A Bilingual Summary Corpus for Information Extraction and other Natural Language Processing Applications*

Un corpus bilingüe para la extracción de información y otras tareas de procesamiento de lenguaje natural.

Horacio Saggion and Sandra Szasz
Universitat Pompeu Fabra
Departament de Tecnologies de la Informació i les Comunicacions
Grupo TALN
C/Tanger 122 - Barcelona - 08018
Spain
horacio.saggion@upf.edu, sandra.szasz@upf.edu

Resumen: Presentamos un corpus bilingüe comparable en español e inglés de pares de resúmenes de tres tipos de eventos: accidentes aéreos, accidentes ferroviarios y terremotos. Cada resumen es un texto que describe de manera sucinta un evento particular. El corpus fue anotado manualmente con información semántica sobre cada evento y resulta apropiado para la experimentación en extracción de información monolingüe así como también cros-lingue.
Palabras clave: Extracción de informaciones, corpus bilingüe, resúmenes

Abstract: Cross-lingual information extraction, the task of extracting information from multiple-multilingual sources, can benefit from the availability of a corpus of equivalent documents in various languages. We present a dataset of pairs of summaries in Spanish and English in various application domains and demonstrate its use in information extraction experiments. The dataset has been manually annotated with semantic information.
Keywords: Cross-lingual information extraction, biligual corpus, summaries

1 Introduction

Cross-lingual information extraction, the task of extracting information from multiple-multilingual sources, is a problem which has received considerably less attention than extraction from mono-lingual sources. In this paper, we are concerned with the creation of a dataset for the development and evaluation of *cross-lingual information extraction* systems. Our corpus is a set of pairs of summaries in Spanish and English in various domains. An example of the dataset is shown below:

> *17 julio 2006 Isla de Java: un maremoto de magnitud 7,7 Richter de magnitud provoca un 'tsunami' que causó la muerte de 596 personas.*

> *On 17 July at 03:19:25 p.m. local time an earthquake measuring 7.7 on the Richter scale struck offshore immediately south of West Java at a depth of 10 km. The areas affected by the earthquake and resultant tsunami included the districts of Taskimalaya, Ciamis, Sukabumi and Garut in West Java province, Cilacap, Kebumen and Banyumas in Central Java and the Gunung Kidul and Bantul districts in the province of Yogyakarta. No. Deaths 500.*

These elements in the dataset are non-translated equivalent summaries which have been found on the Web. They report on the same event, in this case an earthquake, but because they are not translations of one another, they contain different information, for example the Spanish summary reports 596 people dead while the English summary

* We are grateful to Programa Ramón y Cajal from Ministerio de Ciencia e Innovación, Spain.

reports 500 people dead. The English summary is more verbose and contains information about the time of the event and various locations affected by the tremor thus being the two elements complementary. The dataset can be used for training information extraction systems, studying template-to-text bilingual generation, and automatic knowledge modelling.

This paper gives an overview of the dataset and initial experiments showing its potential application. The rest of this paper is structured as follows: Section 2 we explain related work and then, in Section 3 we describe the data set created. After that, in Section 4 we illustrate how we have used the corpus and in Section 5 we present our conclusions.

2 Related Work

There are various multilingual datasets in the machine translation field such as the Europarl Multilingual Corpus (Koehn, 2005) or the United Nations Parallel Corpus (Eisele y Chen, 2010). Related to the work presented here are those datasets prepared for text summarization or information extraction research. Among them we have identified the SummBank corpus (Saggion et al., 2002) created for the study of multi-lingual summarization in Chinese and English. The documents in this corpus are translations of one another and contain announcements of a local administration. The corpus has been used in text summarization and information retrieval experiments (Radev et al., 2003). Because of the content and annotation provided with the dataset, this corpus is probably less suitable for information extraction. The CAST corpus (Orăsan, Mitkov, y Hasler, 2003) contains newswire texts and popular science articles in English where annotations are added to indicate: (i) essential sentences, (ii) unessential fragments in sentences, and (iii) links between sentences when one sentence is needed to understand another. Because of the particular annotation schema used, the corpus has potential applications for sentence compression. The SumTime-Meteo Corpus (Reiter y Sripada, 2002) provides weather summaries in English from numerical data and is potentially useful in data to text generation applications and information extraction. The Ziff-Davis corpus contains technical documents in English and their human created summaries and has been used in text summarization experiments (Knight y Marcu, 2000). The dataset of the Message Understanding Conferences (ARPA, 1993) is probably the best known set for the development of information extraction systems.

