Multi-domain Cross-lingual Information Extraction from Clean and Noisy Texts

Horacio Saggion[1] **and Sandra Szasz**[1]

[1]Departament of Information and Communication Technologies
Universitat Pompeu Fabra
C/Tanger 122 - Campus de la Comunicación
Barcelona - 08018
Spain
{horacio.saggion,sandra.szasz}@upf.edu

Abstract. *We have created a human-annotated, multi-event, cross-lingual corpus of equivalent summaries in Spanish and English to investigate cross-lingual information extraction. The corpus contains, in addition to pairs of equivalent non-translated summaries, automatic translations of each summary produced using an available translation tool. We have developed trainable information extraction systems per language and have applied them to both original summaries and their automatic translations obtaining encouraging results.*

Resumo. *Apresentamos um estudo de extração de informações de um corpus bilíngüe paralelo em espanhol e inglês. O corpus está formado por pares de resumos curtos de eventos em três domínios de aplicação. Temos desenvolvido sistemas de extração de informaçoes para as duas línguas estudadas e avaliado o desempenho do sistema em varias experiências tanto monolíngües como translíngües. Apresentamos uma análise dos resultados obtidos.*

1. Introduction

Creating knowledge repositories in specific application domains and populating them from textual sources would be an impossible task without the appropriate information extraction tools. In this paper we address cross-lingual information extraction, which consists on developing an information extraction system for a given source language and applying it to another target language. Such tool would be appropriate for extracting information which is only available in the target language but for which only limited language processing tools exist. We address the language pair Spanish/English for which, and to the best of our knowledge, there has been no past research. Spanish is widely used with over 500 million speakers worldwide, it is the third Internet language in terms of content, and it is the official language of 21 nations in addition to be spoken in many non-Spanish speaking countries. It it is therefore particularly important to develop information extraction technology for Spanish. The documents we have dealt with in this research are event summaries in Spanish and English which provide in a condensed form information about specific events. There are various reasons to work with summaries:

- they are easily found on the Web and on document collections. For example, it would be difficult to ignore the number of summaries available in public sources such as Wikipedia or in background pieces of news about particular events which frequently include condensed descriptions, i.e., summaries, of past similar events.

Proceedings of the 8th Brazilian Symposium in Information and Human Language Technology, pages 11–19,
Cuiabá, MT, Brazil, October 24–26, 2011.

- they contain the key pieces of information of an event but in natural language instead of in tabular form;
- they still are complex textual units which deserve special attention from the scientific community and, finally,
- non-extractive summaries such as those we present in this paper offer opportunities for research into abstractive summary generation.

In order to carry out this research we have created a comparable corpus of available summaries in Spanish and English (i.e., the clean data). We have also enriched the corpus with automatic translations (i.e., the noisy data) and have manually annotated both the clean and noisy data to create a valuable resource for the scientific community. The automatic translations were produced using the Google Translate software available on the Web (`http://translate.google.com`). In this research we test the performance of trainable information extraction systems in various conditions including: training/testing in clean data, training/testing in noisy data, and training in clean data and testing in noisy data. Because summaries by definition provide key information about a domain, they offer a great potential for the extraction of domain specific information and for the creation of structured sources of knowledge (e.g., ontologies or knowledge repositories). Also because summaries are concise, they offer increased advantages compared to extraction of information from full document collections: knowing that the summary contains just the key elements of an event certainly could reduce extraction mistakes.

The rest of this paper is structured as follows: in Section 2 we describe related work and then, in Section 3 we describe the data set created for the study of cross-lingual extraction. After that, in Section 4 we describe the automatic tools used to process documents. Finally, Section 5 reports experiments and discusses the results and Section 7 closes the paper.

