Spanish Automatic Text Enrichment

Mariano Felice, Fernando R.A. Bordignon y Gabriel H. Tolosa
marianofelice@yahoo.com, {bordi, tolosoft}@unlu.edu.ar

Universidad Nacional de Luján
Departamento de Ciencias Básicas
Laboratorio de Redes

Abstract. Unlike text on paper, hypertext enables the linking of pieces of text with other texts and multimedia resources, which not only improves the way we read but also lays the foundation for new information systems. Specifically, the proliferation of collaborative sites, blogs, online databases, encyclopedias and many other services on the World Wide Web provides an invaluable source of up-to-date information which can be used to aid reading comprehension. As a result, an approach to the automatic extraction, merging and integration of online information is proposed for the purpose of "enriching" texts. This unprecedented text enrichment process allows users to transform ordinary plain texts into self-explanatory hypertexts containing contextual information and resources selected automatically from the Web. Application of such an enrichment process could help students in their scholarly reading, provide users with related multimedia resources and avoid multiple searches for concepts and entities mentioned in a text, among other purposes.

Keywords: automatic text enrichment, named entity recognition, entities, NER, hypertext, web, texts.

1 Introduction

From its origins, hypertext (regarded as a digital document for non-sequential reading) has revolutionized the way we read, becoming an alternative to traditional text on paper which is static and linear by nature. Besides providing a quicker, more interactive and dynamic experience to readers, hypertext introduces an ideal scheme to achieve thorough text comprehension thanks to the linking of phrases with explanatory and illustrative resources. As a result, any simple text could be endowed with self-explanatory qualities, for example by including definitions of key concepts, descriptive data, contextual information on facts or people and related multimedia resources.

Today, the World Wide Web offers its users an extraordinary range of hypertexts on practically any topic, often with rich links and resources that help gain a deeper understanding of its contents. The invaluable contextual framework favored by hypertext has made web searching a quick, practical and advantageous strategy to approach unknown subjects.

However, the availability of texts that users may find interesting on the Web is often limited by a variety of factors, most of which have been explored by many authors:

a. the lack of information on the Web about the requested topic, often due to its great specificity, little worldwide knowledge or spreading of the subject,
b. the absence of hypertexts containing terms or phrases of great significance to the reader,
c. the excessive amount of counterfeit, commercial and irrelevant contents (*noise*) [1],
d. the potentially low or questionable quality of texts about the requested topic [2][3],
e. the availability of hypertexts that, despite being of high quality, fail to interest users or whose links and resources are found to be insufficient, uninsteresting, unreliable or unavailable,
f. the "invisibility" of high quality sites [4],
g. the low availability of Spanish texts (only 4.6% of the entire Web) [5],
h. the influence of search tools on the contents finally accessed by users, either by their ranking of results or users' habit of disregarding results beyond the first result page [6][7],
i. the potential inability of users to make good use of search engines, e.g. by supplying inaccurate or poorly specified queries [8][9], and
j. the fact that hypertexts are generally the result of human production.

To overcome many of these obstacles, users have traditionally turned to search engines [10], although they cannot guarantee user satisfaction or the quality of the resources provided by their results. Alternatively, many practical tips have been given to users [11][12] and some computational models designed [13][14][15] in order to assess the quality of contents on the Web. However, there is yet a better way for users to obtain hypertexts which best meet their needs: let them provide base texts to be transformed into hypertexts by some automatic mechanism. This approach does not necessarily imply that users should write their texts and establish hyperlinks but rather choose a suitable text and submit it to an automatic linking tool.

Such a tool would allow users to generate their own hypertexts from plain texts which are of interest to them, with no need to search for or depend on the existence of hypertexts on the Web that completely fulfil their needs. This task of automatically converting text into hypertext and its consequent search and linking of relevant resources is presented here as *Automatic Text Enrichment* (ATE) and is the subject of this paper.

This enrichment concerns the linking of concepts and phrases in a text with relevant data and resources available on the World Wide Web. All selected textual items, known as *entities*, serve as the starting point for searching and selecting resources and represent the links to the collected information.