3 Data Set Creation and Annotation

The dataset under development is a comparable corpus of Spanish and English summaries for four different domains: aviation accidents, rail accidents, earthquakes, and terrorist acts; this later subset is still under development. 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" could be used to search for documents in the earthquake domain. 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 not unusual 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, this time 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 (event or incident) a set of semantic components (i.e., slots) were identified based on intuition and on the actual data observed in a set of summaries for the domain. The slots/components making

Information	# Spa	# Eng
City	23	16
Country	47	31
DateOfEarthquake	53	36
Depth	1	4
Duration	1	3
Epicentre	7	7
Fatalities	50	35
Homeless	7	11
Injured	9	11
Magnitude	47	32
OtherPlacesAffected	27	29
Province	10	9
Region	25	25
Survivors	1	2
TimeOfEarthquake	4	21
TotalVictims	2	0

Table 3: Number of Semantic Concepts in Spanish and English Earthquake's Summaries

up the templates which model the domain are shown in Table 1.

Corpus examples (pairs of summaries in the two languages) for the three domains are shown in Table 2. 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. The summaries are annotated by one person, however a second person checks the annotations for any inconsistency. Note that because we are dealing with short texts, the annotation process is easier than that of annotating a full event report.

The number of event components found in the set of summaries is reported in Tables 3, 4 and 5.

4 Uses of the Corpus

We have started using the corpus in monolingual as well as in cross-lingual information extraction. 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) such as those we defined in Table 1. Various techniques have been used

Information	# Spa	# Eng
Airline	26	31
Cause	16	13
DateOfAccident	30	29
Destination	8	7
FlightNumber	26	31
NumberOfVictims	21	23
Origin	11	5
Passenger	5	9
Place	24	28
Survivors	5	10
Tripulation	8	6
TypeOfAccident	28	29
TypeOfAircraft	18	32
Year	31	31

Table 4: Number of Semantic Concepts in Spanish and English Aviation Accident's Summaries

Information	# Spa	# Eng
Cause	18	23
DateOfAccident	43	36
Destination	8	12
NumberOfVictims	43	37
Origin	9	13
Place	45	40
Survivors	25	20
TypeOfAccident	41	36
TypeOfTrain	30	33

Table 5: Number of Semantic Concepts in Spanish and English Train Accident's Summaries

in the development of information extraction systems including rule-based approaches relying on robust partial syntactic analysis (Appelt et al., 1993), Hidden Markov Models (Leek, 1997; Freigtag y McCallum, 1999), and a combination of supervised machine learning (Ciravegna, 2001) and weakly supervised machine learning (Yangarber, 2003; Riloff, 1996). In recent years there has been an increasing interest in the application of information extraction for the "Semantic Web" using ontologies as knowledge representation formalisms (Maynard et al., 2007; Saggion et al., 2007) as well as on multilingual and cross-lingual information extraction (Poibeau y Saggion, 2007; Poibeau, Saggion, y Yangarber, 2008). It has been shown that extraction from multiple

Incident	Semantic Schema
Aviation Accident	Airline; Cause; DateOfAccident; Destination; FlightNumber; Origin; Passenger; Place; Survivors; Tripulation; TypeOfAccident; TypeOfAircraft; Victims; Year
Railway Accident	Cause; DateOfAccident; Destination; Origin; Passenger; Survivors; TrainLine; Tripulation; TypeOfAccident; TypeOfTrain; Victims; Year
Earthquake	City; Country; DateOfEarthquake; Depth; Epicentre; Fatalities; Homeless; Injured; Magnitude; OtherPlacesAffected; Province; Region; Survivors; TimeOfEarthquake; TotalVictims