2. Related Work

Information extraction is the mapping of natural language texts (e.g. news articles, web pages, e-mails) into predefined structured representations or templates [Grishman 1997]. Information extraction is a complex task carried out by human analysts on a daily basis. Because it is very time-consuming and labour-intensive, there has been much research over the last 20 years to automate the process. The field of information extraction has been fuelled by two major US international evaluations efforts. From 1987 until 1997 the Message Understanding Conferences (MUC) [Grishman and Sundheim 1996, Cowie and Lehnert 1996] concentrated on template-based information extraction. After MUC, the interest was changed to content extraction in the ACE evaluations [ACE 2004] where semantics more than linguistic analysis was the focus. There was also interest on systems able to easily adapt to new languages and tasks. In recent years there has been an increasing interest in multilingual as well as cross-lingual information extraction with a number of events organized on the subject [Poibeau and Saggion 2007, Poibeau et al. 2008]. Using rule-based information extraction in three different languages and robust graph-based event linking, the MUSING project [Saggion et al. 2003] demonstrated how extraction could be used to improve multimedia indexing in multiple languages. Related to this is work on cross-lingual retrieval: [Hakkani-Tür et al. 2007] use an information extraction system in English as a filtering step to improve retrieval of Chinese documents. As a key technology for information extraction is named entity recogni-

tion, multilingual named entity recognition is also relevant. [Steinberger et al. 2007] use extensive resources and rule-based systems developed through bootstrapping processes to identify and match names in over 8 languages.
Related to the work presented here is also research related to the creation of corpora of summaries for natural language processing applications. We have identified the SummBank corpus [Saggion et al. 2002] created for the study of multi-lingual summarization in Chinese and English which is suitable for cross-lingual summarization but not for information extraction tasks. The SumTime-Meteo Corpus [Reiter and Sripada 2002] provides weather summaries in English from numerical data and are potentially useful in data to text generation applicationsn and might be suitable for summary-to-template applications.

3. Data Set Creation and Annotation

In its current state, the dataset we work with is a corpus of equivalent summaries in Spanish and English in three different domains: aviation accidents, rail accidents, and earthquakes. Further domains will be incorporated in the future for researchers interested in evaluating the robustness and adaptation capabilities of different natural language processing techniques. In order to collect the summaries, a keyword search strategy was used to search for documents on the Internet using Google Search. Keywords per domain were defined and used to select a set of Web pages in Spanish, for example the keywords "lista de terremotos" ("list of earthquakes") could be used to find out pages on earthquakes. The pages returned by the search engine were examined to verify if they actually contained an event summary and in that case a document was created for the summary (it is usual to find multiple summaries in a single Web page). The documents were given names indicating the type of event and the date of the event/incident. A set of around 50 summaries per domain in Spanish were collected in this manner. After this, for each event summary originally in Spanish the Internet was searched for an equivalent English summary (not a translation) using keywords in English, manually derived from the Spanish summary. For example if an earthquake event mentioned a particular date and intensity, then those elements were used as keywords. Following this procedure we found equivalent English summaries for most of the Spanish ones.
For each domain a set of semantic components were identified based on intuition and on the actual data observed in the set of summaries. Some examples of semantic information are as follows:

- For aviation accidents: the airline, the cause of the accident, the date of the accident, the destination, the flight number, the origin of the flight, etc.
- For railway accidents: the cause of the accident, its date, its destination, its origin, the number of passenger, the number of survivors, etc.
- For earthquakes: the city affected, the country affected, its date, its epicentre, the number of fatalities, etc.

Corpus examples (pairs of summaries in the two languages) for the aviation domain are shown in Table 1. In order to manually annotate the summaries with semantic information, we have used the GATE annotation framework [Maynard et al. 2002]. To facilitate the annotation process an annotation schema was used so that in the GATE Graphical User Interface the target text span to be annotated can be selected, and annotated with one valid category from the annotation schema. One annotator was in charge

Aviation Accident

(A1) 2009 4 de agosto: El vuelo 622 de Bangkok Airways, se disponía a enlazar dos de los más importantes centros turísticos: Krabi y Koh Samui. Pero al aterrizar, se sale de pista y se estrella contra la torre de control, y se incendia. Fallece el piloto, y 41 personas resultan heridas. El aparato, un ATR-72, tenía poco más de ocho años, y ya había operado anteriormente para Bangkok Airways. (Spanish original)