The main advantage of Automatic Text Enrichment is its ability to integrate relevant resources into a text in an automatic way. Namely, this process finds entities in a text and links them to contextual boxes with relevant information and resources available on the Web without human intervention.

One of the main features of the proposed architecture lies in the fact that it has been applied to texts in Spanish, which poses many interesting challenges both from a

lexical/semantic view of base texts and the limited availability of resources on the Web.

The final result of the enrichment process is a new hypertext whose entities found in the original text have been transformed into hyperlinks to boxes with detailed contextual information and carefully selected web resources. Consider the following text as an example:

Natalia Oreiro junto con Greenpeace convocan a marcha azul por las ballenas

Buenos Aires, Argentina — La actriz Natalia Oreiro se sumó a la Campaña contra la Caza Comercial de Ballenas de Greenpeace para convocar, a través de un spot que circula por internet, a una gran marcha azul por las ballenas, que se realizará el próximo domingo 27 de mayo a las 14 en el obelisco porteño.

Más información en www.greenpeace.org

The application of the proposed enrichment process to the text above should result in the identification of all entities and provide information on each and every one of them. An illustration of this result is shown in Fig. 1.

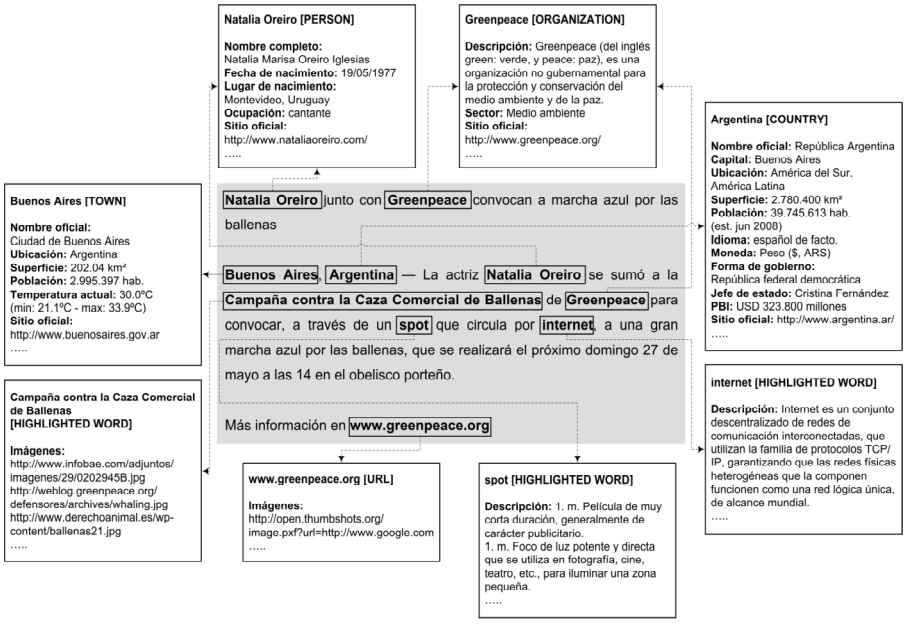

Fig. 1. An enriched document.

The architecture described in this work is principally aimed at enriching informational or explanatory texts, which is why all subsequent development has been optimized to work with those types of text. Anyway, this specification does not necessarily limit the use, application or adaptation of the architecture to other types of

text that could also be satisfactorily enriched. This proposal combines techniques both from Information Retrieval (IR) and Information Extraction (IE). More precisely, the architecture and methodology described here involve three main tasks:

Named Entity Recognition: A subtask of IE aimed at identifying elements in a text which can be classified according to a previously defined taxonomy [16]. These elements are called *entities* and have traditionally represented locations, organizations and people. In this work, however, classification has been extended to include continents, countries, towns, locations, email addresses, URLs, historical events, people, organizations, trademarks and other significant words. Moreover, new recognition techniques have been proposed and tested.

