The CONCISUS Corpus of Event Summaries

Horacio Saggion & Sandra Szasz

Department of Information and Communication Technologies
Universitat Pompeu Fabra
C/Tanger 122
Barcelona - 08018
Spain
horacio.saggion@upf.edu, sandra.szasz@upf.edu

Abstract

Text summarization and information extraction systems require adaptation to new domains and languages. This adaptation usually depends on the availability of language resources such as corpora. In this paper we present a comparable corpus in Spanish and English for the study of cross-lingual information extraction and summarization: the CONCISUS Corpus. It is a rich human-annotated dataset composed of comparable event summaries in Spanish and English covering four different domains: aviation accidents, rail accidents, earthquakes, and terrorist attacks. In addition to the monolingual summaries in English and Spanish, we provide automatic translations and "comparable" full event reports of the events. The human annotations are concepts marked in the textual sources representing the key event information associated to the event type. The dataset has also been annotated using text processing pipelines. It is being made freely available to the research community for research purposes.

Keywords: Text Summarization; Information Extraction; Cross-lingual Text Processing; Annotated Event Summaries

1. Introduction

Text summarization (TS) and Information Extraction (IE) (Poibeau et al., 2012; Saggion and Poibeau, 2012; Piskorski and Yangarber, 2012) are key information access technologies the adaptation of which depends on the availability of language resources such as domain corpora. Here, we are concerned with the creation of resources for the study of multilingual and cross-lingual information extraction (Gaizauskas et al., 1997; Maynard and Cunningham, 2003) and summarization (Saggion, 2006) in Spanish and English. We have created an annotated dataset of comparable Spanish and English event summaries in four application domains with the objective of contributing to the study of TS and IE with a unique resource. We focus our research on short event summaries for various reasons: First, event summaries can be found on the Web and in newspaper collections; Second, event summaries as those we study here are rather concise, therefore being of interest for automatic text generation applications such as non-extractive summarization; Lastly, the summaries we have collected contain the key/essential information of the reported events, therefore being of value for manual or automatic domain modeling.

An example of comparable summaries in the aviation accident domain is shown below:

> 2008 January 17 - British Airways Flight 38, a Boeing 777-200ER, lands short of the runway at London Heathrow Airport in the United Kingdom. Nine of the 152 people on board are treated for minor injuries, but there are no fatalities; this is the first loss of a Boeing 777.

> 2008 17 de enero: el Vuelo 38 de British Airways (Boeing 777) sufrió un accidente al tomar tierra en el Aeropuerto de Londres-Heathrow procedente de Pekín. No hubo víctimas mortales.

An example of comparable summaries in the terrorist attack domain is shown below:

> Monday, February 19, 2007. Around midnight on Sunday, a pair of bombs exploded on the Samjhauta Express (Friendship Express), a night train going from Delhi, India to Lahore, Pakistan. At least 68 fatalities have been reported. Two more suitcases with improvised explosive devices have been found on the train. Some 13 passengers were reported injured, some with severe burns.

> 19 de febrero de 2007: Fallecen 66 personas y más de 60 resultan heridas a consecuencia de la explosión de dos bombas en un tren que enlaza la India con Pakistán.

It is important to note that because these are comparable summaries, the information in one language may differ from the information in the other language, therefore being a specially interesting dataset for applications in cross-language information extraction (Hakkani-Tür et al., 2007) and cross-lingual summarization (Saggion, 2006).

This paper describes the CONCISUS corpus of event summaries and additional resources for IE and TS. The paper also proposes a number of research scenarios for using the dataset. The paper is organized in the following way: The next Section gives an overview of related research in resources for TS and IE. In Sections 3. and 4. we describe the process of data collection and manual annotation. Section 5. gives an overview of the tools used for automatic text processing while Section 6. indicates possible experimentation frameworks. In Section 7. we describe additional resources that make up the dataset and lastly in Section 8. we close the paper with some conclusions and avenues for further developments.

