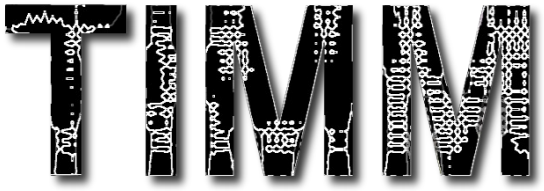
III CATÁLOGO DE RECURSOS EN TECNOLOGÍAS DEL LENGUAJE HUMANO



V Jornadas TIMM

Tratamiento de la Información Multilingüe y Multimodal

12 y 13 de junio de 2014

Cazalla de la Sierra, Sevilla

GOBIERNO DE ESPAÑA
MINISTERIO DE ECONOMÍA Y COMPETITIVIDAD
SECRETARÍA DE ESTADO DE INVESTIGACIÓN, DESARROLLO E INNOVACIÓN

Financiada por el MINECO

III Catálogo de Recursos en Tecnologías del Lenguaje Humano

Editador por L. Alfonso Ureña

Presentación

La Red Temática TIMM (Tratamiento de Información Multilingüe y Multimodal), con referencia TIN2011-13070-E, dentro del programa de acciones complementarias ha facilitado la elaboración del tercer catálogo en Tecnologías del Lenguaje Humano.

Este catálogo se presentará en el marco de las V Jornadas TIMM, en Cazalla de la Sierra, Sevilla.

El objetivo general de las jornadas es promover la difusión de las actividades de investigación, desarrollo e innovación entre los diferentes grupos de investigación de ámbito nacional en el ámbito del Tratamiento de Información Multilingüe y Multimodal. Concretamente se persiguen los siguientes objetivos:

- Crear un foro donde los investigadores en formación puedan presentar y discutir su trabajo en un ambiente que facilite el intercambio de ideas y la colaboración.

- Organización de un seminario y una mesa redonda con el objetivo de hacer una puesta en común para conocer en qué estado se encuentra cada grupo participante y hacia dónde se dirige con el fin de que los grupos puedan interactuar y reutilizar los recursos de cada uno. Concretamente un seminario sobre sistemas de recomendación y minería de opiniones y una mesa redonda sobre proyectos.

- Difusión de los resultados científicos y tecnológicos mediante trabajos presentados.

- Realizar un catálogo de recursos lingüísticos y herramientas desarrolladas en los diferentes grupos de investigación para fomentar su uso y difusión entre otros grupos.

En esta memoria se incluyen las fichas del catálogo.

Agradecer a todos los investigadores miembros que han contribuido con sus fichas a la compilación y actualización del presente catálogo. Asimismo, agradecer a la Red Temática TIMM, en cuyo marco se organiza por quinta vez estas jornadas.

Este catálogo ha sido cofinanciado por la Red Temática (TIN2011-13070-E) del Ministerio de Economía y Competitividad y por el Fondo Europeo de Desarrollo Regional (FEDER).

L. Alfonso Ureña

Índice

1. Córpora, Bases de Datos y otros Recursos Lingüísticos

ADQA (Arabic Definition Question Answering) corpus

Authors: Omar Trigui and Lamia Hadrich Belguith / University of Sfax (Tunisia), Paolo Rosso / NLE Lab. ELiRF

References: http://users.dsic.upv.es/grupos/nle/?file=kop4.php, https://sites.google.com/site/anlprg/outils-et-corpus-realises

Description: ADQA Corpus - Arabic Definition Question Answering corpus. This corpus is constituted of a list of 50 definition questions (ArabicListDefQuest), a set of 50 files containing snippets collected from Wikipedia search engine (ArabicCorpusWikipedia), a set of 50 files containing snippets from Google search engine (ArabicCorpusGoogle) and a set of 50 files which each file contains a question with their answers (ArabicListDefAnsw from -Google+Wikipedia-).

Technical Requirements: No special hardware/software is required. Disk space required: 235 Kbytes.

Modules: No.

Innovation: A first corpus corpus collected from the web and a set of definition questions with their answers. Development: MICINN research project TEXT-ENTERPRISE 2.0 TIN2009-13391-C04-03 (Plan I+D+i). This corpus was generated as part of the Ph.D. work of Omar Trigui under the supervision of Lamia Hadrich Belguith and Paolo Rosso.

Development: MICINN research project TEXT-ENTERPRISE 2.0 TIN2009-13391-C04-03 (Plan I+D+i).

This corpus was generated as part of the Ph.D. work of Diego Ingaramo under the supervision of Marcelo Errecalde (external researcher of TEXT-ENTERPRISE 2.0) and Paolo Rosso.

Publications:

- Trigui O., Hadrich-Belguith L., Rosso P. An Automatic Definition Extraction in Arabic Language. In: Proc. 15th Int. Conf. on Applications of Natural Language to Information Systems, NLDB-2010, Springer-Verlag, LNCS(6177), pp. 240-247, 2010
- Trigui O., Hadrich-Belguith L., Rosso P. DefArabicQA: Arabic Definition Question Answering System. In: Proc. Workshop on LR & HLT for Semitic Languages, 7th Int. Conf. on Language Resources and Evaluation, LREC-2010, Malta, May 17-23, pp. 40-44, 2010

Contact: Omar Trigui omar.trigui@gmail.com (Paolo Rosso prosso@dsic.upv.es)

Amazon Data Sets

Authors: Antonio Reyes / NLE Lab. ELiRF

References: http://users.dsic.upv.es/grupos/nle/?file=kop4.php

Description: This corpus has been created in order to study the figurative language, especially irony, sarcasm and humour, in a context focused on sentiment analysis. It contains approx. 5,000 comments.

Functionality: It allows carrying out experiments on **Irony Detection**.

Innovation: No public corpora are available for irony detection.

Development: MICINN research project TEXT-ENTERPRISE 2.0 TIN2009-13391-C04-03 (Plan I+D+i). Developed as part of the Ph.D. Thesis of Antonio Reyes (writing-up phase).

Contact: Paolo Rosso (prosso@dsic.upv.es)

ANCORA-CA

Authors: CLiC-UB (Centre de Llenguatge i Computació -Universitat de Barcelona: M.Antònia Martí, Mariona Taulé, Lluís Màrquez and Manuel Bertran.

References: http://clic.ub.edu/ancora

Description: AnCora-Ca is a multilevel annotated corpus of Catalan, consisting of 500,000 words mostly from newspaper articles. AnCora-Ca is annotated with morphological (PoS), syntactic (constituents and functions) and semantic (argument structure and thematic roles, semantic class, named entities and WordNet senses) information. All resulting layers are independent of each other, thus making easier the data management. The annotation was performed manually, semiautomatically, or fully automatically, depending on the encoded linguistic information.

Functionality: Annotated corpora constitute a crucial resource to acquire or infer linguistic knowledge about how languages are used. In this line, AnCora-Ca is a very useful resource for computational and linguistic analysis of language, especially necessary for machine learning systems. This corpus is used as source of information for developing POS taggers, syntactic parsers and, Semantic Role Labelling, Word Sense Disambiguation, Named Entity Recognition and Classification systems. This corpus was used in the *SemEval 2007 task*: Multilevel Semantic Annotation of Catalan and Spanish

Technology: Data stored in XML format

Technical Requirements: -

Modules: -

Innovation: At present AnCora-Ca is the largest Catalan corpus annotated at all the linguistic levels described above freely available

Development: The development of AnCora-Ca has been funded by the following projects: 3LB (FIT-150-500-2002-244), CESS-ECE (HUM2004-21127), PRAXEM (HUM2006-27378-E), and Lang2World (TIN2006-15265-C06-06) from the Spanish Ministry of Education and Science, and the funding given by the Catalan Secretary of Linguistic Policy.

Publications:

- Taulé, M., M.A. Martí, M. Recasens (2008) Ancora: Multilevel Annotated Corpora for Catalan and Spanish. Proceedings of 6th International Conference on Language Resources and Evaluation. Marrakesh (Morocco).

Contact: M. Antònia Martí <amarti@ub.edu>

ANCORA-CO-CA

Authors: CLiC-UB (Centre de Llenguatge i Computació -Universitat de Barcelona: M.Antònia Martí, Mariona Taulé, Marta Recasens, Lluís Màrquez and Manuel Bertran

References: http://clic.ub.edu/ancora

Description: AnCora-CO-Ca is a subset of the multilevel annotated corpus AnCora-Ca (for Catalan), consisting of 400,000 words, enriched with coreference information, where all noun phrases (NPs) –pronominal or with a nominal head– pointing to the same entity are linked.

Functionality: AnCora-CO-Ca can be a useful resource for training and evaluating coreference resolution systems for Catalan. From a linguistic point of view, the annotated corpus can be used as a workbench to test and validated hypotheses on coreferential expressions for Catalan. This corpus will be used in SemEval 2010 coreference resolution task: http://stel.ub.edu/semeval2010-coref

Technology: Data stored in XML format

Technical Requirements: -

Modules: -

Innovation: At present AnCora-CO-Ca is the largest Catalan corpus annotated with coreference and freely available.

Development: The development of AnCora-CO-Es has been funded by the following projects: PRAXEM (HUM2006-27378-E) and Lang2World (TIN2006-15265-C06-06) from the Spanish Ministry of Education and Science.

Publications:

- Recasens, M., M.A.Martí, M. Taulé (2008) First-mention Definites: More than Exceptional Cases, S. Featherson & S. Winkler (eds), Fruits: Process and Product in Empirical Linguistics. Berlin: de Gruyter.
- Recasens, M. (2008) Towards Coreference Resolution for Catalan and Spanish. Master Thesis. Universitat de Barcelona.
- Recasens, M., M. A. Martí i M. Taulé (2007) 'Where Anaphora and Coreference Meet. Annotation in the CESS-ECE Corpus'. Recent Advances in Natural language Processing. Borovets, Bulgaria

Contact: Mariona Taulé <mtaule@ub.edu>

ANCORA-CO-ES

Authors: CLiC-UB (Centre de Llenguatge i Computació -Universitat de Barcelona: M.Antònia Martí, Mariona Taulé, Marta Recasens, Lluís Màrquez and Manuel Bertran

References: http://clic.ub.edu/ancora

Description: AnCora-CO-Es is a subset of the multilevel annotated corpus AnCora-Es (for Spanish), consisting of 400,000 words, enriched with coreference information, where all noun phrases (NPs) –pronominal or with a nominal head– pointing to the same entity are linked.

Functionality: AnCora-CO-Es can be a useful resource for training and evaluating coreference resolution systems for Spanish. From a linguistic point of view, the annotated corpus can be used as a workbench to test and validated hypotheses on coreferential expressions for Spanish. This corpus will be used in SemEval 2010 coreference resolution task: http://stel.ub.edu/semeval2010-coref

Technology: Data stored in XML format

Technical Requirements: -

Modules: -

Innovation: At present AnCora-CO-Es is the largest Spanish corpus annotated with coreference and freely available.

Development: The development of AnCora-CO-Es has been funded by the following projects: PRAXEM (HUM2006-27378-E) and Lang2World (TIN2006-15265-C06-06) from the Spanish Ministry of Education and Science.

Publications:

- Recasens, M., M.A.Martí, M. Taulé (2008) First-mention Definites: More than Exceptional Cases, S. Featherson & S. Winkler (eds), Fruits: Process and Product in Empirical Linguistics. Berlin: de Gruyter.
- Recasens, M. (2008) Towards Coreference Resolution for Catalan and Spanish. Master Thesis. Universitat de Barcelona.
- Recasens, M., M. A. Martí i M. Taulé (2007) *'Where Anaphora and Coreference Meet. Annotation in the CESS-ECE Corpus'*. Recent Advances in Natural language Processing. Borovets, Bulgaria

Contact: Mariona Taulé <mtaule@ub.edu>

ANCORA-DEP-CA

Authors: CLiC-UB (Centre de Llenguatge i Computació -Universitat de Barcelona: M.Antònia Martí, Mariona Taulé, Lluís Màrquez and Manuel Bertran

References: http://clic.ub.edu/ancora

Description: AnCora-DEP-Ca is the AnCora-Ca multilevel annotated corpus of Catalan in dependency-based representation, consisting of 500,000 words approximately.

Functionality: AnCora-DEP-Es can be used as source of information for for inducing grammars, developing, improving and/or evaluating syntactic parsers and algorithms for semantic role labelling, dependency-based. This corpus is used in the *CoNLL Shared Task 2009: Syntactic and Semantic Dependencies in Multiple Languages*, where the core of the task is to predict syntactic and semantic dependencies and their labelling.

Technology: Data stored in XML format

Technical Requirements: -

Modules: -

Innovation: At present AnCora-DEP-Ca is the largest corpus multilevel annotated available in dependency format freely downloaded.

Development: The development of AnCora-DEP-Ca has been funded by the following projects: CESS-ECE (HUM2004-21127) and Lang2World (TIN2006-15265-C06-06, and the funding given by the Catalan Secretary of Linguistic Policy.

Publications:

- Civit, M., M.A. Martí & N. Bufí (2006) 'Cat3LB and Cast3LB: from Constituents to dependencies', Springer Verlag, *Advances in Natural Language Processing* (LNAI, 4139), pp. 141-153. Berlin, ISSN: 0302-9743.

Contact: M. Antònia Martí <amarti@ub.edu>

ANCORA-DEP-ES

Authors: CLiC-UB (Centre de Llenguatge i Computació -Universitat de Barcelona: M.Antònia Martí, Mariona Taulé, Lluís Màrquez and Manuel Bertran

References: http://clic.ub.edu/ancora

Description: AnCora-DEP-Es is the AnCora-Es multilevel annotated corpus of Spanish in dependency-based representation, consisting of 500,000 words approximately.

Functionality: AnCora-DEP-Es can be used as source of information for inducing grammars, developing, improving and/or evaluating syntactic parsers and algorithms for semantic role labelling, dependency-based. This corpus is used in the *CoNLL Shared Task 2009: Syntactic and Semantic Dependencies in Multiple Languages*, where the core of the task is to predict syntactic and semantic dependencies and their labelling.

Technology: Data stored in XML format

Technical Requirements: -

Modules: -

Innovation: At present AnCora-DEP-Es is the largest corpus multilevel annotated available in dependency format freely downloaded.

Development: The development of AnCora-DEP-Es has been funded by the following projects: CESS-ECE (HUM2004-21127) and Lang2World (TIN2006-15265-C06-06).

Publications:

- Civit, M., M.A. Martí & N. Bufí (2006) 'Cat3LB and Cast3LB: from Constituents to dependencies', Springer Verlag, *Advances in Natural Language Processing* (LNAI, 4139), pp. 141-153. Berlin, ISSN: 0302-9743.

Contact: M. Antònia Martí <amarti@ub.edu>

ANCORA-ES

Authors: CLiC-UB (Centre de Llenguatge i Computació -Universitat de Barcelona: M.Antònia Martí, Mariona Taulé, Lluís Màrquez and Manuel Bertran.

References: http://clic.ub.edu/ancora

Description: AnCora-Es is a multilevel annotated corpus of Spanish, consisting of 500,000 words mostly from newspaper articles. AnCora-Es is annotated with morphological (PoS), syntactic (constituents and functions) and semantic (argument structure and thematic roles, semantic class, named entities and WordNet senses) information. All resulting layers are independent of each other, thus making easier the data management. The annotation was performed manually, semiautomatically, or fully automatically, depending on the encoded linguistic information.

Functionality: Annotated corpora constitute a crucial resource to acquire or infer linguistic knowledge about how languages are used. In this line, AnCora-Es is a very useful resource for computational and linguistic analysis of language, especially necessary for machine learning systems. This corpus is used as source of information for developing POS taggers, syntactic parsers and, Semantic Role Labelling, Word Sense Disambiguation, Named Entity Recognition and Classification systems. This corpus was used in the *SemEval 2007 task*: Multilevel Semantic Annotation of Catalan and Spanish.