Entity Enrichment: A task whose goal is to characterize entities by linking them to related resources retrieved from the World Wide Web. Although similar works have been developed by companies and universities, this work provides the first formal description of the task, its goals and methodology.

Result Presentation: A task concerning the optimal presentation of enriched texts, according to the principles of Information Visualization [17][18] and Human-Computer Interaction [19].

1.1 Contributions

To the best of our knowledge, this work presents the first formal approach to Automatic Text Enrichment, a task aimed at automatically complementing a text with resources from the Web. Specific contributions are summarized below:

- the first characterization and formal specification of the concept of "Automatic Text Enrichment",
- new guidelines for Named Entity Recognition, like the adoption of dynamically updated gazetteers and reuse of entities previously resolved by coreference,
- the first architecture and methodology for Entity Enrichment,
- *Identity Checking*, a simple solution for the disambiguation of entities in the resource extraction phase,
- the application of the proposed methodology to Spanish texts.

The reminder of this work is organized as follows: Section 2.1 defines Automatic Text Enrichment and proposes an architecture for its development while Section 2.2 describes some implementation details. Section 3 discusses some possible applications and audiences of the enrichment process. Finally, Section 4 concludes while Section 5 discusses future work.

2 Automatic Text Enrichment

2.1 Definition and Architecture

Automatic Text Enrichment is defined as the process of generating a hypertextual version of a linear text by adding related data, images, audio, video, text, services,

hyperlinks, metadata and any other resources associated with each of the entities or concepts identified in the text, for the purpose of providing further information about them.

Such enrichment comprises three main tasks: 1) Named Entity Recognition, 2) Entity Enrichment and 3) Result Presentation. The first of these tasks is carried out to find the items in the text that should be enriched, the second collects resources related to those items and the third shows the results in an appropriate format according to the aims of the enrichment. To sum up:

$$ATE = Recognition + Enrichment + Presentation$$

Each of the tasks involved in the enrichment process is carried out by a module which receives an input document and outputs another. Modules communicate by the results they produce so that the output of one module is the input to the next.

Actually, the Automatic Text Enrichment process begins with an initial plain text, with no markup or format whatsoever. This initial text is considered to be linear as it does not contain any kind of links to navigate through its contents nor does it include references to external information. Generally, initial texts will be provided by users, although they could also be retrieved automatically in cases where the ATE process is implemented as a post-processing tool for other texts, such as emails, web pages, digital documents, etc.

The modules in this architecture interact with auxiliary components as shown in Fig. 2.

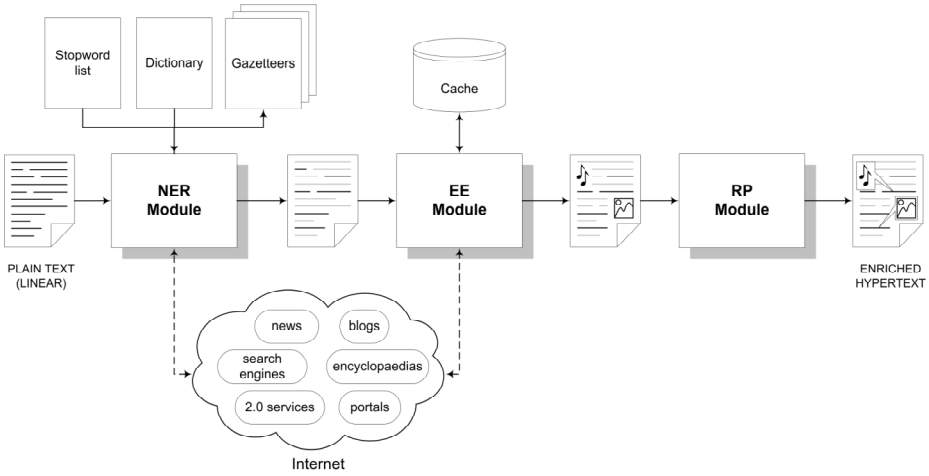

Fig. 2. Automatic Text Enrichment Architecture.