(A2) 2009 August 4 Bangkok Airways Flight 266, an ATR 72-200 carrying 68 passengers crashes in severe weather on landing at Samui airport in the resort island of Ko Samui in Thailand, resulting in at least 1 confirmed death and 37 injuries. (English original)

(A3) 2009 August 4: Flight 622 from Bangkok Airways, was about to link two of the most important tourist centers, Krabi and Koh Samui. But upon landing, exit the track and crashes into the control tower and fire. Pilot dies, and 41 people injured. The device, an ATR-72, had just over eight years and had previously operated for Bangkok Airways. (English translation of A1)

(A4) 2009 04 de agosto - Bangkok Airways Vuelo 266, un ATR 72-200 llevar a 68 pasajeros se estrella en el mal tiempo al aterrizar en el aeropuerto de Samui, en la turística isla de Ko Samui en Tailandia, con al menos una muerte confirmada y lesiones 37. (Spanish translation of A2)

Table 1. Sample of the parallel corpus; Spanish and English parallel texts (A1, A2) and their Google Translate translations (A3, A4).

of the annotations and a curator controlled the annotations for any inconsistency. Note that because we are dealing with short texts, the annotation process is less complex than that of annotating a full event report. Figure 1 shows the two components of a corpus pair annotated in the annotation tool. More detailed information about the corpus is given in [Saggion and Szasz 2011a].

4. Text Analysis Components

All summaries were analysed by automatic processes as described below:

The English summaries were linguistically analysed by the default text analysis and named entity recogniser distributed with the GATE system. Although this is a system not trained on the type of data we are dealing with, we needed an off-the-shelf system to come up with basic linguistic information such as parts-of-speech and general named entities. The components we have used from the GATE system are a sentence identification program, tokenizer, parts-of-speech tagger, rule-based morphological analysis, dictionary lookup, and named entity recognition and classification. The Spanish summaries were linguistically analysed with two components: an adaption of the TreeTagger software [Schmid 1995] so that it can be executed from the GATE system and our own named entity recognizer. TreeTagger provides tokenisation, parts-of-speech tags for each word, and morphological (lemma information) analysis for Spanish (the default trained system was used). Named entity recognition is carried out using a machine learning component developed using Support Vector Machines trained over data from the CoNLL evaluation program [Tjong Kim Sang and De Meulder 2002]. The CoNLL 2002 Spanish dataset which provides information on named entities such as *Location*, *Organization*, *Person*, and *Miscellaneous* was analyzed using parts-of-speech tagging and morphological analysis from the TreeTagger. The named entity recogniser is based on SVMs classi-