Technology: Data stored in XML format

Technical Requirements: -

Modules: -

Innovation: At present AnCora-Es is the largest Spanish corpus annotated at all the linguistic levels described above freely available.

Development: The development of AnCora-Es has been funded by the following projects: 3LB (FIT-150-500-2002-244), CESS-ECE (HUM2004-21127), PRAXEM (HUM2006-27378-E), and Lang2World (TIN2006-15265-C06-06) from the Spanish Ministry of Education and Science.

Publications:

- Taulé, M., M.A. Martí, M. Recasens (2008) Ancora: Multilevel Annotated Corpora for Catalan and Spanish. Proceedings of 6th International Conference on Language Resources and Evaluation. Marrakesh (Morocco).

Contact: M. Antònia Martí <amarti@ub.edu>

ANCORA-Verb-CA

Authors: CLiC-UB (Centre de Llenguatge i Computació -Universitat de Barcelona: M.Antònia Martí, Mariona Taulé, Marta Recasens, Aina Peris, Lluís Màrquez and Manuel Bertran

References: http://clic.ub.edu/ancora

Description: AnCora-Verb-Ca is a verbal lexicon containing 2,141diferent verbs. In AnCora-Verb-Ca lexicon, the mapping between syntactic functions, arguments and thematic roles of each verbal predicate it is established taking into account the verbal semantic class and the diatheses alternations in which the predicate can participate. Each verbal predicate may be divided in different senses where each sense is related to one or more semantic classes (Lexical Semantic Structures), basically differentiated according to the four event classes -accomplishments, achievements, states and activities-, and on the diatheses alternations in which a sense can occur.

Functionality: AnCora-Verb-Ca is the verbal lexicon used as the basis for the semantic annotation with arguments and thematic roles of AnCora-Ca corpus. This lexical resource can be used for syntactic and semantic parsing, lexical representation, etc. This lexicon was used in the *SemEval 2007 task*: Multilevel Semantic Annotation of Catalan and Spanish

Technology: -

Technical Requirements: -

Modules: -

Innovation: At present AnCora-Verb-Ca is the largest verbal lexicon containing information about verbal semantic classes, syntactic subcategorization, argument structure and the corresponding thematic roles, as well as diatheses alternations, freely available.

Development: The development of AnCora-Verb-Ca has been funded by the following projects: PRAXEM (HUM2006-27378-E), and Lang2World (TIN2006-15265-C06-06) from the Spanish Ministry of Education and Science, and the funding given by the Catalan Secretary of Linguistic Policy.

Publications:

- Aparicio, J., M. Taulé, M.A. Martí (2008) AnCora-Verb: A Lexical Resource for the Semantic Annotation of Corpora. Proceedings of 6th International Conference on Language Resources and Evaluation. Marrakesh (Morocco).
- Aparicio, J., M. Taulé, M.A. Martí (2008) AnCora-Verb: Two large-scale lexicons for Catalan and Spanish. Bernal, E. Decesaris, J. (eds.) Proceedings of the XIII Euralex International Congress 2008. Institut Universitari de Lingüística Aplicada, UPF: Barcelona (Spain).

Contact: Mariona Taulé <mtaule@ub.edu>

ANCORA-Verb-ES

Authors: CLiC-UB (Centre de Llenguatge i Computació -Universitat de Barcelona: M.Antònia Martí, Mariona Taulé, Marta Recasens, Aina Peris, Lluís Màrquez and Manuel Bertran

References: http://clic.ub.edu/ancora

Description: AnCora-Verb-Es is a verbal lexicon containing 2,603 different verbs. In AnCora-Verb-Es lexicon, the mapping between syntactic functions, arguments and thematic roles of each verbal predicate it is established taking into account the verbal semantic class and the diatheses alternations in which the predicate can participate. Each verbal predicate may be divided in different senses where each sense is related to one or more semantic classes (Lexical Semantic Structures), basically differentiated according to the four event classes -accomplishments, achievements, states and activities-, and on the diatheses alternations in which a sense can occur.

Functionality: AnCora-Verb-Es is the verbal lexicon used as the basis for the semantic annotation with arguments and thematic roles of AnCora-Es corpus. This lexical resource can be used for syntactic and semantic parsing, lexical representation, etc. This lexicon was used in the *SemEval 2007 task*: Multilevel Semantic Annotation of Catalan and Spanish

Technology: -

Technical Requirements: -

Modules: -

Innovation: At present AnCora-Verb-Es is the largest verbal lexicon containing information about verbal semantic classes, syntactic subcategorization, argument structure and the corresponding thematic roles, as well as diatheses alternations, freely available

Development: The development of AnCora-Verb-Ca has been funded by the following projects: PRAXEM (HUM2006-27378-E) and Lang2World (TIN2006-15265-C06-06) from the Spanish Ministry of Education and Science.

Publications:

- Aparicio, J., M. Taulé, M.A. Martí (2008) AnCora-Verb: A Lexical Resource for the Semantic Annotation of Corpora. Proceedings of 6th International Conference on Language Resources and Evaluation. Marrakesh (Morocco).
- Aparicio, J., M. Taulé, M.A. Martí (2008) AnCora-Verb: Two large-scale lexicons for Catalan and Spanish. Bernal, E. Decesaris, J. (eds.) Proceedings of the XIII Euralex International Congress 2008. Institut Universitari de Lingüística Aplicada, UPF: Barcelona (Spain).

Contact: Mariona Taulé <mtaule@ub.edu>

ANERcorp

Authors: Yassine Benajiba (Ph.D. student) and Paolo Rosso

References: http://www.dsic.upv.es/grupos/nle/

Description: ANERcorp is an Arabic NER corpus which consists of 150,000 tokens (which go up to 200,000 tokens after segmentation).

Functionality: IOB annotated Arabic NER resource.

Technology: The corpus was annotated by one person to ensure annotation coherence. Each named entity is tagged by its class using the IOB annotation scheme following the guidelines of the corpora used in the CoNLL 2002 and 2003 evaluation campaigns.

Technical Requirements: None

Modules: -

Innovation: To our knowledge, it is the only freely available Arabic NER corpus.

Development: Developed as part of Yassine Benajiba's AECI Ph.D. and the MiDES CICYT TIN2006-15265-C06-04 research project, co-funded by the AECI-PCI A01031707 and A706706 projects.

Publications:

- Benajiba Y., Rosso P. Arabic Named Entity Recognition using Conditional Random Fields. In: Proc. Workshop on HLT & NLP within the Arabic world. Arabic Language and local languages processing: Status Updates and Prospects, 6th Int. Conf. on Language Resources and Evaluation, LREC-2008, Marrakech, Morocco, May 26-31, 2008

- Benajiba Y., Diab M. Rosso P. Arabic Named Entity Recognition: An SVM-based approach. In: Proc. Int. Arab Conf. on Information Technology, ACIT-2008, Hammamet, Tunisia, December, 2008.

- Benajiba Y., Rosso P., Benedí J.M. ANERsys: An Arabic Named Entity Recognition system based on Maximum Entropy. In: Proc. 8th Int. Conf. on Comput. Linguistics and Intelligent Text Processing, CICLing-2007, Springer-Verlag, LNCS(4394), pp. 143-153, 2008.

Contact: Yassine Benajiba <benajibayassine@gmail.com>

ANERgazet

Authors: Yassine Benajiba (Ph.D. student) and Paolo Rosso

References: http://www.dsic.upv.es/grupos/nle/

Description: ANERgazet is a set of 3 Arabic gazetteers (people, locations and organizations) which might be used mainly for the Arabic NER task, but still can be used for other Arabic NLP tasks.

Functionality: Each gazetteer contains a list of Arabic names belonging to the concerned class.

Technology: The gazetteers were extracted automatically from Arabic Wikipedia and the Web resources and then manually filtered.

Technical Requirements: None

Modules: -

Innovation: To our knowledge, ANERgazet is the only Arabic gazetteers which are freely available to the research community.

Development: Developed as part of Yassine Benajiba's AECI Ph.D. and the MiDES CICYT TIN2006-15265-C06-04 research project, co-funded by the AECI-PCI A01031707 and A706706 projects.

Publications:

- Benajiba Y., Diab M., Rosso P. Arabic Named Entity Recognition using Optimized Feature Sets. In: Proc. Int. Conf. on Empirical Methods in Natural Language Processing, EMNLP-2008, Waikiki, Honolulu, U.S.A., October, 2008

- Benajiba Y., Rosso P. Arabic Named Entity Recognition using Conditional Random Fields. In: Proc. Workshop on HLT & NLP within the Arabic world. Arabic Language and local languages processing: Status Updates and Prospects, 6th Int. Conf. on Language Resources and Evaluation, LREC-2008, Marrakech, Morocco, May 26-31, 2008

- Benajiba Y., Diab M. Rosso P. Arabic Named Entity Recognition: An SVM-based approach. In: Proc. Int. Arab Conf. on Information Technology, ACIT-2008, Hammamet, Tunisia, December, 2008

- Benajiba Y., Rosso P., Benedí J.M. ANERsys: An Arabic Named Entity Recognition system based on Maximum Entropy. In: Proc. 8th Int. Conf. on Comput. Linguistics and Intelligent Text Processing, CICLing-2007, Springer-Verlag, LNCS(4394), pp. 143-153, 2008

Contact: Yassine Benajiba <benajibayassine@gmail.com>

Arabic QA

Authors: Yassine Benajiba and Lahsen Abouenour (Ph.D. students) and Paolo Rosso

References: http://www.dsic.upv.es/grupos/nle/

Description: This corpus includes Spanish journalistic texts, more precisely, it is a collection of news extracted from El Periódico de Catalunya. It has been manually annotated at a syntactic (phrases and syntactic function) and semantic level (semantic roles, semantic constructions and sense disambiguation). The corpus has approximately 700.000 words. It contains sentences with the 250 more frequent verbs in Spanish.

Functionality: The interface (http://grial.uab.es/search) allows simple and advanced searches on the corpus by different fields, including the negative search. The XML corpus can be downloaded.

Technology: Documents: from the web, Questions and answers: manually built.

Technical Requirements: In order to use these data, it is required to have an Arabic Question system. It is also possible to use these data to test only some modules of the QA system (I.e. the QA system is not required to be fully built).

Modules: None

Innovation: To our knowledge, these data is the only freely available Arabic test-platform for Arabic Question Answering.

Development: Developed as part of Yassine Benajiba's AECI Ph.D. and the MiDES CICYT TIN2006-15265-C06-04 research project, co-funded by the AECI-PCI A01031707 and A706706 projects.

Publications:

- Benajiba Y., Rosso P., Gómez J.M. Adapting JIRS Passage Retrieval System to the Arabic. In: Proc. 8th Int. Conf. on Comput. Linguistics and Intelligent Text Processing, CICLing-2007, Springer-Verlag, LNCS(4394), pp. 530-541, 2007

- Abouenour L., Bouzoubaa K., Rosso P. Improving Q/A using Arabic WordNet. In: Proc. Int. Arab Conf. on Information Technology, ACIT-2008, Hammamet, Tunisia, December, 2008

- Abouenour L., Bouzoubaa K., Rosso P. Construction de l'ontologie Amine Arabic WordNet dans le cadre des systèmes Q/A. (in French) In: Proc. 2nd JOurnées Scientifiques en Technologies de l'Information et de la Communication JOSTIC-2008, Rabat, Marroco, October, 2008

- Abouenour L., Bouzoubaa K., Rosso P. Towards an Arabic Q/A system using a conceptual/lexical ontology. (in Arabic) In: Proc. Proc. 5th Conf. on Scientific Research Outlook & Technology Development in the Arab world, SROV, Fez, Marroco, October, 11-16, 2008

Contact: Yassine Benajiba <benajibayassine@gmail.com>

Arabic WordNet

Authors: Musa Alkhalifa, Manuel Bertran, William J. Black, Sabri Elkateb, Javier Farreres, David Farwell, Christiane Fellbaum, James Kirk, Ma Antònia Martí, Adam Pease, Horacio Rodríguez, Piek Vossen

References: http://www.globalwordnet.org/AWN

Description: The Arabic WordNet (AWN) is a lexical database of the Arabic language following the development process of Princeton English WordNet and Euro WordNet. It utilizes the Suggested Upper Merged Ontology as an interlingua to link Arabic WordNet to previously developed wordnets. Christiane Fellbaum at Princeton was the project lead. The project was sponsored by DOI/REFLEX.

From http://www.globalwordnet.org/AWN/DataSpec.html you can get the XML data exchange specifications of the database. Roughly AWN contains 11,000 synsets (including 1,000 NE). Figures of the current coverage of AWN can be obtained from: http://www.lsi.upc.edu/~mbertran/arabic/awn/query/sug_statistics.php

There are several different ways for accessing the database: 1) The browser package (available at http://sourceforge.net/projects/awnbrowser) includes the AWN data and PWN2.0 mappings in a relational database. You can use the export facilities to export the data as XML or CSV and do what you will with it in textual form. 2) The browser source code contains methods written in Java which access the DB in order to retrieve the data structures the browser displays. The AWN code has been placed on Sourceforge at http://awnbrowser.cvs.sourceforge.net/awnbrowser/source/awnbrowser. There's 2-page document about the source code there too. 3) The database can also be downloaded in xml format (linked to pwn 2.0) from http://www.lsi.upc.edu/~mbertran/arabic/bd/get_bd.php. If you are interested on data connected with other English WN versions you can get it from http://www.lsi.upc.edu/~mbertran/arabic/bd/get_bd.php?ver={20,21,30,awn}. 4) A set of basic python functions for accessing the database can be obtained from: http://www.lsi.upc.edu/~horacio/varios/AWNDatabaseManagement.py.gz

In addition to the AWN database and the AWN browser some other complementary software has been produced including: 1) The AWN Lexicographer's Web Interface (Barcelona) http://www.lsi.upc.edu/~mbertran/arabic/awn/update/synset_browse.php . 2) The AWN User's Web Interface

(Barcelona) http://www.lsi.upc.edu/~mbertran/arabic/awn/index.html . 3) The AWN word spotter can be accessed at: http://www.lsi.upc.edu/~mbertran/arabic/wwwWn7/

Functionality: AWN Browser: Browsing the database . AWN can be downloaded in XML format and access its content be directly accessed.

Technology: Java, Perl, MySQL

Technical Requirements: -

Modules: -

Innovation: One of the most important lexical resources for Arabic language.

Development: -

Publications:

- Christiane Fellbaum, Musa Alkhalifa, William J. Black, Sabri Elkateb, Adam Pease, Horacio Rodríguez, Piek Vossen (2006). Introducing the Arabic WordNet project. Proceedings of the 3rd Global Wordnet Conference, Jeju Island, Korea, January, 2006.

- Christiane Fellbaum, Musa Alkhalifa, William J. Black, Sabri Elkateb, Adam Pease, Horacio Rodríguez, Piek Vossen (2006). Building a WordNet for Arabic. Proceedings of the the 5th Conference on Language Resources and Evaluation LREC2006, May, 2006.

- Horacio Rodríguez, David Farwell, Javi Farreres, Manuel Bertran, Musa Alkhalifa, M. Antonia Martí, William Black, Sabri Elkateb, James Kirk, Adam Pease, Piek Vossen, and Christiane Fellbaum. Arabic WordNet: Current State and Future Extensions in: Proceedings of the Fourth International GlobalWordNet Conference - GWC 2008, Szeged, Hungary, January 22-25, 2008.