Table 1: Conceptual Information in Summaries

Aviation Accident
2009 30 de junio: el vuelo 626 de Yemenia chocó en cercanías a Comoras, en el Océano Indico.
2009 June 30 Yemenia Flight 626, an Airbus A310-300 flying from Sana'a, Yemen to Moroni, Comoros, crashes into the Indian Ocean with 153 people aboard; one 12-year-old is found clinging to the wreckage.
Railway Accident
12 enero 1997 8 muertos y 25 heridos en el descarrilamiento del tren rápido Milán-Roma en las proximidades de Piacenza (Italia).
January 12, 1997 A Pendolino train derails just before a train station at Piacenza, Italy, killing 8 people and injuring 29 others.
Earthquake
27 mayo 2006 Isla de Java (Indonesia): un terremoto de magnitud 6,2 Richter causa al menos 6.234 muertos, 20.000 heridos y 340.000 desplazados.
May 27, 2006 A powerful earthquake struck Indonesia's central province of Java early Saturday morning at 0554 Hrs local time (26 May 2254 Hrs GMT), flattening buildings and killing over 4900 people.

Table 2: Sample of the Parallel Corpus

multilingual sources can lead to improved semantic indexing (Saggion et al., 2003) when compared to monolingual or single source extraction. It has also been shown that cross-lingual extraction (Hakkani-Tür, Ji, y Grishman, 2007) can be used as a filtering step to improve retrieval in a target language.

4.1 Experiments

Our cross-lingual information extraction experiments involve the use of a system trained in a source language to extract information from translations from another language. However, to test how useful the dataset is, we have started with monolingual experiments per domain and language (e.g., six systems in total). The systems are a pipeline of text processing tools followed by a process of token classification based on Support Vector Machines (Li et al., 2002). The machine learning component was adjusted through testing and evaluation cycles. The text analysis components are as follows:

- For English: we used default processors from the GATE system: tokenizer, parts-of-speech tagger, rule-based morphological analysis, dictionary lookup, and named entity recognition and classification;

Event	Prec	Rec	F
Train Accident Spanish	0.49	0.41	0.44
Train Accident English	0.76	0.56	0.64
Aviation Accident Spanish	0.64	0.47	0.53
Aviation Accident English	0.68	0.62	0.65
Earthquake Spanish	0.62	0.48	0.54
Earthquake English	0.49	0.36	0.41

Table 6: Overall Extraction Performance in Spanish and English

- For Spanish: we used the TreeTagger software (Schmid, 1995) and our own trainable named entity recognizer.

Basic linguistic features were used to train Spanish and English extraction systems. Both the Spanish and English systems use for each token to be classified a context window of five positions containing the following token features: orthography (e.g., word capitalization), word root, parts-of-speech, named entity type, and dictionary (gazetteer lookup) information.

Because each dataset is relatively small, we have performed 10-fold cross-validation experiments reporting here aggregated precision, recall, and f-score figures. Table 6 presents the results. The English extraction system performs better than the Spanish system in the train and aviation accident domains, while the Spanish system performs better than the English one in the earthquake domain. This could be due to the fewer human annotations in the English earthquakes compared to the Spanish counterpart. It is worth noting that the English summaries are more verbose in this domain making extraction more difficult. Although the obtained results are modest, they have to be assessed taken into account the limited syntactic and semantic information available from the text processors. In order to test how the systems cope with noisy data we have translated the Spanish summaries into English and the English summaries into Spanish using Google Translator and have applied the information extraction systems to each translation. In these experiments, for each translation T in a domain D, the extraction system is trained with all documents except the document which is equivalent to T and the resulting system is applied to summary T. Evaluation metrics are also computed and aggregated over all documents. In these experiments we have obtained in most domains and languages f-scores over 0.60 which although not directly comparable with the mono-lingual results are certainly encouraging, full details on these experiments can be found in (Saggion y Szasz, 2011).

5 Conclusions

In this paper we have presented an overview of a dataset with potential interest for cross-lingual natural language processing applications. To the best of our knowledge this is one of the few datasets in this field for the pair Spanish/English. We have shown information extraction and cross-lingual extraction as potential applications of the dataset. Our current work involves the expansion of the dataset to cover additional domains such as terrorism and sports. In future work we will address automatic domain modelling from summaries and information extraction induction. We also plan to use the cross-lingual extraction results to improve mono-lingual mono-document extraction.

References

Advanced Research Projects Agency. 1993. *Proceedings of the Fifth Message Understanding Conference (MUC-5)*. Morgan Kaufmann, California.