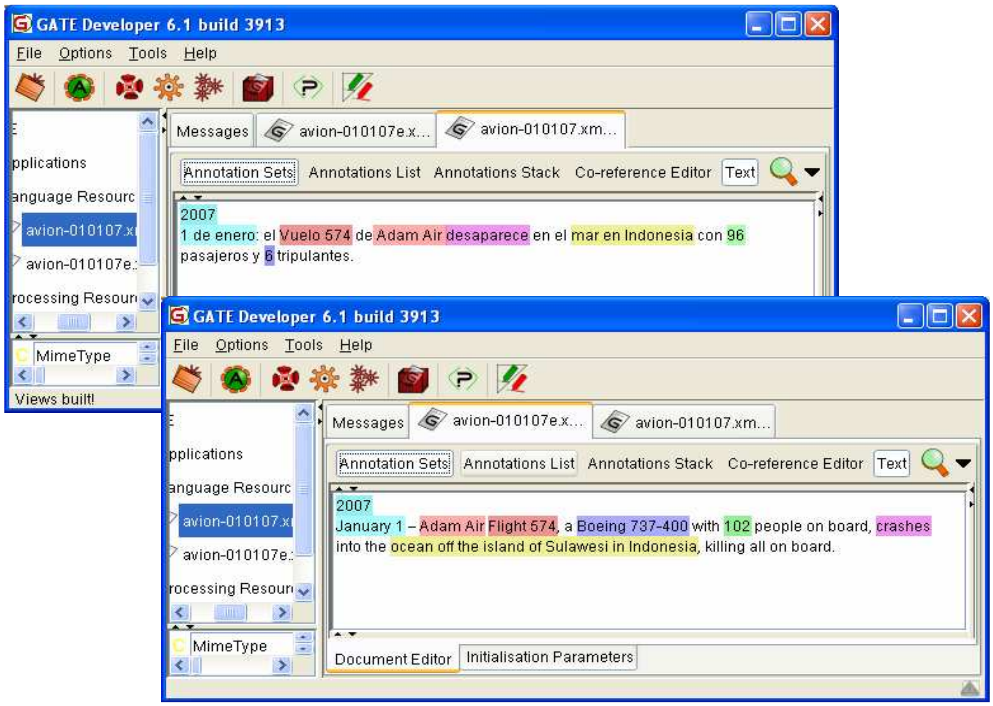

Figure 1. Spanish and English summaries manually annotated with semantic information pertaining to aviation accidents (pair corresponding to an accident on 1 January 2007).

fication [Li et al. 2004] trained over word roots, parts-of-speech, and orthographic information using context windows of 5 words around the token to be classified. It achieves an F-score performance of 68% in cross-validation experiments.

5. Information Extraction System

To develop the Spanish and English information extraction systems we used as machine learning algorithm a Support Vector Machines (SVMs) implementation integrated in the GATE framework [Li et al. 2004]. The SVMs treat the extraction problem as one of chunk-learning, where each token in the document has to be classified into a set of potential begin or end labels pertaining to the different information types defined in the learning task. Taking the aviation domain as an example, the types of information (or concepts) we are interested in identifying in the summaries are: the *Airline*, and the *TypeOfAircraft* among others. In the chunk-learning tasks this is reduced to identifying which tokens have labels *Begin-Airline, End-Airline, Begin-TypeOfAircraft, End-TypeOfAircraft*. Given a set of tokens with their associated labels and their context, the SVMs is trained and used to classify unseen tokens. After "begin" and "end" tokens have been identified for a concept T, the whole concept can be recovered as the shortest span staring at a begin annotation of type T and ending at an end annotation of type T. The SVMs needs two probabilities to be defined for the token classification problem: the probability that a particular token is entity boundary, and the probability that a given sequence of tokens is an entity. These two probabilities are set experimentally through training-testing cycles. In SVMs classification, two different approaches are considered: (i) in a "one vs all" classification approach

each class (e.g., Airline) is compared to all other classes therefore creating different classifiers for each given label; (ii) in a "one vs another" classification approach a given class (e.g., Airline) is compared against each other class (e.g., TypeOfAirchaft) therefore creating a classification problem for each pair of classes. Basic linguistic features were used to train Spanish and English extraction systems:

- The Spanish system uses for each token to be classified a context window of five tokens. Features extracted from each token in the context are: orthographic information (e.g., word capitalization), root information, parts-of-speech tags and named entity type of each token;
- The English system uses for each token to be classified a context window of five tokens. Features extracted from each token are: orthographic information, root information, parts-of-speech tags, type of named entity, and dictionary information (from the gazetteer lookup process).

6. Experiments, Results, and Discussion

Given the relatively small size of the dataset, the following leave-one-out experimental setting was adopted, where one document is left-out for testing and the rest of the documents are used for training:

- *Monolingual experiments in Spanish:* the Spanish extraction system is trained on the original Spanish summaries and applied to the held-out Spanish summary;
- *Monolingual experiments in English:* the English extraction system is trained on the original English summaries and applied to the held-out English summary;
- *Cross-lingual experiments in Spanish:* the Spanish extraction system is trained on the original Spanish summaries and applied to a Spanish translation of an English summary;
- *Cross-lingual experiments in English:* the English extraction system is trained on the original English summaries and applied to an English translation of a Spanish summary;
- *Translation experiments in Spanish:* the Spanish extraction system is trained on the translated Spanish summaries and applied to the held-out translation;
- *Translation experiments in English:* the English extraction system is trained on English translated summaries and applied to an English translation;