The input text is initially processed by the Named Entity Recognition (NER) module, which detects and marks up all entities using any available technique. In fact, a third-party NER module could be used as long as results conform to the format used by the next module. In this work, some common NER techniques such as gazetteers have been used together with other ad hoc strategies. The resulting recognized documents is later processed by the Entity Enrichment (EE) module, which collects

resources related to each of the entities and links them to their corresponding references within the analyzed text. This module uses the World Wide Web as the main source of information although and internal cache is also maintained with every collected data and resource so that repetitions of previously processed entities can be enriched quickly. All the resources collected are integrated with the original text into a new document and forwarded to the Result Presentation (RP) module, which produces a hypertext for an appropriate display and navigation of its contents.

The core component of the architecture is the Entity Enrichment module, which is where the real enrichment of the original text takes place. This is achieved by effectively seeking resources for each entity and integrating them into the text. The steps involved in this process are summarized in the following algorithm:

```
With a text A previously processed by the NER module, do:
FOR EACH entity e of type t recognized in A
  IF e has not been enriched in A yet THEN
    I_t^e <-P_t
    IF the entity e of type t is stored in cache AND is
    not expired THEN
      Fill I_t^e with resources from cache
    ELSE
      FOR each service s that satisifes P_t
        FOR each cacheable resource i in P_t
          IF resource i has not been stored in I_t^e AND S
          is available AND able to provide resource i
          THEN
            r <- f_s(e,i)
            Store r in I_t^e
          END IF
        END FOR
      END IF
    END IF
    FOR each non-cacheable resource i in P_t
      FOR each service s that is able to provide i
        IF resource i has not been stored in I_t^e AND S is
        available THEN
          r <- f_s(e,i)
          Store r in I_t^e
        END IF
      END FOR
    END FOR
    Store I_t^e in the enriched document
    Store I_t^e in cache
  ELSE
    Link e to its previous enrichment within the
    enriched document
  END IF
END FOR
```

This algorithm carries out many search and selection processes in order to collect resources from different complementary web services (like DBpedia[1], Wikipedia[2], KIM[3], YAGO[4] and Freebase[5]) and enrich each entity. To make this process efficient, the aforementioned cache must be implemented and some other problems solved. An example of this is *Identity Checking*, a method intended to disambiguate entities with the same name and type in the resource collection phase. With this technique, entities can be characterized by a very limited dataset (such as date of birth and date of death for people) which allows them to be compared to other exemplars retrieved from the many services consulted and thus determine whether they refer to the same entity. In that case, the resources retrieved are linked to the entity, otherwise they are discarded. This ensures the correct enrichment of ambiguous entities such as "George Bush" or "Roberto Carlos" without mixing resources related to homonymous people.

Additional implementation details have been omitted due to space constraints but are described extensively in [20].

2.2 Implementation

The architecture described above can be implemented in different ways, like a standalone application, a web service for end users or a post-processing tool, depending on who is expected to provide the input text. For the purpose of this work we implemented a web service that takes plain texts as input and produces enriched versions according to the proposed methodology. To make the resource collection process easier, web services providing structured or semi-structured information (such as tables, tagged content, etc) have been preferred over the rest since their results can be easily obtained with parsing techniques. It should be noted that the services chosen for enrichment depend strongly on the availability of Spanish sites on the World Wide Web. Therefore, in cases where we have been unable to find reliable services in that language, sites in English were used instead, together with suitable translation mechanisms. Fig. 3 shows a screenshot with a sample enriched text.

3 Applications and Target Audience

The architecture described here is mainly intended to enrich informational and explanatory texts in Spanish, such as news articles, biographies, historical texts, reviews, essays, encyclopedia articles and short stories to name but a few, although this bias does not necessarily limit its application on other types of texts where this could be useful.