2. Related Work

Over the past few years a number of initiatives have produced valuable resources for TS and IE. For example in the IE context, the Message Understanding Conferences (MUC) (Grishman and Sundheim, 1996; Cowie and Lehnert, 1996; ARP, 1993) and the Automated Content Extraction (ACE) Program (ACE, 2004) have created corpora in English and other languages which are made available to the research community. In TS the Document Understading Conferences (DUC) (Over et al., 2007) and the Text Analysis Conferences (TAC) (Owczarzak and Dang, 2010) have contributed with summarization tasks, documents, and reference summaries for both system adaptation and evaluation. There are however few datasets such as the one we describe here which, although small, provides a variety of domains and text types, and is annotated with rich domain information. 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 and Chen, 2010) or the JRC-Acquis multilingual parallel corpus (Steinberger et al., 2006): none of them is annotated with the necessary information to carry out IE or TS adaptation directly. Attempts have also been made for the automatic creation of comparable corpora in different languages using Wikipedia (Gamallo Otero and González López, 2010): such resource could be used to train statistical machine translation systems for example, but not for TS or IE. The SummBank corpus (Saggion et al., 2002) is a multilingual parallel summarization dataset in Chinese and English which has been used in large scale text summarization and information retrieval experiments (Radev et al., 2003). The 2011 edition of the TAC conference included a multilingual summarization task and as a result a multilingual summarization dataset (Arabic, Czech, English, French, Greek, Hebrew, and Hind) was created containing source documents and human and system summaries. It is however not suitable for IE.

3. Data Collection Methodology

The dataset we have created from available Web sources consist of comparable event descriptions in Spanish and English of four different domains: aviation accidents, rail accidents, earthquakes, and terrorist attacks. The methodology for collecting the data is manual, but we believe it could be implemented in a semi-automatic way. Figure 1 illustrates the data collection and annotation framework: the top of the figure indicates the collection and annotation of comparable summaries in Spanish and English. The bottom part of the figure illustrates the automatic translation of the summaries and their annotation (see Section 7.). 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 "ataques terroristas" ("terrorist attacks") could be used to find pages on terrorist attacks. 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 the event and the date of the event/incident (e.g. terrorist-attack-ddmmyyyy). For each summary in Spanish, the Internet was searched for an equivalent English summary (not a translation) using keywords pertaining to the domain and keywords of the Spanish summary (e.g. the date of the event, the location of the event). This lead to equivalent English summaries for most Spanish datapoints. Summaries in English are also given unique document ids. The ids of the equivalent summaries, the source of the summaries, and additional meta information is recorded. For each domain, semantic information components were defined (e.g. slots describing the event template) as follows:

- Aviation Accidents: Airline, CauseOfAccident, DateOfAccident, Origin, Destination, PlaceOfAccident, etc.
- Rail Accients: TrainLine, CauseOfAccident, TypeOfAccident, Victims, etc.
- Earhquakes: Magnitude, Epicentre, AffectedAreas, Fatalities, Injured, etc.
- Terrorist Attacks: Perpetrator, Victims, DateOfAttack, Injured, etc.

4. Text Annotation

The process of corpus annotation is carried out using the GATE annotation tool (Maynard et al., 2002). A GATE annotation schema per domain was defined as comprising all semantic components of the event type[1]. Text documents are loaded in the GATE annotation tool together with the appropriate schema and manually annotated by one annotator[2], note that the same schema is used for both Spanish and English. The resulting rich structure is saved in XML format. Also part of the CONCISUS Corpus is an instantiated template for each summary, examples of which are shown on Tables 1 and 2.

Tables 3 and 4 provide the dataset statistics: number of documents collected and annotated, the average sentence length, the average number of words, and the average number of domain entities (e.g. slots) in each document.

As can be appreciated, the documents are rather short but packed with semantic information. As a consequence, sentences in the texts will be rather complext with various verbs per sentence and complex syntactic phenomena used to link all semantic information.

[1] The annotation schemas are also part of the CONCISUS Corpus.

[2] The annotations are checked and corrected by a curator.