For the cross-lingual experiments, given the test summary T, the training is the set of all summaries except the summary which is equivalent to T. In this way we make sure that the extraction system has not seen the data in the test set. In each experiment we computed the performance of the extraction system using precision and recall measures. Precision is the proportion of correct answers. Recall is the proportion of correct answers returned by the system. Precision and recall are aggregated in an F-score measure where precision and recall are equally weighted. The final performance is obtained aggregating the F-scores of all datapoints tested. In Table 2 we present information extraction results for the clean data in both languages, more detailed information on monolingual experiments is reported in [Saggion and Szasz 2011b]. For train and aviation accidents, the English system performs better than the Spanish, this is probably because text analysis in English is more robust. Where the earthquake domain is concerned, the Spanish system performs better than its English counterpart perhaps due to the distribution of information

Event	Prec	Rec	F
Train Accident Spanish	0.47	0.41	0.44
Train Accident English	0.65	0.53	0.58
Aviation Accident Spanish	0.64	0.46	0.54
Aviation Accident English	0.68	0.63	0.66
Earthquake Spanish	0.61	0.46	0.53
Earthquake English	0.51	0.37	0.43

Table 2. Mono-lingual extraction performance (training in clean data and evaluating in clean data).

Event	Prec	Rec	F
Train Accident Spanish	0.87	0.60	0.71
Train Accident English	0.88	0.57	0.70
Aviation Accident Spanish	0.89	0.59	0.71
Aviation Accident English	0.80	0.56	0.66
Earthquake Spanish	0.60	0.48	0.53
Earthquake English	0.85	0.60	0.71

Table 3. Cross-lingual extraction performance (training in clean data and evaluating in translated documents).

in the corpus: the English summaries have less annotations and are more verbose, there are therefore less instances to learn from and in less regular contexts. Table 3 presents figures for the cross-lingual experiments. The first striking fact, which is counter-intuitive, is that the perfomance over translated documents is in some cases better than that observed in some monolingual cases. An analysis of the distribution of types of information in the translated documents shows that in some cases there are fewer human annotations in the translations and therefore more chances for the extraction system to get a correct answer. On the other hand the translations contain few of the difficult types of information such as destination or cause of the accident in aviation and railway domains in both Spanish and English. For the earthquake domains the performance of the cross-lingual Spanish system is similar to the mono-lingual system, however the English system is performing better over translations than over source language, here again we believe that the source summaries in English have more human annotations and therefore more chances to learn. Finally, Table 4 shows extraction results for the translation experiments, we notice that some configurations perform better than in the monolingual case, but again the distribution of information types in the summaries may well be the reason for such behavior. Note that in most cases the increase in performance is due to a higher improvement in precision, and this is because there are less information types the extraction system has to recall.

7. Conclusions, Current, and Future Work

In this paper we have presented a set of information extraction experiments over cross-lingual summaries in various domains. To the best of our knowledge this is one of the few studies on this field for the Spanish language. We have shown that our tools are able to extract full event information relying on linguistic annotations produced by off-the-shelf

Event	Prec	Rec	F
Train Accident Spanish	0.70	0.57	0.63
Train Accident English	0.76	0.71	0.73
Aviation Accident Spanish	0.82	0.75	0.78
Aviation Accident English	0.70	0.59	0.64
Earthquake Spanish	0.60	0.46	0.52
Earthquake English	0.73	0.64	0.68

Table 4. Translated extraction performance (training in noisy data and evaluating in noisy data).

robust components.
We have created the first cross-lingual dataset for the study of cross-lingual information extraction in Spanish and English and have carried out a set of experiments to show the value of the dataset, we believe that the obtained results are interesting for further research. Our current work involves the expansion of the dataset to cover additional domains such as terrorism and sports. We are working towards the integration of parsers and semantic analysers into the linguistic pipelines to improve the performance of the extraction systems. In future work we will address automatic clustering-based domain modelling from summaries and information extraction induction. We also plan to use cross-lingual extraction results to improve mon-lingual mono-document extraction.

Acknowledgments

We would like to thank the reviewers for their comments, we have answer all comments to improve the final version of this paper. We are grateful to Programa Ramón y Cajal from Ministerio de Ciencia e Innovación, Spain.