- Horacio Rodríguez, David Farwell, Javi Farreres, Manuel Bertran, Musa Alkhalifa, M. Antonia Martí (2008). Arabic WordNet: Semi-automatic Extensions using Bayesian Inference. Proceedings of the the 6th Conference on Language Resources and Evaluation LREC2008. Marrakech (Morocco), May 2008.

Contact: Horacio Rodrígez <horacio@lsi.upc.edu>

Author Profiling @ PAN-2013

Authors: Francisco Rangel, Paolo Rosso, Giacomo Inches, Moshe Koppel y Efstathios Stamatatos

References:http://www.webis.de/research/corpora/corpus-pan-labs-09-today/pan-13/pan13-data/pan13-author-profiling-training-corpus-2013-01-09.zip

Description: This corpus consists of ocuments written in both English and Spanish. With regard to age, we will consider posts of three classes: 10s (13-17), 20s (23-27), and 30s (33-47). Moreover, documents from authors who pretend to be minors will be included (e.g., documents composed of chat lines of sexual predators will be also considered).

Functionality: -

Technology: -

Technical Requirements: -

Modules: -

Innovation: -

Development: -

Publications: -

Contact: http://www.uni-weimar.de/medien/webis/research/events/pan-13/pan13-web/author-profiling.html

Author Profiling @ PAN-2014

Authors: Francisco Rangel, Paolo Rosso, Giacomo Inches, Benno Stein, Martin Potthast, Walter Daelemans, Efstathios Stamatatos, Fabio Crestani

References:http://www.uni-weimar.de/medien/webis/research/events/pan-14/pan14-web/author-profiling.html

Description: Twitter tweets and social media texts written in both English and Spanish as well as hotel reviews written in English. With regard to age, we will consider the following classes: 18-24, 25-34, 35-49, 50-64, 65-xx.

Functionality: -

Technology: -

Technical Requirements: -

Modules: -

Innovation: -

Development: -

Publications: -

Contact: -

Blogs Analysis corpus

Authors: Antonio Reyes / NLE Lab. ELiRF

References: http://users.dsic.upv.es/grupos/nle/?file=kop4.php

Description: The corpus is integrated by 8 sets. Every set contains 2,400 documents automatically retrieved from LiveJournal and Wikipedia. The corpus is organised as follows: i) The [mfs] versions contain the documents labelled with POS tags and the mosf frequent sense according to WordNet. ii) The [xml] versions

contain the sets converted into the Senseval-2 formatted XML. The corpus has been designed for analysing humour features in the Blogosphere.

Functionality: It allows carrying out experiments on Automatic Humour Recognition.

Innovation: Documents retrieved from LiveJournal and Wikipedia for **Automatic Humour Recognition**

Development: MICINN research project TEXT-ENTERPRISE 2.0 TIN2009-13391-C04-03 (Plan I+D+i).

Developed as part of the Ph.D. Thesis of Antonio Reyes (writing-up phase).

Publications:

- Reyes A., Rosso P., Buscaldi D. Finding Humour in the Blogosphere: The Role of WordNet Resources. In: Proc. 5th Global WordNet Int. Conf., GWN-2010, Bombay, India, January 31-February 4, 2010
- Reyes A., Rosso P., Buscaldi D. Affect-based Features for Humour Recognition. In: Proc. 7th Int. Conf. on Natural Language Processing, ICON-2009, Hyderabad, India, December 15-17, pp. 364-369, 2009
- Reyes A., Rosso P. Linking Humour to Blogs Analysis: Affective Traits in Posts. In: Proc. 1st Workshop on on Opinion Mining and Sentiment Analysis (WOMSA), CAEPIA-TTIA Conference, Seville, Spain, November 13, pp. 205-212, 2009

Contact: Paolo Rosso (prosso@dsic.upv.es)

Blogs Clustering Corpus

Authors: Daniel Pérez (M.Sc. student) and David Pinto

References: http://www.dsic.upv.es/grupos/nle/downloads.html

Description: This is a set of corpora made up of discussion lines extracted from two blogs websites: boing-boing and slashdot.

Functionality: The aim of this corpus is to support experiments of supervised and unsupervised classifiers with narrow domain short texts, especifically in the medicine field, with documents related with the “cancer” topic.

Technology: The corpus (raw-text blogs) and the gold standard are provided. The discussion lines are intended as categories or classes of the gold standard, whereas posts are the target documents.

Technical Requirements: No special requirements are needed in order to use the corpus.

Innovation: The aim of this corpus is to manually verifty the results of different classifiers on the blogs clustering task.

Development: Developed as part of David Pinto Ph.D. and the MiDES CICYT TIN2006-15265-C06-04 research project.

Contact: David Eduardo Pinto Avendaño <dpinto@cs.buap.mx>

CESCA

Authors: Liliana Tolchinsky, M.Antònia Martí, Mariona Taulé (CliC-UB)

References: http://clic.ub.edu/cesca

Description: CESCA is a Catalan corpus consisting of scholar writing text elaborated by 2,400 scholars between the ages of five and sixteen. Each informant has written different types of text: vocabularies, narrative and definition texts as well as jokes.

Functionality: This corpus allows studies about language development; literacy; relationship between oral text and spelling; linguistic analysis of spontaneous language, etc.

Technology: -

Technical Requirements: -

Modules: -

Innovation: It does not exist an another corpus of this size and characteristics for Catalan. It allows studies about evolution of language taking into account the same parameters among a huge amount of population.

Development: The development of CESCA has been funded by the following projects: 2006ARIE-10058, 2007ARIE-00005, 2008ARIE-00053 from the Generalitat de Catalunya (AGAUR).

Publications: -

Contact: M. Antònia Martí <amarti@ub.edu>

CICLing-2002 Clustering Corpus

Authors: Mikhail Alexandrov and Alexander Gelbukh (Instituto Politécnico Nacional, Mexico). Pre-processed by David Pinto; Héctor Jiménez (Universidad Autónoma Metropolitana, México)

References: http://www.dsic.upv.es/grupos/nle/downloads.html

Description: This a pre-processed version of 48 scientific abstracts from the CICLing 2002 conference (computational linguistics).

Functionality: The aim of this corpus is to support experiments of supervised and unsupervised classifiers with narrow domain short texts.

Technology: The corpus (raw text) and the gold standard are provided.

Technical Requirements: No special requirements are needed in order to use the corpus.

Innovation: A very small collection which may be used to manually verify the results obtained in the clustering task of narrow domain short texts.

Development: Developed as part of David Pinto Ph.D. and the MiDES CICYT TIN2006-15265-C06-04 research project.

Publications:

- David Pinto, Alfons Juan, Paolo Rosso: A Comparative Study of Clustering Algorithms on Narrow-Domain Abstracts. Procesamiento del Lenguaje Natural 37(1): 43-49, 2006
- Héctor Jiménez-Salazar, David Pinto, Paolo Rosso: Uso del Punto de Transición en la Selección de Términos Índice para Agrupamiento de Textos Cortos, Procesamiento del Lenguaje Natural 35(1): 114-118, 2005
- Diego Ingaramo, David Pinto, Paolo Rosso, Marcelo Errecalde: Evaluation of Internal Validity Measures in Short-Text Corpora. CICLing 2008. Lecture Notes in Computer Science 4919, Springer-Verlag: 555-567, 2008
- David Pinto, Paolo Rosso: On the Relative Hardness of Clustering Corpora. TSD 2007. Lecture Notes in Artificial Intelligence 4629, Springer-Verlag: 155-161, 2007
- David Pinto, José-Miguel Benedí, Paolo Rosso: Clustering Narrow-Domain Short Texts by Using the Kullback-Leibler Distance. CICLing 2007. Lecture Notes in Computer Science 4394, Springer-Verlag: 611-622, 2007
- David Pinto, Héctor Jiménez-Salazar, Paolo Rosso: Clustering Abstracts of Scientific Texts Using the Transition Point Technique. CICLing 2006. Lecture Notes in Computer Science 3878, Springer-Verlag:536-546, 2006

Contact: David Eduardo Pinto Avendaño <dpinto@cs.buap.mx>

CL!NSS PAN@FIRE corpus

Authors: Parth Gupta, Paolo Rosso, Paul Clough, Mark Stevenson, Rafael E. Banchs

References: http://users.dsic.upv.es/grupos/nle/clinss.html

Description: The corpus contains source (Hindi) and target (English) news stories partition. The documents are marked up with news story metadata such as title, date of publication and content.

Functionality: -

Technology: -

Technical Requirements: -

Modules: -

Innovation: -

Development: -

Publications: -

Contact: -

CL!TR corpus

Authors: Alberto Barrón-Cedeño, Paolo Rosso, Sobha Lalitha Devi, Paul Clough, Mark Stevenson

References: http://users.dsic.upv.es/grupos/nle/fire-workshop-clitr.html

Description: The corpus contains a set of potential source documents D, written in English, and set of suspicious documents S, written in Hindi. In the corpus you will find plain text files encoded in UTF-8. The source documents are taken from English Wikipedia. The source documents include Wiki-mark up.

Functionality: -

Technology: -

Technical Requirements: -

Modules: -

Innovation: -

Development: -

Publications: -

Contact: -

CLPD

Authors: Alberto Barrón-Cedeño, Parth Gupta, Paolo Rosso

References: http://www.dsic.upv.es/grupos/nle/resources/clpd-data.tar.gz

Description: Este corpus ha sido utilizado en los experimentos de detección de plagio translingüe realizados en esta publicación : Barrón-Cedeño A., Gupta P., Rosso P. Methods for Cross-Language Plagiarism Detection . In: Knowledge-Based Systems, vol. 50, pp. 11-17 DOI: 10.1016/j.knosys.2013.06.018 http://dx.doi.org/10.1016/j.knosys.2013.06.018

Functionality: -

Technology: -

Technical Requirements: -

Modules: -

Innovation: -

Development: -

Publications: -

Contact: -

Co-derivatives corpus

Authors: Alberto Barrón-Cedeño / NLE Lab. ELiRF

References: http://users.dsic.upv.es/grupos/nle/resources/abc/download-coderiv.html

Description: This corpus has been generated for the analysis of co-derivatives, text reuse and plagiarism (of course, simulated). It is composed of more than 20,000 documents from Wikipedia in German, English, Hindi and Spanish (around 5,000 documents per language). For each language, some of the most frequently consulted articles in Wikipedia have been considered as pivot and ten of its revisions were downloaded, which compose the set of co-derivatives. The corpus has three versions: (*i*) original (articles without further manipulation); (*ii*) clean (articles after case folding and punctuation marks elimination); and (*iii*) stopwords free (articles after case folding and punctuation marks and stopwords elimination).

Functionality: It allows carrying out experiments on co-derivatives and text similarity analysis in the following languages: German, English, Hindi and Spanish

Innovation: The publicly available corpus for **co-derivatives and text similarity analysis** in German, English, Hindi and Spanish

Development: MICINN research project TEXT-ENTERPRISE 2.0 TIN2009-13391-C04-03 (Plan I+D+i).

Developed as part of the Ph.D. Thesis of Alberto Barrón-Cedeño (writing-up phase).

Publication:

- Barrón-Cedeño A., Eiselt A., Rosso P. Monolingual Text Similarity Measures: A Comparison of Models over Wikipedia Articles Revisions. In: Proc. 7th Int. Conf. on Natural Language Processing, ICON-2009, Hyderabad, India, December 15-17, pp. 29-38, 2009

Contact: Paolo Rosso (prosso@dsic.upv.es)

Colección HEP

Authors: Montejo-Ráez, A. and Steinberger, R. and Ureña-López, L. A.

References:

Description: Este corpus está orientado al estudio de clasificadores de texto multi-etiquetado. Está compuesto por artículos científicos en el área de la Física de Altas Energías (HEP – High Energy Physics) obtenidos del servidor de documentos CDS del Laboratorio de Física Nuclear Europeo (CERN). El corpus está dividido en tres subconjuntos (denominadas particiones), donde cada partición se compone, a su vez, de dos ficheros: uno que contiene los registros de cada artículo (con información como los abstract, los autores y, por supuesto, las clases o palabras clave) en formato XML comprimido, y otro que contiene una versión en texto plano del artículo completo generado a partir del PDF disponible en las bases de datos del CERN (en formato tar + gzip) Las clases están delimitadas por la marca XML KEYWORD. Estas son las etiquetas

del tesauro de DESY asignadas manualmente. Puede obtener más información sobre el tesauro de DESY. Descarga: Versión 2.1 del corpus HEP

- Partición hepth: 18,114 documentos de Física Teórica (metadatos - 5,3 Mb) (artículos - 226 Mb)
- Partición hepex: 2,599 documentos de Física Experimental (metadatos - 1,6 Mb) (artículos - 28 Mb)
- Partición astroph: 2,716 documentos de Astrofísica (metadatos - 1,1 Mb) (artículos - 29 Mb)

Actualizado el 23.04.2007: Gracias a Ioannis Katakis, de la Aristotle University of Thessaloniki, (Grecia) por corregir algunos problemas en el XML proporcionado.

Functionality: -

Technology: -

Technical Requirements: -

Modules: -

Innovation: -

Development: -

Publications:

- Montejo-Ráez, A. and Steinberger, R. and Ureña-López, L. A. Adaptive selection of base classifiers in one-against-all learning for large multi-labeled collections. Advances in Natural Language Processing: 4th International Conference, EsTAL 2004

Contact: Montejo-Ráez, A. <amontejo@ujaen.es> and Ureña-López, L. A. <laurena@ujaen.es>

Computer Science Tri-lingual Corpus

Authors: Grzegorz Chrumpala, Ana Fernández, Elisabet Comelles, Marta Coll-Florit, Glòria Vàzquez, Nerea Achutegui , Irene Castellón, Mercè Coll, Marta Prim

References: Grup de Recerca Interuniversitari d'Aplicacions Lingüístiques: http://grial.uab.es

Description: This corpus was developed as part of a project for teaching innovation whose objective was the improvement of the processes for teaching/learning technical English by using a new and different perspective from the traditional language class, by means of the creation of a parallel corpus. This corpus is a collection of texts available in different languages (English-Catalan-Spanish), which includes approximately 2.257.498 words and exemplifies the usage of the language within a technical register, more specifically, in the computer science domain.

This corpus is very useful for classes since it can be used as a dynamic resource in the teaching, both for the teacher and for the student. For the teacher, it can be used as a source of material for the creation of exercises, texts for the classes and examples. It can also be used as a guide to develop the class syllabus. As the corpus is annotated morphologically and syntactically, this resource is transformed into a very useful instrument for the language class, because it allows us to perform searches not only of collocations but also of words with a given category, or even of lemmas.

Functionality: The interface allows the simple search (form), and the advanced search (form, lemma and POS) and the output is showed in the three languages.

Technology: Wrapper program (Java class) which uses external annotation software (FreeLing and Connexor) to perform annotation of the aligned text while preserving the alignement. It then formats and stores annotated segments in a specialized CPG XML format. eXist database which stores the XML-encoded text. XQuery scripts implementing specialized search functions. There are two basic search modes: *quicksearch*, for searching by sequences of word forms, and *fullsearch* for searching by sequences of token specifications including the token form, lemma and morphological annotation. The Web User Interface, in whose implementation use is made of all the major web techologies: XQuery, XSLT, CSS, XHTML and Ecmascript. The MVC Cocoon web application framework bundled with eXist is used to implement this component.

Technical Requirements: -

Modules: -

Innovation: Using computational linguistics techniques for the production of materials for second language learning.

Development: The trilingual corpus has been developed in the project: Millora de la qualitat docent de la Generalitat de Catalunya (194 MQD 2002).