---

1 http://dbpedia.org
2 http://es.wikipedia.org
3 http://ontotext.com/kim/
4 http://www.mpi-inf.mpg.de/yago-naga/yago/index.html
5 http://www.freebase.com

This kind of text enrichment is especially beneficial in cases where a thorough and deep understanding of texts is essential. Some typical users and scenarios are:

- students who wish to enrich scholarly texts in order to gain a deeper understanding and get information on unknown concepts,
- people who read news articles and demand detailed and updated information on people or facts mentioned in the texts,
- users who wish to avoid repetitive searches for definitions or resources related to key concepts in their texts,
- people interested in complementing reading with multimedia resources,
- users in search of a contextual information tool.

Although the architecture described here has been built into a functional prototype for end users, it could well be implemented as a plugin for existing applications or processes, making up a software layer for enrichments. This kind of integration would allow users to enrich online news articles or personal emails just as they read them, without having to export them to external applications.

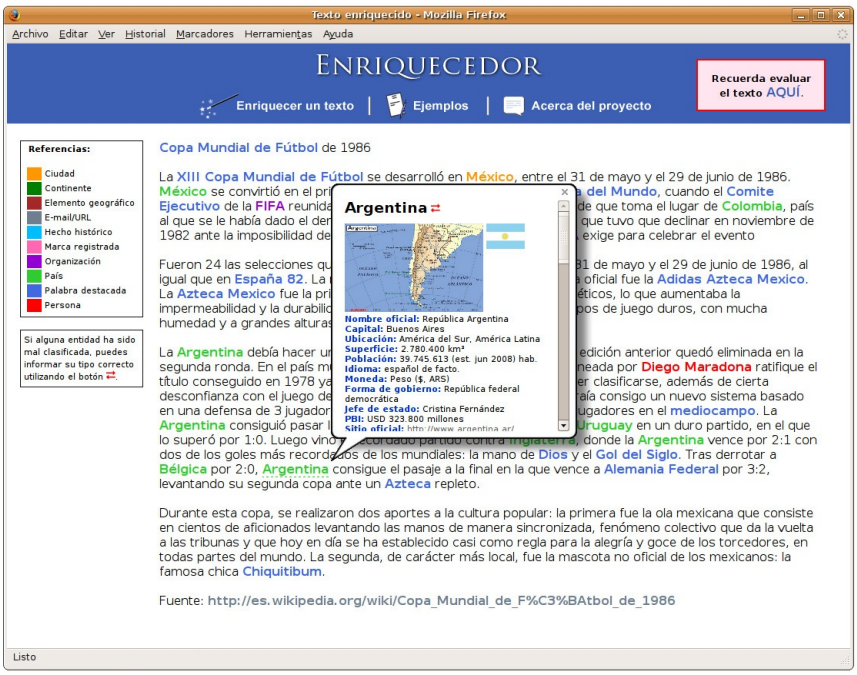

Fig. 3. An example of enriched text.

Besides, implementations could be tailored for specific purposes. Commercial implementations, for instance, might enrich entities with links to shopping websites or even include “sponsored” enrichments, acting as a new advertising platform for companies. One typical example could be the enrichment of singers or bands mentioned in a text with links to websites offering music downloads, tickets for their shows, CDs and DVDs and related merchandising. However, business is not the only

alternative purpose. Entertainment, leisure or academia are only a few examples of the many areas enrichment could be applied to, providing a careful selection of specific services which are available for each case.

4 Final Thoughts

Automatic Text Enrichment adds a new dimension to reading owing to the use of hypertexts and the ability to generate enriched versions of linear plain texts. The advantages of hypertext over paper have been widely reported by authors, philosophers and researchers but with the advent of new technologies many new applications remain to be explored. Text enrichment is one of them, specifically designed to improve reading. With this technique, users can focus on entities mentioned in a text and get concise up-to-date information about them without leaving the original reading space.

Unlike other existing services (such as Evri[6] or Yahoo! Glue[7]), the enrichment proposed in this paper produces dynamic, up-to-date, heterogeneous contents on top of an initial base text provided by users.

5 Future Work

First of all, an scheme for the evaluation of efficiency of the enrichment methodology should be designed and carried out in order to measure performance. These evaluations could be done from different perspectives, such as the number of resurces collected, number of correctly recognized entities, resource retrieval precision, data freshness, disambiguation efficiency, user satisfaction, etc. Knowledge of these measures is essential to fine-tune the techniques used in each case and improve performance.