Appelt, D.E., J.R. Hobbs, J. Bear, D. Israel, M. Kameyama, y M. Tyson. 1993. Description of the JV-FASTUS system as used for MUC-5. En *Proceedings of the Fourth Message Understanding Conference MUC-5*, páginas 221–235. Morgan Kaufmann, California.

Ciravegna, F. 2001. Adaptive information extraction from text by rule induction and generalisation. En *Proceedings of the 17th International Joint Conference on Artificial Intelligence (IJCAI 2001)*.

Eisele, Andreas y Yu Chen. 2010. MultiUN: A Multilingual Corpus from United

Nation Documents. En Nicoletta Calzolari (Conference Chair) Khalid Choukri Bente Maegaard Joseph Mariani Jan Odijk Stelios Piperidis Mike Rosner, y Daniel Tapias, editores, *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may. European Language Resources Association (ELRA).

Freigtag, D. y A. K. McCallum. 1999. Information Extraction with HMMs and Shrinkage. En *Proceesings of Workshop on Machine Learnig for Information Extraction*, páginas 31–36.

Grishman, R. 1997. Information extraction: Techniques and challenges. En Maria Teresa Pazienza, editor, *Information Extraction: A Multidisciplinary Approach to an Emerging Information Technology, International Summer School (SCIE-97)*, volumen 1299 de *Lecture Notes in Computer Science*, páginas 10–27, Frascati, Italy, Jul. Springer Verlag.

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

Knight, K. y M. Marcu. 2000. Statistics-based summarization - step one: Sentence compression. En *AAAI/IAAI*, páginas 703–710, Austin, Texas.

Koehn, Philipp. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. En *Conference Proceedings: the tenth Machine Translation Summit*, páginas 79–86, Phuket, Thailand. AAMT, AAMT.

Leek, T.R. 1997. Information Extraction Using Hidden markov Models. Informe técnico, University of California, San Diego, USA.

Li, Y., H. Zaragoza, R. Herbrich, J. Shawe-Taylor, y J. Kandola. 2002. The Perceptron Algorithm with Uneven Margins. En *Proceedings of the 9th International Conference on Machine Learning (ICML-2002)*, páginas 379–386.

Maynard, D., H. Saggion, M. Yankova, K. Bontcheva, y W. Peters. 2007. Natural Language Technology for Information Integration in Business Intelligence. En W. Abramowicz, editor, *10th International Conference on Business Information Systems*, Poland, 25-27 April.

Maynard, D., V. Tablan, H. Cunningham, C. Ursu, H. Saggion, K. Bontcheva, y Y. Wilks. 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.

Orăsan, C., R. Mitkov, y L. Hasler. 2003. CAST: a Computer-Aided Summarisation Tool. En *Proceedings of EACL2003*, páginas 135 – 138, Budapest, Hungary, April.

Poibeau, T. y H. Saggion, editores. 2007. *1st International Workhop on Multi-Source, Multi-Lingual Information Extraction and Summarization*. RANLP, September.

Poibeau, T., H. Saggion, y R. Yangarber, editores. 2008. *2nd International Workhop on Multi-Source, Multi-Lingual Information Extraction and Summarization*. COLING, September.

Radev, Dragomir Radev, Wai Lam, Arda C Elebi, Simone Teufel, John Blitzer, Danyu Liu, Horacio Saggion, Hong Qi, Elliott Drabek, y Johns Hopkins U. 2003. Evaluation challenges in large-scale document summarization. En *In: Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, páginas 375–382.

Reiter, E. y S. Sripada. 2002. Squibs and discussions: human variation and lexical choice. *Computational Linguistics*.

Riloff, E. 1996. Automatically generating extraction patterns from untagged text. *Proceedings of the Thirteenth Annual Conference on Artificial Intelligence*, páginas 1044–1049.

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

Saggion, H., A. Funk, D. Maynard, y K. Bontcheva. 2007. Ontology-based information extraction for business applications. En *Proceedings of the 6th International Semantic Web Conference (ISWC 2007)*, Busan, Korea, November.

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

Saggion, H. y S. Szasz. 2011. Multi-domain cross-lingual information extraction from clean and noisy texts. En *Proceedings of the Brazilian Symposium on Information and Human Language Technology*, Cuiabá, Brazil, 24-26 October. SBC.

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

Yangarber, R. 2003. Counter-Training in Discovery of Semantic Patterns. En *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL'03)*.