Publications:

- Castellón, I., A. Fernández, G. Vázquez (2005) "Creación de un recurso textual para el aprendizaje del inglés", *NOVATICA Revista de la Asociación de técnicos de informática*, 177, p. 51-54. ISSN: 0211-2124

Contact: A. M. Fernández Montraveta <ana.fernandez@uab.es>

Corpus EasyAbstracts

Authors: Diego Ingaramo and Marcelo Errecalde / Universidad Nacional de San Luis (Argentina), Paolo Rosso / NLE Lab. ELiRF

References:http://users.dsic.upv.es/grupos/nle/?file=kop4.php, https://sites.google.com/site/merrecalde/resources

Description: Corpus EasyAbstracts. This collection can be considered harder than collections of long documents such as Micro4News because its documents are scientific abstracts (same characteristic as CiCling-2002) and, as a consequence, are short documents. It differs from the CiCling-2002 collection with respect to the overlapping degree of the documents' vocabulary. EasyAbstracts documents also refer to a shared thematic (intelligent systems) but its groups are not so closely related as the CiCling-2002 ones are. EasyAbstracts was constructed with abstracts publicly available on Internet that correspond to articles of four international journals in the following fields: 1) Machine Learning, 2) Heuristics in Optimization, 3) Automated reasoning and 4) Autonomous intelligent agents. It is possible to select abstracts for these disciplines in a way that two abstracts of two different categories are not related at all. However, some degree of complexity can be introduced if abstracts of articles related to two or more EasyAbstracts' categories are used. EasyAbstracts includes a few documents with these last features in order to increase the complexity with respect to the Micro4News corpus. Nevertheless, a majority of documents in this collection clearly belong to a single group. This last fact allows us to assume that this collection has a lower complexity

degree than the CiCling2002 corpus used in different works on short-text clustering. Features of EasyAbstracts: Number of groups = 4, Number of documents = 48, number of terms = 9261, vocabulary size = 2169, (average) number of terms per document = 192.93.

Functionality: This corpus is intended to be used in supervised or unsupervised categorization tasks which mainly involve working with short length texts. The idea in this case was to provide a collection with scientific abstracts but with a lower complexity degree than the CiCling2002 corpus.

Technology: The development of this corpus did not require any special development tool. All the documents in this collection were manually selected.

Technical Requirements: No special hardware/software is required. Disk space required: 62 Kbytes.

Modules: No.

Innovation: This corpus allows to work with a small short-text collection that contains (as CiCling2002 does) scientific abstracts. However, this collection can be considered as a collection with a lower complexity degree than the CiCling2002 corpus.

Development: MICINN research project TEXT-ENTERPRISE 2.0 TIN2009-13391-C04-03 (Plan I+D+i).

This corpus was generated as part of the Ph.D. work of Diego Ingaramo under the supervision of Marcelo Errecalde (external researcher of TEXT-ENTERPRISE 2.0) and Paolo Rosso.

Publications:

- Ingaramo D., Errecalde M., Rosso P.A general bio-inspired method to improve the short-text clustering task. In: Proc. 10th Int. Conf. on Comput. Linguistics and Intelligent Text Processing, CICLing-2010, Springer-Verlag, LNCS(6008), pp. 661-672, 2010
- Errecalde M., Ingaramo D., Rosso P. ITSA*: An Effective Iterative Method for Short-Text Clustering Tasks. In: Proc. 23rd Int. Conf. on Industrial, Engineering & Other Applications of Applied Intelligent Systems , IEA-AIE-2010, Springer-Verlag, LNAI(6096), pp. 550-559, 2010
- Ingaramo D., Cagnina L., Errecalde M., Rosso P. A Particle Swarm Optimizer to cluster short-text corpora: a performance study. In: Proc. Workshop on Natural Language Processing and Web-based Technologies, 12th edition of the Ibero-American Conference on Artificial Intelligence, IBERAMIA-2010, Bahía Blanca, Argentina, November 1-5, pp. 71-79, 2010
- Ingaramo D., Rosas M.V., Errecalde M., Rosso P. Clustering Iterativo de Textos Cortos con Representaciones basadas en Conceptos. In: Proc. Workshop on Natural Language Processing and Web-based Technologies, 12th edition of the Ibero-American Conference on Artificial Intelligence, IBERAMIA-2010, Bahía Blanca, Argentina, November 1-5, pp. 80-89, 2010
- Errecalde M., Ingaramo D., Rosso P. Proximity estimation and the hardness of short-text corpora. In: 5th Workshop on Text-based Information Retrieval, TIR-2008, In: Proc. Database and Expert Systems Applications, DEXA-2008, IEEE Press, Turin, Italy, September 1-5, pp. 15-19, 2008
- Cagnina L., Errecalde M., Ingaramo D., Rosso P. A discrete particle Swarm optimizer for clustering short-text corpora. In: Bioinspired Optimization Methods and their Applications, BIOMA-2008, Ljubljana, Slovenia, October 13-14, pp. 93-103, 2008

Contact: Marcelo Errecalde merreca@unsl.edu.ar (Paolo Rosso prosso@dsic.upv.es)

Corpus Micro4News

Authors: Diego Ingaramo and Marcelo Errecalde / Universidad Nacional de San Luis (Argentina), Paolo Rosso / NLE Lab. ELiRF

References: http://users.dsic.upv.es/grupos/nle/?file=kop4.php

https://sites.google.com/site/merrecalde/resources

Description: Corpus Micro4News. The Micro4News collection was constructed with medium-length documents that correspond to four very different topics of the popular 20Newsgroups corpus: 1) sci.med, 2) soc.religion.christian, 3) rec.autos and 4) comp.os.ms-windows.misc. For each topic, the largest documents in the corresponding group were selected. Thus, the length of the selected documents was, on average, seven times (or more) the length of the abstracts of corpora such as EasyAbstracts and CICLing-2002 which were usually used in comparative studies with Micro4News. Features of Micro4News: Number of groups = 4, Number of documents = 48, number of terms = 125614, vocabulary size = 12785, (average) number of terms per document = 2616.95.

Functionality: This corpus is intended to be used in supervised or unsupervised categorization tasks which mainly involve working with short length texts. The idea in this case was to provide a low complexity small collection with well differentiated categories and relatively long documents. It has been used in comparative studies with high complexity small collections which consist of short texts (for example, the CICLing-2002 collection of scientific abstracts).

Technology: The development of this corpus did not require any special development tool. All the documents in this collection were manually selected.

Technical Requirements: No special hardware/software is required. Disk space required: 706 Kbytes.

Modules: No.

Innovation: This corpus allows to work with a small collection that should not be difficult to clustering or categorization purposes. Clustering or categorization algorithms should not have any problem in obtain high quality results with Micro4News. In that way, Micro4News becomes an interesting alternative as baseline collection in comparative studies that involve difficult short-text collections.

Development: MICINN research project TEXT-ENTERPRISE 2.0 TIN2009-13391-C04-03 (Plan I+D+i). This corpus was generated as part of the Ph.D. work of Diego Ingaramo under the supervision of Marcelo Errecalde (external researcher of TEXT-ENTERPRISE 2.0) and Paolo Rosso.

Publications:

- Ingaramo D., Errecalde M., Rosso P.A general bio-inspired method to improve the short-text clustering task. In: Proc. 10th Int. Conf. on Comput. Linguistics and Intelligent Text Processing, CICLing-2010, Springer-Verlag, LNCS(6008), pp. 661-672, 2010
- Errecalde M., Ingaramo D., Rosso P. ITSA*: An Effective Iterative Method for Short-Text Clustering Tasks. In: Proc. 23rd Int. Conf. on Industrial, Engineering & Other Applications of Applied Intelligent Systems , IEA-AIE-2010, Springer-Verlag, LNAI(6096), pp. 550-559, 2010
- Ingaramo D., Cagnina L., Errecalde M., Rosso P. A Particle Swarm Optimizer to cluster short-text corpora: a performance study. In: Proc. Workshop on Natural Language Processing and Web-based Technologies, 12th edition of the Ibero-American Conference on Artificial Intelligence, IBERAMIA-2010, Bahía Blanca, Argentina, November 1-5, pp. 71-79, 2010

- Ingaramo D., Rosas M.V., Errecalde M., Rosso P. Clustering Iterativo de Textos Cortos con Representaciones basadas en Conceptos. In: Proc. Workshop on Natural Language Processing and Web-based Technologies, 12th edition of the Ibero-American Conference on Artificial Intelligence, IBERAMIA-2010, Bahía Blanca, Argentina, November 1-5, pp. 80-89, 2010
- Errecalde M., Ingaramo D., Rosso P. Proximity estimation and the hardness of short-text corpora. In: 5th Workshop on Text-based Information Retrieval, TIR-2008, In: Proc. Database and Expert Systems Applications, DEXA-2008, IEEE Press, Turin, Italy, September 1-5, pp. 15-19, 2008
- Cagnina L., Errecalde M., Ingaramo D., Rosso P. A discrete particle Swarm optimizer for clustering short-text corpora. In: Bioinspired Optimization Methods and their Applications, BIOMA-2008, Ljubljana, Slovenia, October 13-14, pp. 93-103, 2008

Contact: Marcelo Errecalde merreca@unsl.edu.ar (Paolo Rosso prosso@dsic.upv.es)

Corpus Plagiarism Competition PAN-PC-2010

Authors: Alberto Barrón-Cedeño / NLE Lab. ELiRF + rest of organisers of the PAN competition on plagiarism detection: http://pan.webis.de/

References: http://users.dsic.upv.es/grupos/nle/resources/abc/download-panpc10.html

Description: This corpus contains documents in which artificial plagiarism has been inserted automatically: 8.4 GB, 162,000 cases of plagiarism

Functionality: It allows carrying out experiments on plagiarism detection.

Innovation: The only publicly available corpus **for plagiarism detection.**

Development: This is the corpus developed for the 2nd competition on plagiarism detection. MICINN research project TEXT-ENTERPRISE 2.0 TIN2009-13391-C04-03 (Plan I+D+i). Developed as part of the Ph.D. Thesis of Alberto Barrón-Cedeño (writing-up phase).

Publication:

- Potthast M., Barrón-Cedeño A., Stein B., Rosso P. An Evaluation Framework for Plagiarism Detection. In: Proc. of the 23rd International Conference on Computational Linguistics, COLING-2010, Beijing, China, August 23-27, 2010

Contact: Paolo Rosso (prosso@dsic.upv.es)

Corpus R8+

Authors: Diego Ingaramo and Marcelo Errecalde / Universidad Nacional de San Luis (Argentina), Paolo Rosso / NLE Lab. EliRF

References: http://users.dsic.upv.es/grupos/nle/?file=kop4.php

https://sites.google.com/site/merrecalde/resources

Description: Corpus R8+. Subset of documents of the R8-Test corpus, a sub-collection of the well-known Reuters-21578 dataset. R8+ has the same number of groups as R8-Test (eight groups), but they differ in the number of documents per group. Each group of R8+ only contains the largest documents in the corresponding group of R8-Test (a 20% of the documents of each original group). Features of R8+: Number of groups = 8, Number of documents = 445, number of terms = 66314, vocabulary size = 7797, (average) number of terms per document = 149.02.

Functionality: This corpus is intended to be used in supervised or unsupervised categorization tasks which mainly involve working with short length texts. However, R8+ only contains the largest documents in R8-Test in order to analyze how difficult a collection with these particularities is, with respect to collections with "very short" length documents like R8-, a collection similar to R8+, but generated with the shortest documents of R8-Test. Both collections were simultaneously generated in previous studies to consider the "shortest length" and the "largest length" versions of R8-Test.

Technology: The development of this corpus did not require any special development tool beyond the very simple routines to select a 20% of the largest documents in each R8-Test's group.

Technical Requirements: No special hardware/software is required. Disk space required: 440 Kbytes.

Modules: No.

Innovation: Unlike R8-Test, which contains relatively short documents, this corpus allows to focus on the particularities that present working with the largest documents of this collection. This study, is usually complemented with results obtained with R8-, which was generated with the shortest documents of R8-Test.

Development: MICINN research project TEXT-ENTERPRISE 2.0 TIN2009-13391-C04-03 (Plan I+D+i). This corpus was generated as part of the Ph.D. work of Diego Ingaramo under the supervision of Marcelo Errecalde (external researcher of TEXT-ENTERPRISE 2.0) and Paolo Rosso.

Publications:

- Ingaramo D., Errecalde M., Rosso P.A general bio-inspired method to improve the short-text clustering task. In: Proc. 10th Int. Conf. on Comput. Linguistics and Intelligent Text Processing, CICLing-2010, Springer-Verlag, LNCS(6008), pp. 661-672, 2010
- Errecalde M., Ingaramo D., Rosso P. ITSA*: An Effective Iterative Method for Short-Text Clustering Tasks. In: Proc. 23rd Int. Conf. on Industrial, Engineering & Other Applications of Applied Intelligent Systems , IEA-AIE-2010, Springer-Verlag, LNAI(6096), pp. 550-559, 2010
- Rosas M., Errecalde M., Rosso P. Un Análisis Comparativo de Estrategias para la Categorización Semántica de Textos Cortos. In: Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN), num. 44, pp. 11-18, 2010
- Ingaramo D., Rosas M.V., Errecalde M., Rosso P. Clustering Iterativo de Textos Cortos con Representaciones basadas en Conceptos. In: Proc. Workshop on Natural Language Processing and Web-based Technologies, 12th edition of the Ibero-American Conference on Artificial Intelligence, IBERAMIA-2010, Bahía Blanca, Argentina, November 1-5, pp. 80-89, 2010

Contact: Marcelo Errecalde merreca@unsl.edu.ar (Paolo Rosso prosso@dsic.upv.es)

Corpus R8-

Authors: Diego Ingaramo and Marcelo Errecalde / Universidad Nacional de San Luis (Argentina), Paolo Rosso / NLE Lab. ELiRF

References: http://users.dsic.upv.es/grupos/nle/?file=kop4.php

https://sites.google.com/site/merrecalde/resources

Description: Corpus R8-. Subset of documents of the R8-Test corpus, a sub-collection of the well-known Reuters-21578 dataset. R8- has the same number of groups as R8-Test (eight groups), but they differ in the number of documents per group. Each group of R8- only contains the shortest documents in the corresponding group of R8-Test (a 20% of the documents of each original group). Features of R8-: Number of groups = 8, Number of documents = 445, number of terms = 8481, vocabulary size = 1876, (average) number of terms per document = 19.06.

Functionality: This corpus is intended to be used in supervised or unsupervised categorization tasks which mainly involve working with short length texts. In particular, this collection has been used in studies related to the difficulties that collections with short documents (like R8-) present to clustering algorithms, with respect to arbitrary-size document collections.

Technology: The development of this corpus did not require any special development tool beyond the very simple routines to select a 20% of the shortest documents in each R8-Test's group.

Technical Requirements: No special hardware/software is required. Disk space required: 44.3 Kbytes.

Modules: No.

Innovation: Unlike R8-Test, which contains relatively short documents, this corpus allows to focus on the particularities that present working with extremely short documents.

Development: MICINN research project TEXT-ENTERPRISE 2.0 TIN2009-13391-C04-03 (Plan I+D+i).

This corpus was generated as part of the Ph.D. work of Diego Ingaramo under the supervision of Marcelo Errecalde (external researcher of TEXT-ENTERPRISE 2.0) and Paolo Rosso.