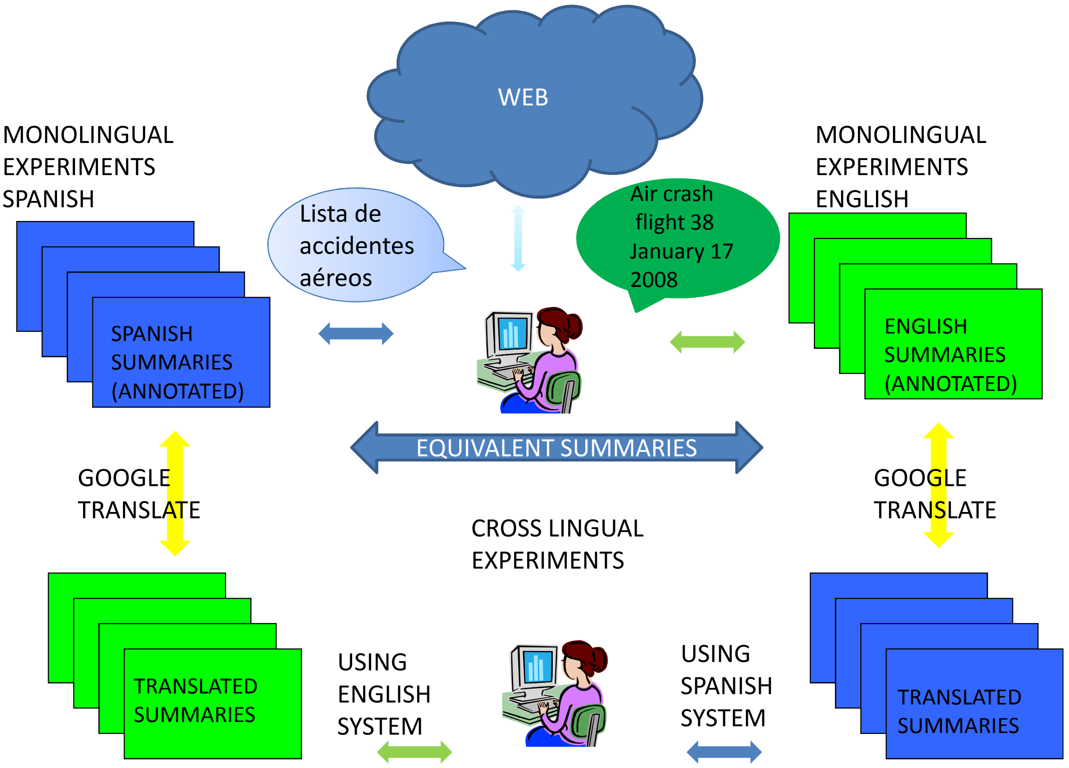

WEB
MONOLINGUAL EXPERIMENTS SPANISH
MONOLINGUAL EXPERIMENTS ENGLISH
Lista de accidentes aéreos
Air crash flight 38 January 17 2008
SPANISH SUMMARIES (ANNOTATED)
ENGLISH SUMMARIES (ANNOTATED)
EQUIVALENT SUMMARIES
GOOGLE TRANSLATE
GOOGLE TRANSLATE
CROSS LINGUAL EXPERIMENTS
TRANSLATED SUMMARIES
USING ENGLISH SYSTEM
USING SPANISH SYSTEM
TRANSLATED SUMMARIES

Figure 1: Corpus Creation Framework

English Terrorist Attack Template	
City:	Colombo
Country:	Sri Lanka
DateOfAttack:	January 31
Injured:	1,400
Target:	Central Bank
Perpetrator:	LTTE
TotalVictims:	90
TypeOfAttack:	suicide bomber

Table 1: Instantiated Template for a Terrorist Attack Event (English dataset)