As for the components of the ATE architechture, different NER approaches should be tested and evaluated in order to achieve optimal recognition.

Another key aspect that needs further study is user interaction. In this regard, many features could be explored, such as the adoption of user profiles to refine enrichment. This is tightly related to the different applications and purposes of enrichement, all of which must be analyzed in detail. As mentioned earlier, this work presents a generic architecture with an informational bias which can be easily adapted to meet different needs, ranging from educational to commercial applications.

Finally, new processes for the inteligent selection and filtering of the best resources for each entity remain to be explored.

6 http://www.evri.com
7 http://glue.yahoo.com

References

1. Piper P.S.: Better read that again: Web hoaxes and misinformation. Searcher 8, vol. 8, 40--49 (2000)
2. Martin-Facklam M., Kostrzewa M., Martin P., Haefeli W.E.: Quality of drug information on the World Wide Web and strategies to improve pages with poor information quality. An intervention study on pages about sildenafil. Br J Clin Pharmacol 2004, vol. 57, nro. 1, 80--85 (2004)
3. Bedell S.E., Agrawal A., Petersen L.E.: A systematic critique of diabetes on the World Wide Web for patients and their physicians. Int J Med Inform. 2004, vol. 73, nro. 9, 687--694 (2004)
4. Sherman C., Price G.: The Invisible Web: Uncovering Information Sources Search Engines Can't See. Cyberage Books, Medford, NJ (2001)
5. Accenture: La difusión del español en Internet. Fundación Caja de Burgos, Burgos (2006)
6. iProspect Search Engine User Behavior Study (April 2006). Technical report, iProspect.com, Inc. (2006)
7. Jansen B.J., Spink A.: How are we searching the world wide web? A comparison of nine search engine transaction logs. Information Processing & Management, vol. 42, nro. 1, 248--263 (2006)
8. Sisson D.: Assumptions About User Search Behavior. A thoughtful approach to web site quality, http://www.philosophe.com/search/user_behavior.html (2003)
9. Rose D.E., Levinson D.: Understanding user goals in web search. In: WWW '04: Proceedings of the 13th international conference on World Wide Web, pp. 13—19. ACM Press (2004)
10. Fallows D.: Search Engine Use. Memo, Pew Internet & American Life Project (2008)
11. Tillman H.N.: Evaluating Quality on the Net, http://www.hopetillman.com/findqual.html (2003)
12. Harris R.: Evaluating Internet Research Sources, http://www.virtualsalt.com/evalu8it.htm (2007)
13. Dondio P., Barret S.: Computational Trust in Web Content Quality: A Comparative Evalutation on the Wikipedia Project. Informatica, vol. 31, nro. 2, 151--160 (2007)
14. Herrera-Viedma E.: Fuzzy Qualitative Models to Evaluate the Quality on the Web. In: Torra, V., Narukawa, Y. (eds) MDAI 2004. LNAI, vol. 3131, pp. 15--27. Springer, Heidelberg (2004)
15. van Gils B., Proper E., van Bommel P., van der Weide T.P.: On the quality of resources on the web: An information retrieval perspective. Information Sciences, vol. 177, nro. 21, 4566--4597 (2007)
16. Chinchor, N: Named Entity Task Definition (Version 2.1). In: Proceedings of the Sixth Message Understanding Conference (MUC-6), pp. 317--332 (1995)
17. Spence R.: Information Visualization. ACM Press, New York (2000)
18. Fluit C., Wester J.: Using visualization for information management tasks. In: Proceedings of the Sixth International Conference on Information Visualisation, pp. 447--454. IEEE Press (2002)

19. Myers, B.A.: A Brief History of Human Computer Interaction Technology. ACM interactions, vol. 5, nro. 2, 44--54 (1998)
20. Felice, M: Enriquecimiento Automático de Textos. Trabajo Final de Licenciatura, Universidad Nacional de Luján (2009)