Publications:

- Ingaramo D., Errecalde M., Rosso P.A general bio-inspired method to improve the short-text clustering task. In: Proc. 10th Int. Conf. on Comput. Linguistics and Intelligent Text Processing, CICLing-2010, Springer-Verlag, LNCS(6008), pp. 661-672, 2010
- Errecalde M., Ingaramo D., Rosso P. ITSA*: An Effective Iterative Method for Short-Text Clustering Tasks. In: Proc. 23rd Int. Conf. on Industrial, Engineering & Other Applications of Applied Intelligent Systems , IEA-AIE-2010, Springer-Verlag, LNAI(6096), pp. 550-559, 2010
- Rosas M., Errecalde M., Rosso P. Un Análisis Comparativo de Estrategias para la Categorización Semántica de Textos Cortos. In: Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN), num. 44, pp. 11-18, 2010
- Ingaramo D., Rosas M.V., Errecalde M., Rosso P. Clustering Iterativo de Textos Cortos con Representaciones basadas en Conceptos. In: Proc. Workshop on Natural Language Processing and Web-based Technologies, 12th edition of the Ibero-American Conference on Artificial Intelligence, IBERAMIA-2010, Bahía Blanca, Argentina, November 1-5, pp. 80-89, 2010

Contact: Marcelo Errecalde merreca@unsl.edu.ar (Paolo Rosso prosso@dsic.upv.es)

Corpus R8B

Authors: Diego Ingaramo and Marcelo Errecalde / Universidad Nacional de San Luis (Argentina), Paolo Rosso / NLE Lab. ELiRF

References: http://users.dsic.upv.es/grupos/nle/?file=kop4.php, https://sites.google.com/site/merrecalde/resources

Description: Corpus R8B. Subset of documents of the R8-Test corpus, a sub-collection of the well-known Reuters-21578 dataset. R8B has the same number of groups as R8-Test (eight groups), but they differ in the number of documents in two specific groups. R8B can be considered as a "balanced" version of R8-Test, with respect to the number of documents per group. Two of the eight groups of R8-Test, contains almost 70% of all the documents in the collection. R8B on the other hand, is intended to provide a collection as similar to R8-Test as possible but fixing this imbalance produced by these two "big" groups. In order to obtain a more balanced collection, those groups were reduced in size by removing a specific number of documents and obtaining in that way a collection without the differences in the size of groups that R8-Test exhibited. Features of R8B: Number of groups = 8, Number of documents = 816, number of terms = 71842, vocabulary size = 5854, (average) number of terms per document = 88.04.

Functionality: This corpus is intended to be used in supervised or unsupervised categorization tasks which mainly involve working with short length texts. The idea in this case was to provide a more balanced variant of R8-Test without the differences in size that two of its groups presented.

Technology: The development of this corpus did not require any special development tool beyond the very simple routines to reduce the size of the two biggest groups in R8-Test.

Technical Requirements: No special hardware/software is required. Disk space required: 415 Kbytes.

Modules: No.

Innovation: Unlike R8-Test, an unbalanced document collection, this corpus allows to work with a short-text collection similar to R8-Test but with groups of comparable size.

Development: MICINN research project TEXT-ENTERPRISE 2.0 TIN2009-13391-C04-03 (Plan I+D+i).

This corpus was generated as part of the Ph.D. work of Diego Ingaramo under the supervision of Marcelo Errecalde (external researcher of TEXT-ENTERPRISE 2.0) and Paolo Rosso.

Publications:

- Ingaramo D., Cagnina L., Errecalde M., Rosso P. A Particle Swarm Optimizer to cluster short-text corpora: a performance study. In: Proc. Workshop on Natural Language Processing and Web-based Technologies, 12th edition of the Ibero-American Conference on Artificial Intelligence, IBERAMIA-2010, Bahía Blanca, Argentina, November 1-5, pp. 71-79, 2010

Contact: Marcelo Errecalde merreca@unsl.edu.ar (Paolo Rosso prosso@dsic.upv.es)

Cross-Lingual Plagiarism Corpus

Authors: Luis Alberto Barrón Cedeño (Ph.D. student) and Paolo Rosso

References: http://www.dsic.upv.es/grupos/nle/downloads.html

Description: The CliPA corpus has been created as a resource for the design and test of methods for the automatic detection of cross-lingual plagiarism cases. It contains a set of original text fragments in English and around twelve different plagiarised versions of them in Spanish (Italian will be added soon). The plagiarised text fragments were obtained by both "human plagiarisers" and Machine Translators. In order to create a realistic plagiarism detection environment, the corpus includes a set of text fragments on the same topic but originally written in Spanish.

Functionality: Due to the facts that all the text fragments in the corpus are identified as original or plagiarised and that the plagiarised fragments are linked to their actual source, the corpus can be used to develop and test cross-lingual plagiarism detection methods.

Technology: The corpus is codified in XML.

Technical Requirements: There are not special requirements to use the corpus. It can be accessed via any XML parser.

Innovation: To our knowledge, With respect, CLiPA corpus is the only freely available corpus cross-lingual plagiarism analysis.

Development: Developed as part of the MiDES CICYT TIN2006-15265-C06-04 research project.

Publications:

- Barrón-Cedeño A., Rosso, P., Pinto, D. and Juan, A. On cross-lingual plagiarism analysis using a statistical model. In: Proceedings of the ECAI'08 PAN Workshop: Uncovering Plagiarism, Authorship and Social Software Misuse, pp. 9-13. Patras, Greece, 2008.

- Pinto D., Civera J., Juan A., Rosso P., Barrón-Cedeño A. A statistical approach to crosslingual natural language tasks. In: Proc. 4th Latin American Workshop on Non-Monotonic Reasoning, LANMR-2008, Puebla, Mexico, October 22-24, 2008

- Pinto D., Civera J., Barrón-Cedeño A., Juan A., Rosso P. A statistical approach to crosslingual natural language tasks (selected and enhanced version; accepted and to be published). In: Journal of Algorithms in Cognition, Informatics and Logic, 2009

Contact: Luis Alberto Barrón Cedeño <lbarron@dsic.upv.es>

DDI corpus

Authors: Isabel Segura-Bedmar, Paloma Martínez, María Herrero Zazo

References: http://labda.inf.uc3m.es/doku.php?id=en:labda_ddicorpus

Description: DDI corpus: An annotated corpus with pharmacological substances and drug-drug interactions.

The management of drug-drug interactions (DDIs) is a critical issue resulting from the overwhelming amount of information available on them. Natural Language Processing (NLP) techniques can providean interesting way to reduce the time spent by healthcare professionals on reviewing biomedical literature. However, the shortage of annotated corpora for DDI extraction is the main bottleneck in the development of NLP systems for this area of Pharmacovigilance. So precisely for this reason, we are pleased to announce that the DDI corpus, an annotated corpus with pharmacological substances and drug-drug interactions (DDIs), is now available at http://labda.inf.uc3m.es/ddicorpus.

The DDI corpus is made up of 792 texts selected from the DrugBank database and other 233 Medline abstracts on the subject of DDIs. The corpus was annotated with a total of 18,502 pharmacological substances and 5028 DDIs, including both pharmacokinetic (PK) as well as pharmacodynamic (PD) interactions. To date, the corpora annotated with DDIs have focused in PK DDIs, but not in PD DDIs.

Functionality: The DDI corpus was developed for the SemEval 2013-DDIExtraction 2013 task (http://www.cs.york.ac.uk/semeval-2013/task9/), whose main goal was to provide a common framework for the evaluation of information extraction techniques applied to the recognition and classification of pharmacological substances (DrugNER subtask) and the detection and classification of drug-drug interactions (DDIExtraction subtask) from biomedical texts. The DDI corpus is a valuable gold-standard for those research groups interested in the recognition of pharmacological active substances, including drugs, groups of drugs, toxins, etc. or those specifically working in the field of DDI relation extraction.

The DDI corpus is divided into two datasets: training and test. The training dataset is the same for both subtasks and contains gold-standard annotations of pharmacological substances and their interactions. It consists of 714 texts (572 from DrugBank and 142 MedLIne abstracts) annotated with a total of 13029 pharmacological substances (13029 from DrugBank and 1826 from MedLine) and 4037 DDIs (3805 from DrugBank and 232 from MedLine). The test dataset for the Drug NER subtask consists of 52 DrugBank texts (annotated with 303 pharmacological substances) and 58 MedLine abstracts (with 382 pharmacological substances). The test dataset for the subtask of DDI extraction consists of 158 DrugBank Texts (annotated with 889 DDIs) and 33 MedLine abstracts (with 95 DDIs).

Technology: -

Technical Requirements:

Modules:

Innovation: -

Development: -

Publications:

- María Herrero-Zazo, Isabel Segura-Bedmar, Paloma Martínez, Thierry Declerck, The DDI corpus: An annotated corpus with pharmacological substances and drug?drug interactions, Journal of Biomedical Informatics, Volume 46, Issue 5, October 2013, Pages 914-920, ISSN 1532-0464, http://dx.doi.org/10.1016/j.jbi.2013.07.011.)
- Isabel Segura-Bedmar, Paloma Martínez, María Herrero-Zazo. SemEval-2013 Task 9 : Extraction of Drug-Drug Interactions from Biomedical Texts (DDIExtraction 2013). In Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval 2013).

Contact: Isabel Segura-Bedmar (isegura@inf.uc3m.es), Paloma Martínez (pmf@inf.uc3m.es)

DeliciousT140

Authors: Arkaitz Zubiaga, Alberto P. García-Plaza, Víctor Fresno, Raquel Martínez (UNED)

References: http://nlp.uned.es/~azubiaga/delicioust140

Description: DeliciousT140: Colección de 144.574 documentos web en inglés, con su correspondiente información de tags extraída de Delicious en junio de 2008, a partir de los feeds ofrecidos por este sitio web para los 140 tags más populares.

Functionality: Análisis y minería de etiquetas sociales, Recuperación de Información, Clasificación Automática.

Technology: Se ofrecen los documentos web en su formato original (html, pdf,...), junto con un documento xml que contiene la información de tags.

Technical Requirements: 2 GB de espacio en disco

Modules: -

Innovation: Primera colección disponible de información de etiquetas sociales que incluye también el contenido de los documentos web.

Development: Desarrollo de un artículo de investigación, desarrollo de una tesis.

Publications: No disponible aún.

Contact: Arkaitz Zubiaga <azubiaga@lsi.uned.es>

Diccionario de colocaciones del Español (DICE)

Authors: Margarita Alonso Ramos

References: http://www.dicesp.com/

Description: El Diccionario de Colocaciones del Español es un diccionario que proporciona información sobre la coocurrencia restringida de las palabras del español, de manera similar que diccionarios comerciales del inglés como el Oxford Collocations Dictionary o Macmillan Collocations Dictionary, y, además también contiene derivados semánticos. El usuario puede, por ejemplo, consultar qué adjetivos se utilizan con el nombremiedo para encontrar combinaciones como miedo cerval, miedo atroz, miedo ancestral, miedo infundado, miedo escénico, etc., o encontrar que una persona que tiende a sentir o es susceptible de sentir miedo puede designarse mediante los adjetivos miedoso o pavoroso.

Functionality: El DiCE es libremente accesible en la web desde el 2004 y su base de datos se está mejorando constantemente. La descripción especialmente detallada aplicada a las colocaciones representa una implementación práctica de los fundamentos teóricos introducidos por la Lexicología Explicativa y Combinatoria de Igor Mel'čuk. Además, nuestro objetivo ha sido que el diccionario sirva como herramienta útil no solo para la investigación lingüística, sino también para un público menos especializado. Para ello, hemos tratado de adaptar los términos y anotaciones especializados del marco teórico de manera que sean accesibles a cualquier usuario, asimismo hemos diseñado una interfaz de búsqueda que permite distintas maneras de acceso a la información léxica almacenada.

Technology: El DiCE está desarrollado en PHP utilizando el framework CakePHP, con una base de datos MySQL Server y está corriendo en un servidor web Apache.

Technical Requirements: Las consultas al DiCE se realizan vía web
http://www.dicesp.com/accesodiccionario/lemas

Modules: -

Innovation: -

Development: El desarrollo del DiCE ha sido parcialmente financiado por el Ministerio de Ciencia e Innovación dentro del marco de los proyectos Creación de contenidos. Un entorno de aprendizaje de colocaciones basado en la web (FFI2008-06479-C02-01) y Herramienta de ayuda a la redacción en español: fundamentos lingüísticos para el procesamiento de colocaciones (FFI2011-30219-C02-01).

Publications:

- Orsolya Vincze and Margarita Alonso, Incorporating Frequency Information in a Collocation Dictionary: Establishing a Methodology, Procedia – Social and Behavioral Sciences, 95:241-248, 2013. ISSN 1877-0428. DOI 10.1016/j.sbspro.2013.10.644

- Margarita Alonso, De diccionarios a herramientas interactivas de aprendizaje: colocaciones en español, 19. Hispanistentag – XIX Congreso de la Asociación Alemana de Hispanistas, Münster, Germany, 2013.

- Orsolya Vincze and Margarita Alonso, Testing an electronic collocation dictionary interface: Diccionario de Colocaciones del Español, en Proceedings of eLex 2013: electronic lexicography in the 21st century: thinking outside the paper, pp. 328-337, Tallin, Estonia, 2013.

- Margarita Alonso, Explorando la frecuencia léxica para el Diccionario de colocaciones del español, in Tomás Jiménez Juliá, Belén López Meirama, Victoria Vázquez Rozas and Alexandre Veiga (eds.), Cum corde et in nova grammatica: Estudios ofrecidos a Guillermo Rojo, pp. 19-40, Servizo de Publicacións e Intercambio científico, Universidade de Santiago de Compostela, Santiago de Compostela, Spain, 2012. ISBN 978- 84-9887-914-8.

- Vincze, O., E. Mosqueira y M. Alonso Ramos, An online collocation dictionary of Spanish, en Boguslavsky, I. y L. Wanner, eds.Proceedings of the 5th International Conference on Meaning-Text Theory Barcelona, September 8-9, 2011, pp. 275-286.

- Margarita Alonso, Alfonso Nishikawa and Orsolya Vincze, DiCE in the web: An online Spanish collocation dictionary, en S. Granger, M. Paquot (Eds.), eLexicograpy in the 21st century: New Challenges, New Applications. Proceedings of eLex 2009, pp. 369-374, Cahiers du Cental 7, Louvain-la-Neuve, Presses universitaires de Louvain, Belgium, 2010. ISBN 978-2-87463-211-2.

Contact: margarita.alonsoudc.es

DrugNer

Authors: Isabel Segura-Bedmar, Paloma Martínez, María Segura-Bedmar

References : Advanced Databases Group (Labda) of Universidad Carlos III de Madrid is a research group with an extensive activity in several Natural Language Processing and Information Retrieval projects: http://basesdatos.uc3m.es/index.php?id=202&L=0

Description: DrugNer Corpus: a corpus annotated with generic drug names and other biomedical concepts by ISABEL SEGURA-BEDMAR is licensed under a Creative Commons Reconocimiento-No comercialCompartir bajo la misma licencia 3.0 Unported License. http://labda.inf.uc3m.es/DrugDDI/.

Functionality: -

Technology: DrugNer is an XML database. An XML database allows data to be stored in XML format. This data can then be queried, exported and serialized into the desired format. An XML database defines a logic model from an XML document and stores and retrieves information according to this model. An XML database does not use SQL like query language. The XML database supports at least one form of querying syntax. Minimally, just about all of them support XPath for performing queries against documents or collections of documents. XPath provides a simple pathing system that allows users to identify nodes that match a particular set of criteria. In addition, it supports XSLT as a method of transforming documents or query-results retrieved from the database. XSLT provides a declarative language written using an XML grammar. The main systematic criteria and methodology for ordering the data are the attributes ID for each element in XML format. Also, every element or attribute can be queried by some queryng language.

Technical Requirements: -

Modules: -

Innovation: This is the first biomedical corpus annotated with drug and their pharmacological families. DrugNer, could encourage research on automatic extraction information of drug interactions, adverse drug events and other drug information, occurring in biomedical and pharmacological texts.