5. Automatic Document Processing

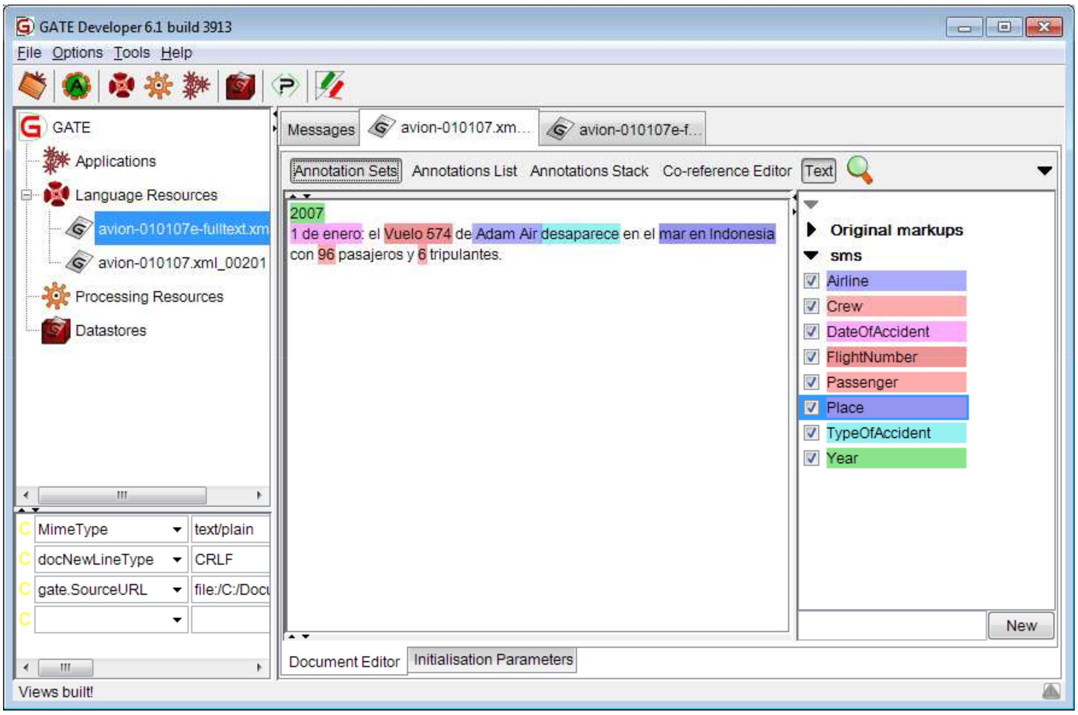
All summaries were analysed by automatic processes as described below:

5.1. English Text Analysis

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.

5.2. Spanish Text Analysis

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 (SVMs) trained over data from the CoNLL evaluation program (Sang, 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 package. The named entity



MimeType: text/plain
docNewLineType: CRLF
gate.SourceURL: file:/C:/Docu

2007
1 de enero: el Vuelo 574 de Adam Air desaparece en el mar en Indonesia
con 96 pasajeros y 6 tripulantes.

Figure 2: Aviation Accident Summary Annotated with Semantic Information

Spanish Terrorist Attack Template	
City:	Colombo
DateOfAttack:	enero de 1996
Injured:	centenares
Target:	Banco Central de Sri Lanka
Perpetrator:	Tigres de la Liberación Tamil Eelam
TotalVictims:	90
TypeOfAttack:	explotar un camion-bomba

Table 2: Instantiated Templates for a Terrorist Attack Event (Spanish dataset)

recogniser is based on SVMs classification (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.

