whole set of AQUAINT corpus a total number of 576,880 candidates has been obtained (note that for getting this list we are interested on having a good recall not a good precision).

Obviously, the list of candidates includes a lot of noise. After pre-processing the corpus as in section 2.1, we have applied a decision tree technique for learning a classifier that classifies the candidates as acronyms or not. We have used 88 features for this classification task (see table 4). These features include orthographic information (length, number of points, digits, quotes, vowels, case, etc. of the candidate), morpho-syntactic information of the candidate and its immediate context (up to two tokens before and after the candidate), regular patterns satisfied by the candidate (he have used the set of ten canonical and contextual patterns proposed in *Acrophile*), etc. For instance, '^[A-Z]+[/-][A-Z]+$' matches 'AFL-CIO'.

For training we have started with a subset of our initial list of 8,000 items (a shallow manual revision for removing unclear occurrences was performed). About 10% was reserved for testing. We have looked for all the occurrences of these acronyms in two available corpora: the AQUAINT corpus and the British National Corpus (BNC[3]). From these two corpora we have extracted the immediate context of each occurrence as well as the needed morphological information. This procedure resulted in a total of positive examples with a similar number of (assumed to be) negative examples, randomly extracted from our set of candidates, not belonging to the initial list and with a shallow manual revision. The whole set of examples was sent to a C4.5 classifier (we have used *SIPINA*[4] for this purpose) for obtaining a decision tree that was translated into a set of 29 Prolog rules. An accuracy of about 93% was obtained on the test corpus.

As a result of the performance of the classifier over the set of 576,880 candidates, a total of 85,112 acronyms were found in AQUAINT.

For expanding the acronyms we have used a set of rules that are applied on NP candidates provided by a chunker (Ageno, 2003) and occurring within a predefined window (in our experiments the window is reduced to the paragraph

[3] http://www.natcorp.ox.ac.uk/

[4] http://eric.univ-lyon2.fr/~ricco/sipina.html

Feature type	Features
orthographic information	- length. - number of dots. - number of uppercase letters. - number of digits. - has dots. - last char is dot. - all letters are uppercase. - number of dots is equal to number of letters. - some letters are uppercase. - only has a final dot. - only first char is uppercase. - only has uppercases and digits. - num. of uppercases in previous word. ...
morpho-syntactic information	- actual word is a proper noun. - actual word is a common noun. - previous word is a proper noun. - previous word POS is DT. - previous word POS is (. - following word POS is IN. ...
patterns	- matches *Acrophile* pattern 1 ('^[A-Z][,\.-/_])+$'). - matches *Acrophile* pattern 2 ('^([A-Z])+$'). ...

Table 4: Some examples of features used by the acronym classifier.

where the acronym occurs). These rules basically measure the similarity between the acronym and its possible expansion. They have been manually built and own a credibility score. For instance, the highest scored rule looks for a NP spanning a number of capitalized words equal to the length of the acronym and beginning each word with the corresponding letter in the acronym (e.g. "International Business Machines" could be a good expansion of "IBM"). These constraints are relaxed in other (up to 11) less scored rules for allowing, for instance, the introduction of noisy material (as the preposition "of" in "United States of America") that could, in this way, co-refer to "USA". Although the recall increases using these less scored rules, the drop in precision is dramatic. Using finer grained rules and reducing the size of the window can be good lines for further investigation. Good results have been obtained, for instance, using co-reference in terms occurring in appositions.

4. Conclusions and Further Work

Two methods for building gazetteers for co-referring NEs have been presented. The first tries to cluster together groups of NEs that can be co-refered by the same alias. The second deals with the extraction of acronym-expansion pairs. The two methods have been applied to the AQUAINT corpus for building a set of 260,244 clusters and 85,112 acronyms. The future work includes three basic objectives:

- Applying these methods to other languages (Spanish and Catalan) and collections (BNC, EFE).
- Using other similarity measures (such as edit distances (Arslan et al., 2003)) in the clustering process.
- Improving the expansion rules of the acronym-expansion module experimenting with different windows and weighting schemas.

5. Acknowledgments

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

6. References

Ageno, A. 2003. *An Island-Driven Parsing System.* PhD thesis, Departament de Llenguatges i Sistemes Informàtics, Universitat Politècnica de Catalunya.

Arslan, A.N., Egecioglu, O. 2000. Efficient Algorithms For Normalized Edit Distance. *Journal of Discrete Algorithms,* 1,1, pp.3-20.

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

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

Carreras, X., Márquez, L., Padró, L. 2002. Named Entity Extraction Using Adaboost. *Proceedings of the 6th conference on Computational Natural Language Learning (CoNLL 2002). Shared Task Contribution,* Taipei, Taiwan.

Everitt, B. 1993. *Cluster Analysis.* Edward Arnold corp.

Larkey, L. S., Ogilvie, P., Price, M.A., Tamilio, B. 2000. Acrophile: An Automated Acronym Extractor and Server. *Proceedings of the Fifth ACM Conference on Digital Libraries,* ACM Press, pp. 205-214, San Antonio, USA.

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

Toole, J. 2000. A hybrid approach to the identification and expansion of abbreviations. *Proceedings of RIAO 2000,* vol. 1, pp. 725-736, Paris, France.

WWWAAS. World Wide Web Acronym and Abbreviation Server, http://www.ucc.ie/cgi-bin/acronym.