Development: This corpus was part of the thesis "Application of information extraction techniques to pharmacological domain: extracting drug-drug interactions" Isabel Segura-Bedmar, Advisor: Paloma Martínez. Recently, this thesis has been granted with the Extraordinary PhD award 2011. This work has been partially supported by the Spanish research projects: MA2VICMR consortium (S2009/TIC-1542, www.mavir.net), a network of excellence funded by the Madrid Regional Government and TIN2007-67407-C03-01 (BRAVO: Advanced Multimodal and Multilingual Question Answering).

Publications:

- Isabel Segura-Bedmar, Paloma Martínez, Maria Segura-Bedmar. Drug name recognition and classification in biomedical texts A case study outlining approaches underpinning automated systems. Drug Discovery Today, Volume 13, Issues 17-18, September 2008, Pages 816-823

- Isabel Segura-Bedmar, Paloma Martínez, Doaa Samy. A preliminary approach to recognize generic drug names by combining UMLS resources and USAN naming conventions. In Proceedings of BIONLP'08, Association for Computational Linguistics (ACL) . Columbus, Ohio, 2008.

- Isabel Segura-Bedmar, Paloma Martínez, Doaa Samy. Detección de fármacos genéricos en textos biomédicos. Revista Española para el procesamiento del lenguaje natural (SEPLN), Marzo 2008. nº 40. pp 27-34

Contact: Isabel Segura-Bedmar <isegura@inf.uc3m.es>, Paloma Martínez <pmf@inf.uc3m.es>

DrugNerAr corpus

Authors: Isabel Segura-Bedmar, Mario Crespo, Paloma Martínez, César de Pablo-Sánchez

References: http://labda.inf.uc3m.es/

Description: There is no corpus dedicated to the resolution of the anaphoric expressions occurring in drug interaction descriptions in pharmacological documents. A collection of 49 unstructured and plain documents was taken randomly from the field 'interactions' in the DrugBank database. Documents have on average 40 sentences, 716 words and 331 anaphoric expressions. Documents were downloaded by using an automatic

robot developed with the free tool openKapow. Each document was subsequently preprocessed by MMTx and the DrugNer system. The corpus was annotated manually by a linguist with the assistance of a pharmaceutical expert over the output of MMTx and DrugNer.

Functionality:

Technology: DrugNer is an XML database. An XML database allows data to be stored in XML format. This data can then be queried, exported and serialized into the desired format. An XML database defines a logic model from an XML document and stores and retrieves information according to this model. An XML database does not use SQL like query language. The XML database supports at least one form of querying syntax. Minimally, just about all of them support XPath for performing queries against documents or collections of documents. XPath provides a simple pathing system that allows users to identify nodes that match a particular set of criteria. In addition, it supports XSLT as a method of transforming documents or query-results retrieved from the database. XSLT provides a declarative language written using an XML grammar. The main systematic criteria and methodology for ordering the data are the attributes ID for each element in XML format. Also, every element or attribute can be queried by some queryng language.

Technical Requirements:

Modules:

Innovation: There is no corpus dedicated to the resolution of the anaphoric expressions occurring in drug interaction descriptions in pharmacological documents. The DrugNerAr corpus is the only annotated resource for drug anaphoric expressions built to date. This corpus is free for academic research and is available in http://labda.inf.uc3m.es/DrugDDI/.

Development: This corpus was part of the thesis "Application of information extraction techniques to pharmacological domain: extracting drug-drug interactions" Isabel Segura-Bedmar, Advisor: Paloma Martínez. Recently, this thesis has been granted with the Extraordinary PhD award 2011. This work has been partially supported by the Spanish research projects: MA2VICMR consortium (S2009/TIC-1542, www.mavir.net), a network of excellence funded by the Madrid Regional Government and TIN2007-67407-C03-01 (BRAVO: Advanced Multimodal and Multilingual Question Answering).

Publications:

- Isabel Segura-Bedmar, Mario Crespo, César de Pablo-Sánchez, Paloma Martínez, (2010). Resolving anaphoras for the extraction of drug-drug interactions in pharmacological documents. , April, 2010, BMC BioInformatics, ISSN: 1471-2105, Volumen: 11, Número: (Suppl 2).
- Isabel Segura-Bedmar, Mario Crespo, César de Pablo-Sánchez, Paloma Martínez, (2010). Score-based approach for Anaphora Resolution in Drug-Drug Interactions Documents, April, 2010, Natural Language Processing and Information Systems, Springer Berlin / Heidelberg, ISBN: 978-3-642-125, ISSN: 0302-9743, Volumen: 5723/2010, Páginas: 91-102, url.
- Sergio Aparicio, Isabel Segura-Bedmar, (2009). Resolución de expresiones anafóricas en textos biomédicos., Colmenarejo, España, February, 2009, III Jornadas PLN-TIMM, Páginas: 47-48.

Contact: Isabel Segura-Bedmar (isegura@inf.uc3m.es), Paloma Martínez (pmf@inf.uc3m.es)

EDBL lexical database

Authors: IXA group. Propiedad intelectual: GI-97/1596

References: http://ixa2.si.ehu.es/demo/edbl.jsp

Description: EDBL (Euskararen Datu-Base Lexikala) is a general-purpose lexical database used in Basque text-processing tasks. It is a large repository of lexical knowledge (currently around 80,000 entries) that acts as basis and support in a number of different NLP tasks, thus providing lexical information for several language tools: morphological analysis, spell checking and correction, lemmatization and tagging, syntactic analysis, and so on. It has been designed to be neutral in relation to the different linguistic formalisms, and flexible and open enough to accept new types of information. A browser-based user interface makes the job of consulting the database, correcting and updating entries, adding new ones, etc. easy to the lexicographer.

Functionality: Lexical database for Basque

Technology: Oracle. XML exportation.

Technical Requirements: -

Modules: -

Innovation: First Lexical database for Basque

Development: Different projects funded by the Basque government.

Publications:

- Aduriz I., Aldezabal I., Ansa O., Artola X., Díaz de Ilarraza A. EDBL: a Multi-Purposed Lexical Support for the Treatment of Basque. Proceedings of the First International Conference on Language Resources and Evaluation.Vol II. pp 821-826. Granada. May 28-30, 1998.
- Aldezabal I., Ansa O., Arrieta B., Artola X., Ezeiza A., Hernández G., Lersundi M. EDBL: a General Lexical Basis for the Automatic Processing of Basque. IRCS Workshop on linguistic databases. Philadelphia (USA). 2001.

Contact: Iñaki Alegria <i.alegria@ehu.es>

EmIroGeFB

Authors: Francisco Rangel, Irazú Hernández, Paolo Rosso, Antonio Reyes

References: http://ow.ly/uQWEs

Description: Corpus de comentarios de Facebook en español sobre 3 dominios (política, fútbol, celebrities) que ha sido etiquetado con las 6 emociones básicas joy, surprise, fear, anger, disgust, sadness), ironía y género (masculino/feminino) por tres etiquetadores. Los tres etiquetadores independientes etiquetaron 1200 documentos.

Functionality: -

Technology: -

Technical Requirements: -

Modules: -

Innovation: -

Development: -

Publications: -

Contact:

EmotiCorpus

Authors: Davide Buscaldi (Ph.D. student) and Paolo Rosso

References: http://www.dsic.upv.es/grupos/nle/resources/emoticorpus.xml.bz2

Description: This resource is an annotated corpus of quotes from the Italian wikiquote collection.

Functionality: This corpus can be used in order to carry out experiments over Automatic Humour Recognition (AHR).

Technology: It is a set of XML-formatted files. It was produced with a Java interface.

Technical Requirements: None.

Modules: -

Innovation: This is currently the only available resource for AHR in Italian.

Development: Developed as part of the MiDES CICYT TIN2006-15265-C06-04 research project.

Publications:

- Davide Buscaldi and Paolo Rosso, Some Experiments in Humour Recognition Using the Italian Wikiquote Collection. Applications of Fuzzy Sets Theory, LNAI vol. 4578, Springer, pp. 464-468, 2007.

Contact: Davide Buscaldi <dbuscaldi@dsic.upv.es>

English-Spanish dictionary of weighted morphological forms

Authors: Alberto Barrón-Cedeño / NLE Lab. ELiRF – Grigori Sidorov / Instituto Politécnico Nacional (Mexico).

References: http://users.dsic.upv.es/grupos/nle/resources/abc/download-morphdict.html

Description: This dictionary contains an exhaustive list of forms weighted according to the distributions of corresponding grammar classes in reference corpora.

Functionality: It can be useful for tasks such as: Cross-Language Information Retrieval, Cross-Language Plagiarism Detection, and Machine Translation

Innovation: The produced statistical bilingual dictionary represents a useful resource for various NLP **cross-language** tasks.

Development: MICINN research project TEXT-ENTERPRISE 2.0 TIN2009-13391-C04-03 (Plan I+D+i). Developed as part of the Ph.D. Thesis of Alberto Barrón-Cedeño (writing-up phase).

Publication:

- Sidorov G., Barrón-Cedeño A., Rosso P. English-Spanish Large Statistical Dictionary of Inflectional Forms. In: Proc. 7th Int. Conf. on Language Resources and Evaluation, LREC-2010, Malta, May 17-23, pp. 277-281, 2010

Contact: Paolo Rosso (prosso@dsic.upv.es)

Enriched List of Questions in Arabic

Authors: Paolo Rosso / NLE Lab. EliRF, Lahsen Abouenour, Ecole Mohammadia d'Ingenieurs Rabat (Morocco).

References: http://users.dsic.upv.es/grupos/nle/?file=kop4.php

Description: Set of TREC and CLEF questions in Arabic enriched with a query expansion process. These questions have been expanded using an Arabic WordNet-based semantic Query Expansion process divided into four types: by Synonyms, by Definitions, by Subtypes and by Supertypes.

Functionality: Useful for **Question Answering in Arabic**

Innovation: List of TREC and CLEF questions in Arabic. Moreover they have been enriched semantically (query expansion process).

Development: MICINN research project TEXT-ENTERPRISE 2.0 TIN2009-13391-C04-03 (Plan I+D+i); PCI-AECID C/026728/09 research project. Developed as part of the Ph.D. Thesis of Lahsen Abouenour (writing-up phase).

Publication:

- Abouenour L., Bouzoubaa K., Rosso P. An evaluated semantic QE and structure-based approach for enhancing Arabic Q/A. In: IEEE Int. Journal on Information and Communication Technologies, Vol. 3, No. 3, pp.37-51, 2010

Contact: Paolo Rosso (prosso@dsic.upv.es)

EPEC-DEP

Authors: IXA group

References: http://ixa.si.ehu.es/Ixa/resources/Treebank

Description: EPEC is a corpus of standard written Basque that has been manually tagged at different levels (morphology, surface syntax, phrases) and is currently being hand tagged at deep syntax level following the

Dependency Structure-based Scheme. It is aimed to be a "reference" corpus for the development and improvement of several NLP tools for Basque. This corpus has already been used for the construction of some tools such as a morphological analyser, a lemmatiser, or a shallowsyntactic analyser.

The EPEC-DEP corpus is the EPEC (Reference Corpus for the Processing of Basque) corpus manually tagged with dependency relations. Part of this work was developed in the CESS-ECE project (HUM2004-21127). Since in this project the constituents based syntactic formalism was chosen for consulting the corpus of all languages, the conversion from the dependencies to the constituents had to be done. In this way, it is possible to get the EPEC corpus tagged either with dependency relations or with constituent relations.

Functionality: -

Technology: -

Technical Requirements: -

Modules: -

Innovation: First Basque corpus manually tagged at different levels (morphology, surface syntax, phrases).

Development:

Publications:

- Aduriz I., Aranzabe M., Arriola J., Atutxa A., Díaz de Ilarraza A., Ezeiza N., Gojenola K., Oronoz M., Soroa A., Urizar R. 2003. Methodology and steps towards the construction of EPEC, a corpus of written Basque tagged at morphological and syntactic levels for the automatic processing. Proceedings of the Corpus Linguistics 2003. Lancaster

- Aldezabal I., Aranzabe M.J., Arriola J.M., Díaz de Ilarraza A., Estarrona A., Fernandez K., Uria L., Quintian M.. 2007. EPEC (Euskararen Prozesamendurako Erreferentzia Corpusa) dependentziekin etiketatzeko eskuliburua. UPV/EHU / LSI / TR 12-2007

Contact: Iñaki Alegria <i.alegria@ehu.es>

EPEC-Eusemcor

Authors: IXA group. Propiedad intelectual: SS-411-08

References: http://sisx04.si.ehu.es:8080/eusemcor

Description: Eusemcor is a hand annotated corpora for Basque (the Basque Semcor). This joint development allows for better motivated sense distinctions, and a tighter coupling between both resources. Themethodology involves edition, tagging and refereeing tasks. We are currently half way through the nominal part of the 300.000 word corpus (roughly equivalent to a 500.000 word corpus for English).

Functionality: -

Technology: -

Technical Requirements: -

Modules: -

Innovation: First Basque corpus manually tagged at semantic level

Development: -

Publications:

- Agirre E., Aldezabal I., Etxeberria J., Izagirre E., Mendizabal K., Quintian. M., eta Pociello E. A methodology for the joint development of the Basque WordNet and Semcor. Proceedings of the 5th International Conference on Language Resources and Evaluations (LREC), Genoa (Italia). 2006.

- Agirre E., Aldezabal I., Etxeberria J., Izagirre E., Mendizabal K., Quintian. M., eta Pociello E. Eusemcor: euskarako corpusa semantikoki etiketatzeko eskuliburua: editatze- etiketatze- eta epaitze-lanak. Barne-txostena, Euskal Herriko Unibertsitatea, 2005.

Contact: Iñaki Alegria <i.alegria@ehu.es>

eSOL

Authors: M. Dolores Molina-González, Eugenio Martínez-Cámara, María-Teresa Martín-Valdivia y José M. Perea-Ortega

References: http://sinai.ujaen.es/wp-content/uploads/2013/05/esol.tar.gz

Description: eSOL es una lista de palabras indicadoras de opinión en español dependientes del dominio. El dominio del conjunto de palabras es el de críticas de cine.

Para la elaboración de la lista se ha seguido un enfoque basado en corpus. En este caso se ha seleccionado el corpus de críticas de cine en español Spanish Movie Reviews. La lista está formada por 2.535 palabras positivas y 5.639 palabras negativas. Para más información sobre como se ha elaborado la lista puede consultar el artículo: Semantic Orientation for Polarity Classification in Spanish Reviews (In revision).

Functionality: -

Technology: -

Technical Requirements: -

Modules: -

Innovation: -

Development: -

Publications:

- M Dolores Molina-González, Eugenio Martínez-Cámara, María-Teresa Martín-Valdivia, José M Perea-Ortega. 2013. Semantic orientation for polarity classification in Spanish reviews. Expert Systems with Applications, 40(18): 7250-7257.

Contact: M. Dolores Molina-González <mdmolina@ujaen.es>, Eugenio Martínez-Cámara <emcamara@ujaen.es>, María-Teresa Martín-Valdivia <maite@ujaen.es> y José M. Perea-Ortega<jmperea@ujaen.es>

EuroWordNet

Authors: TALP, CLIC and UNED.