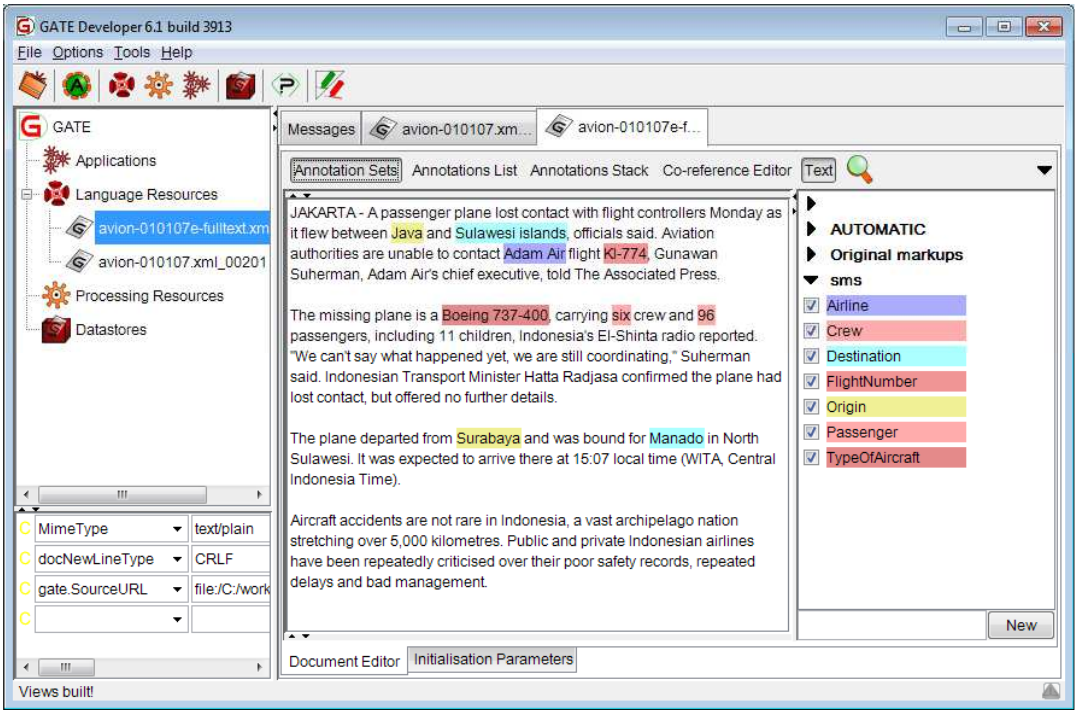
6. Uses of the Corpus

We believe this dataset is rich enough to carry out experiments in information extraction – identifying the key semantic elements of each event type – in various conditions, such as:

- training and testing in summaries;
- training in summaries in one language and testing in comparable full documents; and
- training in original summaries and testing in translations.

We have carried out experiments in each of the above scenarios and results have been reported elsewhere (Saggion and Szasz, 2011). Here, and to give an idea of the obtained performance, we describe monolingual IE experiments. We have developed domain independent information extraction systems for English and Spanish. The systems are again based on SVMs (Li et al., 2002) which are trained on the human and machine annotated summaries. The features used to represent the learning instances are: words, word orthography, lemmas, parts-of-speech tags, and named entity information. Windows of 5 words around each target token are used to represent the learning instances.

Because the CONCISUS dataset is rather small we have carried out 10-fold cross-validation experiments per domain and language, adjusting the parameters of the SVM to obtain an optimal system. Monolingual information extraction results in terms of standard precision, recall, and



MimeType	text/plain
docNewLineType	CRLF
gate.SourceURL	file:/C:/work

JAKARTA - A passenger plane lost contact with flight controllers Monday as it flew between Java and Sulawesi islands, officials said. Aviation authorities are unable to contact Adam Air flight KI-774, Gunawan Suherman, Adam Air's chief executive, told The Associated Press.

The missing plane is a Boeing 737-400, carrying six crew and 96 passengers, including 11 children, Indonesia's El-Shinta radio reported. "We can't say what happened yet, we are still coordinating," Suherman said. Indonesian Transport Minister Hatta Radjasa confirmed the plane had lost contact, but offered no further details.

The plane departed from Surabaya and was bound for Manado in North Sulawesi. It was expected to arrive there at 15:07 local time (WITA, Central Indonesia Time).

Aircraft accidents are not rare in Indonesia, a vast archipelago nation stretching over 5,000 kilometres. Public and private Indonesian airlines have been repeatedly criticised over their poor safety records, repeated delays and bad management.