References

ACE (2004). *Annotation Guidelines for Event Detection and Characterization (EDC)*. Available at http://www.ldc.upenn.edu/Projects/ACE/.

Cowie, J. and Lehnert, W. (1996). Information Extraction. *Communications of the ACM*, 39(1):80–91.

Grishman, R. (1997). Information extraction: Techniques and challenges. In Pazienza, M. T., editor, *Information Extraction: A Multidisciplinary Approach to an Emerging Information Technology, International Summer School (SCIE-97)*, volume 1299 of *Lecture Notes in Computer Science*, pages 10–27, Frascati, Italy. Springer Verlag.

Grishman, R. and Sundheim, B. (1996). Message Understanding Conference - 6: A Brief History. In *Proceedings of the 16th International Conference on Computational Linguistics*, Copenhagen. Association for Computational Linguistics, Morristown, NJ, USA.

Hakkani-Tür, D., Ji, H., and Grishman, R. (2007). Using Information Extraction to Improve Cross-lingual Document Retrieval. In *Proceedings of the 1st Intl. Workshop on Multi-source Multi-lingual Information Extraction and Summarization Workshop*.

Li, Y., Bontcheva, K., and Cunningham, H. (2004). An SVM Based Learning Algorithm for Information Extraction. Machine Learning Workshop, Sheffield.

Maynard, D., Tablan, V., Cunningham, H., Ursu, C., Saggion, H., Bontcheva, K., and Wilks, Y. (2002). Architectural Elements of Language Engineering Robustness. *Journal of Natural Language Engineering – Special Issue on Robust Methods in Analysis of Natural Language Data*, 8(2/3):257–274.

Poibeau, T. and Saggion, H., editors (2007). *1st International Workhop on Multi-Source, Multi-Lingual Information Extraction and Summarization*, Borovets, Bulgaria. RANLP.

Poibeau, T., Saggion, H., and Yangarber, R., editors (2008). *2nd International Workhop on Multi-Source, Multi-Lingual Information Extraction and Summarization*, Manchester, UK. COLING.

Reiter, E. and Sripada, S. (2002). Squibs and discussions: human variation and lexical choice. *Computational Linguistics*, (4):545–553.

Saggion, H., Cunningham, H., Bontcheva, K., Maynard, D., Hamza, O., and Wilks, Y. (2003). Multimedia Indexing through Multisource and Multilingual Information Extraction; the MUMIS project. *Data and Knowledge Engineering*, 48:247–264.

Saggion, H., Radev, D., Teufel, S., Wai, L., and Strassel, S. (2002). Developing Infrastructure for the Evaluation of Single and Multi-document Summarization Systems in a Cross-lingual Environment. In *3rd International Conference on Language Resources and Evaluation (LREC 2002)*, pages 747–754, Las Palmas, Gran Canaria, Spain.

Saggion, H. and Szasz, S. (2011a). A Bilingual Summary Corpus for Information Extraction and other Natural Language Processing Applications. In *Proceedings of the Workshop on Iberian Cross-Language NLP Tasks*, Huelva, Spain.

Saggion, H. and Szasz, S. (2011b). Extracting Information from a Parallel Spanish-English Summary Corpus. In *Proceedings of the XXVII Conference of the Spanish Society for Natural Language Processing (SEPLN)*, Huelva, Spain.

Schmid, H. (1995). Improvements in part-of-speech tagging with an application to german. In *In Proceedings of the ACL SIGDAT-Workshop*, pages 47–50.

Steinberger, Ralf, Pouliquen, and Bruno (2007). Cross-lingual Named Entity Recognition. *Linguisticae Investigationes*, 30(1):135–162.

Tjong Kim Sang, E. F. and De Meulder, F. (2002). Introduction to the CoNLL-2002 Shared Task: Language-Independent Named Entity Recognition. In *proceedings of the 6th conference on Natural language learning - Volume 20*, COLING-02, pages 1–4, Stroudsburg, PA, USA. Association for Computational Linguistics.