References: http://www.lsi.upc.edu/~nlp/projectes/ewn.html , http://www.illc.uva.nl/EuroWordNet

Description: EuroWordNet is a multilingual database with wordnets for several European languages (Dutch, Italian, Spanish, German, French, Czech and Estonian). The wordnets are structured in the same way as the American wordnet for English (Princeton WordNet, Miller et al. 1990) in terms of synsets (sets of synonymous words) with basic semantic relations between them. Each wordnet represents a unique language-internal system of lexicalizations. In addition, the wordnets are linked to an Inter-Lingual-Index, based on the Princeton wordnet. Via this index, the languages are interconnected so that it is possible to go from the words in one language to similar words in any other language. The index also gives access to a shared top ontology of 63 semantic distinctions. This top ontology provides a common semantic framework for all the languages, while language specific properties are maintained in the individual wordnets. The database can be used, among others, for monolingual and cross-lingual information retrieval, which was demonstrated by the users in the project.

Functionality: To add Semantic Knowledge to any Resource.

Technology: Java, Php, Perl are used for the Demos. MySQL is used by the Database containing MCR.

Technical Requirements: Java, Perl, Apache and MySqL are required for the use of MCR.

Modules: -

Innovation: This tool was created in the framework of the EuroWordNet project. The most important results were: 1) the wordnets for each separate language (Dutch, Italian, Spanish, French, German, Czech and Estonian) linked to the English WordNet,; 2) the addition to Wordnet1.5 of relations not covered in the Princeton Wordnet for English,; 3) WordNet1.5 in EuroWordNet format; 4) an Inter-Lingual-Index (based on WordNet1.5) to connect the different wordnets and other ontologies; 5) the shared top ontology; 6) Polaris: a wordnet editor to create and edit wordnets linked to EuroWordNet; 7) Periscope: the EuroWordNet viewer in which all this can be viewed and selections can be exported; 8) a report on the demonstration of the results in information retrieval tasks.

Development: It was developed in the framework of an European Project called EuroWordNet: Building a multilingual wordnet with semantic relations between words, which lasted 3 years, from 1996 to 1999.

Publications: There are a lot of related publications, including Deliverables and Reports that can be found at http://www.illc.uva.nl/EuroWordNet/docs.html

Contact: Lluís Padró <padro@lsi.upc.edu >

EuskalWordnet

Authors: IXA group. Propiedad intelectual: SS-323-03

References: http://ixa2.si.ehu.es/mcr/wei.html

Description: The Basque WordNet follows the EuroWordNet framework and, basically, it is produced using a semi-automatic method that links Basque words to the English WordNet. We have found that in order to ensure proper linguistic quality and avoid excessive English bias, a double manual pass on the automatically produced Basque synsets is desirable: a first concept-to-concept pass to ensure correctness of the Basque words linked to the synsets, and a word-to-word pass to ensure the completeness of the word senses linked to the words. By this method, we expect to combine quick progress (as allowed by a development based on the English WordNet) with quality (as provided by a development based on a native dictionary). We have completed the concept-to-concept review of the automatically produced links for the nominal concepts, and are currently performing the word-to-word review.

Functionality: Semantic database for Basque

Technology: MySQL

Technical Requirements: -

Modules: -

Innovation: First Semantic database for Basque

Development: Different projects funded by the Basque government.

Publications:

- Agirre E., Ansa O., Arregi X., Arriola J., Díaz de Ilarraza A., Pociello E., Uria L. Methodological issues in the building of the Basque WordNet: quantitative and qualitative analysis. Proceedings of First International WordNet Conference. pp. 32-40. Mysore (India). 2002.

Contact: Iñaki Alegria <i.alegria@ehu.es>

EVOCA Corpus

Authors: Mohammed Rushdi-Saleh, M Teresa Martín-Valdivia, L Alfonso Ureña-López, José M Perea-Ortega

References: http://sinai.ujaen.es/wp-content/uploads/2013/11/EVOCA-corpus.rar

Description: EVOCA (English Version of OCA) es un corpus en inglés generado a partir de la traducción del corpus OCA en árabe. Este corpus contiene comentarios de películas y está dividido en 250 comentarios considerados positivos y 250 negativos. Algunas estadísticas sobre EVOCA corpus. Este corpus fue traducido en Abril de 2011.

Functionality: -

Technology: -

Technical Requirements: -

Modules: -

Innovation: -

Development: -

Publications:

- Rushdi Saleh, M., Martín-Valdivia, M. T., Ureña-López, L. A. & Perea-Ortega, J. M. (2011). Bilingual Experiments with an Arabic-English Corpus for Opinion Mining. Proceedings of Recent Advances in Natural Language Processing, pages 740–745.

Contact: José M. Perea <jmperea@ujaen.es>

Features Inventory

Authors: Antonio Reyes / NLE Lab. ELiRF

References: http://users.dsic.upv.es/grupos/nle/resources/Signatures.components

Description: File containing the elements to represent all the dimensions regarding signatures features (emoticons, counter-factuality items, temporal compression items)

Functionality: It allows carrying out experiments on **Irony Detection**

Innovation: Publicly available information wrt signature features for irony

Development: MICINN research project TEXT-ENTERPRISE 2.0 TIN2009-13391-C04-03 (Plan I+D+i). Developed as part of the Ph.D. Thesis of Antonio Reyes (writing-up phase).

Contact: Paolo Rosso (prosso@dsic.upv.es)

Geo-WordNet

Authors: Davide Buscaldi (Ph.D. student) and Paolo Rosso

References: http://www.dsic.upv.es/grupos/nle/resources/geo-wn/download.html

Description: This is a semi-automatically generated mapping from WordNet 2.0 to geographical coordinates.

Functionality: It allows to assign a point in the map (geo-reference) to all WordNet synsets that are related to a geographical entity.

Technology: It was developed using a MySQL database and WordNet, however it is merely a list of WordNet synsets and their coordinates.

Technical Requirements: It requires WordNet 2.0 in order to be used.

Modules: -

Innovation: This is the first time that an effort is done to connect WordNet to geographical databases, therefore connecting two different type of research communities (Natural Language research groups with GIS research groups).

Development: Developed as part of the MiDES CICYT TIN2006-15265-C06-04 research project, co-funded by the AECI-PCI A01031707 project.

Publications:

- Davide Buscaldi and Paolo Rosso, Geo-WordNet: Automatic Georeferencing of WordNet. In: Proceedings of the Sixth International Language Resources and Evaluation (LREC'08), Marrakech, Morocco, 2008.

Contact: Davide Buscaldi <dbuscaldi@dsic.upv.es>

Geo-WordNet 3.0

Authors: Davide Buscaldi, Paolo Rosso / NLE Lab. ELiRF

References: http://www.dsic.upv.es/grupos/nle/resources/geo-wn/download.html , http://wordnet.princeton.edu/wordnet/related-projects/

Description: Geo-WordNet 3.0 connects WordNet synsets with their geographical coordinates (latitude and longitude). In the new 3.0 version, the source of geographical data was Geonames (http://www.geonames.org). Therefore, it was possible to assign to every synset a Geonames ID, together with the coordinates. Geo-WordNet is constituted of a plain text file where every line contains the following fields: <synset offset> <geonames ID> <latitude> <longitude>, separated by a tabulation character. The synsetoffsets used in Geo-WordNet 3.0 correspond to those included in WordNet 3.0.

Functionality: Geo-WordNet extends WordNet with geographical data, allowing all WordNet-based applications to associate spatial information to un-structured texts.

Technology: Geo-WordNet 3.0 has been developed in Java using the MIT Java WordNet Interface (MIT JWI), with data from WordNet 3.0 and Geonames. Geographical data were loaded into a PostgreSQL database.

Technical Requirements: The only requirement is WordNet 3.0 as a reference for synset offsets.

Modules: The distribution is a tar.gz file including a folder named "GeoWN3.0", which contains the following files: 00README.txt, LICENSE, mapping.dat, mapping.txt, not_mapped.txt. Main data are contained in the "mapping.dat" file, while "mapping.txt" and "not_mapped.txt" contain explications on the associations between WordNet synsets and Geonames Ids.

Innovation: Geo-WordNet is the only resource associating geographical information to synsets.

Development: MICINN research project TEXT-ENTERPRISE 2.0 TIN2009-13391-C04-03 (Plan I+D+i).

Developed as part of the Ph.D. Thesis of Davide Buscaldi "Toponym Disambiguation in Information Retrieval", Universidad Politécnica de Valencia, 2010

Publications:

- Buscaldi, D., Rosso, P. Geo-WordNet: Automatic Georeferencing of WordNet. In: Proc. 5th Int. Conf. on Language Resources and Evaluation, LREC-2008, Marrakech, Morocco, May 2008
- Buscaldi D., Rosso P. Using GeoWordNet for Geographical Information Retrieval. In: Revised Selected Papers CLEF-2008, Springer-Verlag, LNCS(5706), pp. 863-866
- Buscaldi D., Toponym Disambiguation in Information Retrieval, Ph.D. Thesis, Universidad Politécnica de Valencia, 2010

Contact: Davide Buscaldi davide.buscaldi@univ-orleans.fr (Paolo Rosso prosso@dsic.upv.es)

GeoSemCor2.0

Authors: Davide Buscaldi (Ph.D. student) and Paolo Rosso

References: http://www.dsic.upv.es/grupos/nle/resources/geosemcor2.0.tar.gz

Description: This resource is the SemCor corpus labeled with WordNet 2.0 synsets, enriched with the addition of labels for synsets that are related to geographical entities.

Functionality: It is a corpus that can be used to evaluate Toponym Disambiguation methods.

Technology: It is a set of SGML-formatted files.

Technical Requirements: It needs WordNet 2.0 in order to be used.

Modules: -

Innovation: This is the first time that a resource aimed at the evaluation of Toponym Disambiguation is produced.

Development: Developed as part of the MiDES CICYT TIN2006-15265-C06-04 research project.

Publications:

- Davide Buscaldi and Paolo Rosso, A conceptual density-based approach for the disambiguation of toponyms. International Journal of Geographical Information Science, 22(3), pages 301-313, Taylor and Francis, 2008.
- Davide Buscaldi and Paolo Rosso, Map-based vs. Knowledge-based Toponym Disambiguation., in: Proc. 5th Int. Workshop on Geographical Information Retrieval, GIR-2008, CIKM-2008, pages 19-22, 2008.

Contact: Davide Buscaldi <dbuscaldi@dsic.upv.es>

Ironic Quotes

Authors: Antonio Reyes / NLE Lab. ELiRF

References: http://users.dsic.upv.es/grupos/nle/resources/ironicQuotes.source

Description: This corpus has been manually created on the basis of the **irony tag** that users employ in their posts in blogs on the Web. It contains comments related to the irony tag.

Functionality: It allows carrying out experiments on Irony Detection.

Innovation: No public corpora are available for **Irony Detection.**

Development: MICINN research project TEXT-ENTERPRISE 2.0 TIN2009-13391-C04-03 (Plan I+D+i). Developed as part of the Ph.D. Thesis of Antonio Reyes (writing-up phase).

Contact: Paolo Rosso (prosso@dsic.upv.es)

iSOL

Authors: M. Dolores Molina-González, Eugenio Martínez-Cámara, María-Teresa Martín-Valdivia y José M. Perea-Ortega

References: http://sinai.ujaen.es/wp-content/uploads/2013/05/isol.tar.gz

Description: iSOL es una lista de palabras indicadoras de opinión en español independiente del dominio.

Para la elaboración del recurso se ha partido de la lista de palabras que mantiene el profesor Bing Liu (Bing Liu's Opinion Lexicon). La lista de palabras ha sido traducida automáticamente usando el traductor Reverso y posteriormente se han corregido manualmente.

La lista está formada por 2.509 palabras positivas y por 5.626. Para más información sobre como se ha desarrolla la lista puede consultar el artículo: Bilingual Experiments on an Opinion Comparable Corpus (in press).

Functionality: -

Technology: -

Technical Requirements: -

Modules: -

Innovation: -

Development: -

Publications:

- M Dolores Molina-González, Eugenio Martínez-Cámara, María-Teresa Martín-Valdivia, José M Perea-Ortega. 2013. Semantic orientation for polarity classification in Spanish reviews. Expert Systems with Applications, 40(18): 7250-7257.

Contact: M. Dolores Molina-González <mdmolina@ujaen.es>, Eugenio Martínez-Cámara <emcamara@ujaen.es>, María-Teresa Martín-Valdivia <maite@ujaen.es> y José M. Perea-Ortega<jmperea@ujaen.es>

Lexicon of Prototypical Discourse Markers

Authors: Laura Alonso

References: http://russell.famaf.unc.edu.ar/~laura/shallowdisc4summ/discmar

Description: This is the seminal discourse marker lexicon used in the thesis Representing discourse for automatic text summarization via shallow NLP techniques. The discourse markers listed here were the primary source of evidence to draw the semantic maps to obtain an inventory of basic discursive meanings. This lexicon is also the basis for the implementations of a discourse segmenter and for the discourse analysis exploited by the e-mail summarizer Carpanta.

The lexicon is parallel in three languages: Catalan, Spanish and English. Therefore, in this starting version of the lexicon we have only included those discourse markers that have a near-synonym in one of the other languages. The lexicon is formed by 84 discourse markers, representing different discursive meanings. Some discourse markers have been assigned to more or less than one meaning per dimension, because they are ambiguous or underspecified, respectively. In this lexicon, discourse markers are characterized by their structural (continuation or elaboration) and semantic (revision, cause, equality, context) meanings, and they are also associated to a morphosyntactic class (part of speech, PoS), one of adverbial (A), phrasal (P) or conjunctive (C).

Functionality: The semantic information associated to discourse markers can be integrated into any tool that exploits these lexical items as source of evidence of discursive structure in texts. It has been integrated in a segmenter in discursive units and in some automatic summarizers.

Technology: The lexicon is in raw text.

Technical Requirements: -

Modules: -

Innovation: It constitutes a lexico-semantic resource that, to our knowledge, was not existing for Spanish and Catalan.

Development: This lexicon represents part of the work carried out in writing the doctoral dissertation "Representing discourse for automatic text summarization via shallow NLP" by Laura Alonso.

Publications:

- Laura Alonso i Alemany, Ezequiel Andújar Hinojosa and Robert Sola Salvatierra, (2004), A framework for feature-based description of low level discourse, in *Discourse Annotation*, workshop at the ACL'04

- Laura Alonso, Jennafer Shih, Irene Castellón, Lluís Padró, (2003), An Analytic Account of Discourse Markers for Shallow NLP, in *The Meaning and Implementation of Discourse Particles* , whorkshop held as part of the Fifteenth European Summer School in Logic, Language and Information, ESSLLI'03, 18-19 August, Vienna, Austria

Contact: Laura Alonso Alemany. Facultad de Matemática, Astronomía y Física. Universidad Nacional de Córdoba, Argentina.

LibiXaml

Authors: IXA group. Propiedad intelectual: SS-412-08

References: -

Description: It is a framework for creating, browsing and editing linguistic annotations generated by a set of different linguistic processing tools.

Functionality: The aim is to establish a flexible and extensible infrastructure which follows a coherent and general representation scheme. This proposal provides us with a well-formalized basis for the exchange of linguistic information. It uses TEI-P4 conformant feature structures as a representation schema for linguistic analyses.

Technology: XML, C++, Berkeley DBXML and DB

Technical Requirements: -

Modules: -

Innovation: -

Development: -

Publications:

- Artola X., Díaz de Ilarraza A., Ezeiza N., Gojenola K., Sologaistoa A., Soroa A. 2004. EULIA: a graphical web interface for creating, browsing and editing linguistically annotated corpora. LREC 2004. ISBN 2-9517408-1-6

- Artola X., Díaz de Ilarraza A., Ezeiza N., Gojenola K., Labaka G., Sologaistoa A., Soroa A. 2005. A framework for representing and managing linguistic annotations based on typed feature structures. RANLP 2005. ISBN: 954-91743-3-6

Contact: Iñaki Alegria <i.alegria@ehu.es>

MCE Corpus

Authors: Martín-Valdivia, M. T., Martínez-Cámara, E., Perea-Ortega, J. M., & Alfonso Ureña-López, L.