Figure 3: Aviation Accident Full Report Annotated with Semantic Information

Event	Docs	Sents.	Words	Entities
Airplane Accident	32	2	48	8
Rail Accident	43	1.4	25	6
Earthquake	56	1.6	41	4.8
Terrorist Attack	53	2.2	52	6

Table 3: Spanish dataset statistics.

f-score (Piskorski and Yangarber, 2012) are presented in Table 5: these results aggregate all semantic types into a single figure. The results are still modest especially if compared with systems trained over larger datasets (i.e. over 400 documents (Li et al., 2005)) but they are somehow expected, not only because of the reduced amount of training data but also because of the uneven distribution of each semantic type in the dataset, with some types (such as the date of the event) being present in most summaries and others (such as "flight origin" and "flight destination" in the aviation accident domain) being present in just a few summaries.

7. Translations, Full Documents, and Man-Machine Annotation

In addition to summaries of events, and to provide infrastructure and support for cross-lingual IE experiments, we have also produced translation of each summary using Google Translate. For each summary in Spanish a translation into English is produced and for each English translation an automatic translation into Spanish is produced. These translations are also annotated with the annotation tool, but note that because of the noisy status of automatic translations not all information in the translations can be accurately identified. Examples of automatic translations of the summaries introduced above (see Section 1.) are:

> 2008 January 17: The British Airways Flight 38 (Boeing 777) crashed **at the airport to land at** London Heathrow from Beijing. There were no casualties
>
> 2008 17 de enero - British Airways Vuelo 38, un Boeing 777-200ER, **las tierras por debajo de la pista** en el aeropuerto Heathrow de Londres en el Reino Unido. Nueve de las 152 personas a bordo son tratados por lesiones menores, pero no hay víctimas mortales, **lo que es la primera pérdida de un Boeing 777**.

As it can be appreciated (in boldface in the examples) the translations can be considered noisy data because of the

Event	Docs	Sents.	Words	Entities
Airplane Accident	32	1.5	46	9
Rail Accident	36	1.3	30	7
Earthquake	44	2.8	71	7
Terrorist Attack	47	1.8	48	7

Table 4: English dataset statistics.

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
Terrorist Attack Spanish	0.64	0.48	0.55
Terrorist Attack English	0.61	0.50	0.54

Table 5: Mono-lingual Information Extraction Experiment Results (Summaries)

errors they contain.

For each reported event we also provide, whenever possible, a full event report in Spanish and English containig details of the event beyond the information of the summaries. These comparable full documents are very useful for experimentation in monolingual and cross-lingual TS.

The human annotation of translations and full documents is done in a human-computer collaborative way. The methodology is illustrated in the bottom part of Figure 1: the information extraction systems described in Section 6. are first applied to the automatic translations and the full documents, and then extraction results are corrected by a human annotator. We are currently evaluating the performance of the extraction from full documents given systems trained on summaries. In Table 6 we report preliminary results for extraction from full documents in the aviation domain: results are comparable to extraction from summaries.

8. Conclusions

Corpora and language resources for the adaptation of natural language processing systems are of paramount importance, especially with the current need to develop extraction and summarization technology to distill the increasing volume of online text. We believe this work contributes with a rich cross-lingual dataset to the study of cross-lingual information extraction and summarization. The corpus covers four application domains and two languages and contains monolingual comparable summaries in Spanish and English, summary translations, and full documents. The documents have been annotated by a human annotator following an annotation schema per application domain. Translations and full documents have been annotated with the help of an information extraction system trained on monoligual summaries. We have also carried out a set of machine learning experiments to show the value of the dataset. The results are still modest but they should be assessed considering the limited syntactic and semantic infomation used in the experiments. Our current work involves the expansion of the dataset to cover additional domains, languages, and data-points. Our future work will focus on semi-automatic domain modelling for information extraction and summarization. The corpus is made available to the research community through the CONCISUS Web page at `http://www.taln.upf.edu/pages/concisus/`.

Acknowledgments

We would like to thank the reviewers for their comments and suggestions which have helped improve the final version of this paper. We are grateful to the Advanced Research Fellowship RYC-2009-04291 from Programa Ramón y Cajal 2009, Ministerio de Economía y Competitividad, Secretaría de Estado de Investigación, Desarrollo e Innovación, Spain. We thank Biljana Drndarevic who helped us prepare the final version of the paper.

9. References

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

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

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

Andreas Eisele and Yu Chen. 2010. MultiUN: A Multilingual Corpus from United Nation Documents. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may. European Language Resources Association (ELRA).

Event	Prec	Rec	F
Aviation Accident Spanish	0.68	0.47	0.56
Aviation Accident English	0.53	0.48	0.50

Table 6: Mono-lingual Information Extraction Experiment Results (Full Documents)

R. Gaizauskas, K. Humphreys, S. Azzam, and Y. Wilks. 1997. Concepticons *vs.* lexicons: An architecture for multilingual information extraction. In M.T. Pazienza, editor, *Proceedings of the Summer School on Information Extraction (SCIE-97)*, LNCS/LNAI, pages 28–43. Springer-Verlag.

P. Gamallo Otero and I. González López. 2010. Wikipedia as multilingual source of comparable corpora. In *Proceedings of the 3rd Workshop on Building and Using Comparable Corpora, LREC 2010*, Malta, 22 May 2010.

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

D. Hakkani-Tür, Heng Ji, and R. Grishman. 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*.

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

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

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

Y. Li, K. Bontcheva, and H. Cunningham. 2005. Using Uneven Margins SVM and Perceptron for Information Extraction. In *Proceedings of Ninth Conference on Computational Natural Language Learning (CoNLL-2005)*.

D. Maynard and H. Cunningham. 2003. Multilingual Adaptations of a Reusable Information Extraction Tool. In *Proceedings of the Demo Sessions of EACL'03*, Budapest, Hungary. ACL.

D. Maynard, V. Tablan, H. Cunningham, C. Ursu, H. Saggion, K. Bontcheva, and 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.

P. Over, H. Dang, and D. Harman. 2007. DUC in context. *Inf. Process. Manage.*, 43:1506–1520, November.

K. Owczarzak and H.T. Dang. 2010. Overview of the tac 2010 summarization track. In *Proceedings of TAC 2010*. NIST.

J. Piskorski and R. Yangarber. 2012. Information extraction: Past, present, and future. In T. Poibeau, H. Saggion, J. Piskorski, and R. Yangarber, editors, *Multi-source, Multilingual Information Extraction and Summarization*, Theory and Applications of Natural Language Processing. Springer.

T. Poibeau, H. Saggion, J. Piskorski, and R. Yangarber, editors. 2012. *Multi-source, Multilingual Information Extraction and Summarization*. Theory and Applications of Natural Language Processing. Springer.

Dragomir R. Radev, Simone Teufel, Horacio Saggion, Wai Lam, John Blitzer, Hong Qi, Arda Çelebi, Danyu Liu, and Elliott Drábek. 2003. Evaluation Challenges in Large-Scale Document Summarization. In *ACL*, pages 375–382.

H. Saggion and T. Poibeau. 2012. Automatic text summarization: Past, present, and future. In T. Poibeau, H. Saggion, J. Piskorski, and R. Yangarber, editors, *Multi-source, Multilingual Information Extraction and Summarization*, Theory and Applications of Natural Language Processing. Springer.

H. Saggion and Sandra Szasz. 2011. Multi-domain cross-lingual information extraction from clean and noisy texts. In *8th Brazilian Symposium in Information and Human Language Technology*, Cuiabá, Brazil, 10/2011. SBC, SBC.

H. Saggion, D. Radev, S. Teufel, L. Wai, and S. Strassel. 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.

H. Saggion. 2006. Multilingual Multidocument Summarization Tools and Evaluation. In *Proceedings of LREC 2006*.

Erik F. Tjong Kim Sang. 2002. Introduction to the conll-2002 shared task: Language-independent named entity recognition. *CoRR*.

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

E Steinberger, B Pouliquen, A. Widiger, C. Ignat, T. Erjavec, and D. Tufis. 2006. The jrc-acquis: A multilingual aligned parallel corpus with 20+ languages. In *In Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'2006*, pages 2142–2147.