References: http://sinai.ujaen.es/wp-content/uploads/2013/05/MCE-corpus.tar.gz

Description: MuchoCine corpus en Inglés (MCE) es la versión traducida del corpus MuchoCine (Spanish Movies Reivews). El corpus de MuchoCine fue elaborado por el investigador Fermín Cruz Mata y presentado en el año 2008 en el número 41 de la revista Procesamiento del Legujae Natural en el artículo titulado Clasificación de documentos basada en la opinión: experimentos con un corpus de críticas de cine en español.

En el artículo Sentiment polarity detection in Spanish reviews combining supervised and unsupervised approaches comprueba la validez de una metodología para la clasificación de la polaridad en español que consiste en combinar tres clasificadores, dos supervisdados (sobre textos en inglés y en otro idioma) y otro

no supervisado usando algún recurso lingüístico en inglés para análisis de opiniones. Esta metodología fue propuesta previamente para opiniones en árabe en el artículo Improving Polarity Classification of Bilingual Parallel Corpora combining Machine Learning and Semantic Orientation approaches (in press).

Functionality: -

Technology: -

Technical Requirements: -

Modules: -

Innovation: -

Development: -

Publications: Solo se permite el uso de este corpus para investigación. En tal caso, debe citar el siguiente artículo:

- Martín-Valdivia, M. T., Martínez-Cámara, E., Perea-Ortega, J. M., & Alfonso Ureña-López, L. (2012). Sentiment polarity detection in Spanish reviews combining supervised and unsupervised approaches. Expert Systems with Applications. (http://dx.doi.org/10.1016/j.eswa.2012.12.084)

Contact: José M. Perea <jmperea@ujaen.es> y Eugenio Martínez Cámara <emcamara@ujaen.es>

MCR: Multilingual Central Repository

Authors: TALP, and the other members of the MEANING project.

References: http://garraf.epsevg.upc.es/cgi-bin/wei4/public/wei.consult.perl , http://www.lsi.upc.es/~nlp/meaning/demo/demo.html

Description: The Multilingual Central Repository (MCR) follows the model proposed by the EuroWordNet project. EuroWordNet (Vossen, 1998) is a multilingual lexical database with wordnets for several European languages, which are structured as the Princeton WordNet. The Princeton WordNet contains information about nouns, verbs, adjectives and adverbs in English and is organized around the notion of a synset.

The current version of the MCR (Atserias et al., 2004) is a result of the 5th Framework MEANING project. The MCR integrates into the same EuroWordNet framework wordnets from five different languages (together with four English WordNet versions). The MCR also integrates WordNet Domains (Magnini and Cavaglià, 2000) and new versions of the Base Concepts and Top Concept Ontology. The final version of the MCR contains 1,642,389 semantic relations between synsets, most of them acquired by automatic means. This is almost one order of magnitude larger than the Princeton WordNet (204,074 unique semantic relations in WordNet 2.0).

Functionality: To add Semantic Knowledge to any Resource.

Technology: Java, Php, Perl are used for the Demos. MySQL is used by the Database containing MCR.

Technical Requirements: Java, Perl, Apache and MySqL are required for the use of MCR.

Modules: -

Innovation: Currently, MCR integrates into the EuroWordNet framework five local wordnets (including four versions of the English WordNet from Princeton), the EuroWordNet Top Concept ontology, MultiWordNet Domains, and hundreds of thousand of new semantic relations and properties automatically acquired from corpora. MCR constitutes the largest and richest multilingual resource for lexical knowledge ever build.

Development: MCR (Multilingual Central Repository) was the result of the 5th Framework European MEANING Project (2002-2005).

Publications: There are a lot of related publications, including Deliverables and Reports that can be found here: http://www.lsi.upc.es/~nlp/meaning/documentation/3rdYear

- Cuadros M. and Rigau G. KnowNet: Building a Large Net of Knowledge from the Web. 22nd International Conference on Computational Linguistics COLING'08. Manchester, UK. 2008.
- Álvez J., Atserias J., Carrera J., Climent S., Laparra E., Oliver A. and Rigau G. Complete and Consistent Annotation of WordNet using the Top Concept Ontology. 6th international conference on Language Resources and Evaluation, LREC'08, Marrakesh, Morroco. 2008.
- Cuadros M. and Rigau G. Quality Assessment of Large-Scale Knowledge Resources. Proceedings of Joint SIGDAT Conference on Empirical Methods in Natural Language Processing (EMNLP'06). Sydney, Australia. 2006.
- Atserias J., Villarejo L., Rigau G., Agirre E., Carroll J., Magnini B., Vossen P. The MEANING Multilingual Central Repository. In Proceedings of the Second International Global WordNet Conference (GWC-2004). ISBN 80-210-3302-9. Brno, Czech Republic. Enero, 2004.

Contact: German Rigau <german.rigau@ehu.es>

ML-SentiCon: A Layered, Multilingual Sentiment Lexicon

Authors: Fermín L. Cruz, José A. Troyano, Beatriz Pontes, F. Javier Ortega

References: http://www.lsi.us.es/~fermin/index.php?title=Datasets

Description: Se trata de varias listas de lemas positivos y negativos para inglés, español, catalán, gallego y vasco. Cada lema viene acompañado de una estimación numérica de su polaridad (entre -1.0 y 1.0) así como de un valor de desviación típica de dicha polaridad. Las listas están organizadas en varias capas, de manera que las primeras capas contienen estimaciones más precisas de los valores anteriores, aunque contienen menos elementos que las capas posteriores.

Además de las listas de lemas, el recurso también contiene un lexicón a nivel de synsets para el inglés, con el mismo formato de SentiWordNet (SWN). Este lexicón fue obtenido a partir de una versión mejorada del método original de SWN, y utilizado para la generación de las listas de lemas anteriores.

Functionality: -

Technology: -

Technical Requirements: -

Modules: -

Innovation: -

Development: -

Publications:

- Cruz, Fermín L., José A. Troyano, Beatriz Pontes, F. Javier Ortega. Building layered, multilingual sentiment lexicons at synset and lemma levels, Expert Systems with Applications, 2014.

Contact: Fermín L. Cruz: fcruz@us.es

OCA Corpus

Authors: Mohammed Rushdi-Saleh, M Teresa Martín-Valdivia, L Alfonso Ureña-López, José M Perea-Ortega

References: http://sinai.ujaen.es/wp-content/uploads/2013/11/OCA-corpus.zip

Description: OCA es un corpus en árabe sobre comentarios de películas. Este corpus ha sido generado a partir de comentarios en árabe obtenidos de diferentes páginas web.

Functionality: -

Technology: -

Technical Requirements: -

Modules: -

Innovation: -

Development: -

Publications:

- Rushdi Saleh, M., Martín-Valdivia, M. T., Ureña-López, L. A. & Perea-Ortega, J. M. (2011). OCA: Opinion Corpus for Arabic. Journal of the american society for information science and technology, 62(10): 2045-2054.

Contact: José M. Perea <jmperea@ujaen.es>

Opinion analysis corpus

Authors: Enrique Vallés, Paolo Rosso / NLE Lab. ELiRF

References: http://users.dsic.upv.es/grupos/nle/?file=kop4.php

Description: The corpus contains 3,000 opinions on the domain of tourism. These opinions have been obtained from the TripAdvisor blog.

Functionality: It allows carrying out experiments on **Opinion Mining**

Innovation: Opinion retrieved from the TripAdvisor blog

Development: MICINN research project TEXT-ENTERPRISE 2.0 TIN2009-13391-C04-03 (Plan I+D+i).

Developed as part of the M.Sc. Thesis of Enrique Vallés "Empresa 2.0: Detección de plagio y análisis de opiniones ", Universidad Politécnica de Valencia, 2010

Publications:

- Vallés E., Rosso P., Locoro A., Mascardi V. Análisis de opiniones con Ontologías. In: POLIBITS, Research Journal on Computer Science and Computer Engineering with Applications, num. 41, pp. 29-37, 2010
- Vallés E., Rosso P. Empresa 2.0: Detección de plagio y análisis de opiniones. In: Proc. SEPLN Workshop on NLP in the Enterprise: Envisioning the Next 10 Years (PLN-E), CEUR-WS.org, vol. 697, pp. 9-12, 2010

Contact: Paolo Rosso (prosso@dsic.upv.es)

SENSEM Corpus

Authors: Glòria Vàzquez, Ana Fernández, Laura Alonso , Irene Castellón

References: Grup de Recerca Interuniversitari en Aplicacions Lingüístiques (GRIAL): http://grial.uab.es

Description: This corpus includes Spanish journalistic texts, more precisely, it is a collection of news extracted from *El Periódico de Catalunya.* It has been manually annotated at a syntactic (phrases and syntactic function) and semantic level (semantic roles, semantic constructions and sense disambiguation). The corpus has approximately 700.000 words. It contains sentences with the 250 more frequent verbs in Spanish.

Functionality: The interface (http://grial.uab.es/search) allows simple and advanced searches on the corpus by different fields, including the negative search. The XML corpus can be downloaded.

Technology: The corpus is stored in a Mysql database and the interface has been developed in PHP.

Technical Requirements: -

Modules: -

Innovation: The main innovation is the amount of examples (100 examples for each verb), in addition to the syntactic and semantic annotation, since there are few corpus annotated with such information.

Development: The Corpus has been developed with the Sensem project: *Banco de datos sintáctico y semántico del español. 2004-2006 - Ministerio de Ciencia y Tecnología (BFF2003-06456)* and at this moment it is under development thanks to the project: *Ampliación de la BD léxica y el corpus sintáctico-semántico de semántica oracional del español SenSem Ministerio de educación y Ciencia HUM2007-65267*

Publications:

- Alonso, L., I. Castellón, N. Tincheva (2006). "Detección automática de errores en el Corpus Sensem", Congreso de la Asociación Española de Lingüística Aplicada (AESLA)

- Alonso, L., J.A. Capilla, I. Castellón, A. Fernández, G. Vázquez (2005). "The Sensem Project: Syntactico-Semantic Annotation of Sentences in Spanish", Proceedings of the International Conference RANLP, p. 39-46. Borovets, Bulgaria. ISBN: 954-91743-3-6

- Alonso, L., J.A. Capilla, I. Castellón, A. Fernández, G. Vázquez (2007). "The Sensem Project: Syntactico-Semantic Annotation of Sentences in Spanish". N.Nikolov, K. Bontcheva, G.Angelova and R. Mitkov. (ed.), Recent Advances in Natural Language Processing IV. Selected papers from RANLP 2005. Current Issues in Linguistic The*ory 292*John Benjamins Publishing Co, p. 89-98. ISBN: 978 90 272 4807 7

- Castellón, I., A. Fernández, G. Vázquez (2005). "La semántica oracional del español: perspectiva desde el léxico". G. Wotjak, J. Cantero (ed.), Entre semántica léxica, teoría del léxico y sintaxis. Frankfurt:Leipzig. Peter Lang, Europaishcher Verlag der Wissenschaften, p. 113-122. ISBN: 3-631-53207-5. ISSN: 1436-1914

- Castellón, I., A. Fernández, G. Vázquez, L. Alonso, J.A. Capilla (2006). "The Sensem Corpus: a Corpus Annotated at the Syntactic and Semantic Level", Fifth International Conference on Language Resources and Evaluation (LREC), p. 355-359

- Fernádez, A., G. Vázquez, D. Teruel (2006). "Interfaz de explotación del corpus SenSem", Congreso de la Asociación Española de Lingüística Aplicada (AESLA)

- Fernández, A., G. Vázquez (2007). "Problemas sobre la distinción entre argumentos y adjuntos en el corpus SenSem". Castellón, I., A. Fernández (ed.), Perspectivas de análisis de la unidad verbal. Seres. Barcelona:Publicacions i Edicions de la Universitat de Barcelona, p. 35-48. ISBN: 978-84-475-3177-6

- Fernández, A., G. Vázquez, I. Castellón (2004). "Sensem: base de datos verbal del español". G. de Ita, O. Fuentes, M. Osorio (ed.), IX Ibero-American Workshop on Artificial Intelligence, IBERAMIA. Puebla de los Ángeles, Mexico:, p. 155-163. ISBN: 968-863-786-6

- Fernández, A., G. Vázquez, I. Castellón (2006). "SenSem: a Databank for Spanish Verbs", Proceedings of the X Ibero-American Workshop on Artificial Intelligence, IBERAMIA.. Ribeirão Preto, Brasil

- Vázquez, G., A. Fernández (2008). "Annotation de corpus: Sur la délimitation des arguments et des adjoints", SKY Journal of Linguistics

- Vázquez, G., A. Fernández, L. Alonso (2005). "Description of the Guidelines for the Syntactico-semantic Annotations of a Corpus in Spanish". Angelova, G., K. Bontcheva, R. Mitkov, N. Nicolov (ed.), International Conference Recent Advances in Natural Language. Shoumen (Bulgaria):, p. 603-607. ISBN: 954-91743-3-6

- Vázquez, G., A. Fernández, L. Alonso (2005). "Description of the Guidelines for the Syntactico-semantic Annotations of a Corpus in Spanish". Angelova, G., K. Bontcheva, R. Mitkov, N. Nicolov (ed.), International Conference Recent Advances in Natural Language. Shoumen (Bulgaria):, p. 603-607. ISBN: 954-91743-3-6

- Vázquez, G., L. Alonso, J.A. Capilla, I. Castellón, A. Fernández (2006). "SenSem: sentidos verbales, semántica oracional y anotación de corpus", Procesamiento del Lenguaje Natural, 37, p. 113-120. ISSN: 1135-5948

Contact: Ana Maria Fernández Montraveta <ana.fernandez@uab.es>

SENSEM Verbal DB

Authors: Glòria Vázquez, Ana Fernández, Laura Alonso , Irene Castellón

References: Grup de Recerca Interuniversitari en Aplicacions Lingüístiques (GRIAL): http://grial.uab.es

Description: The lexical database contains the most frequent 250 Spanish verbs, a total of 1000 senses. These senses are described from a syntactic and semantic perspective: semantic roles, eventive class, synonyms, and includes a mapping with the 1.5, 1.6, 2.1 and 3.0 WordNet versions. Other information provided in the lexical database has been inferred from the annotation of a journalistic corpus of over 700,000 words: examples, subcategorization patterns and semantic constructions with frequency information.

Functionality: The query interface displays the examples of corpus and its associated annotation. In addition, the interface displays information in a compact or expanded mode, depending on the user's needs.

Technology: The information is stored in a Mysql database and the interface has been developed in php.

Technical Requirements: -

Modules: -

Innovation: The database contains information that has been acquired from corpus and the examples can be consulted from database. The senses are mapping with WordNet .

Development: The Database has been developed with the Sensem project: *Banco de datos sintáctico y semántico del español. 2004-2006 - Ministerio de Ciencia y Tecnología (BFF2003-06456)* and at this moment it is under development thanks to the project Ampliación de la BD léxica y el corpus sintáctico-semántico de semántica oracional del español SenSem Ministerio de educación y Ciencia HUM2007-65267

Publications:

- Alonso, L., I. Castellón and N. Tincheva (2007). "Obtaining coarse-grained classes of subcategorization patterns for Spanish", *Proceedings of the International Conference RANLP*

- Alonso,L.,I. Castellón, N. TInkova (2007). "Adquisición de subcategorizaciones verbales mediante un clasificador automático", *Revista de la SEPLN*

- Carrera J., I. Castellón, S. Climent and M. Coll-Florit (2008). "Towards Spanish verbs' selectional preferences automatic acquisition. Semantic annotation of SenSem corpus", *Proceedings of The 6th international conference on Language Resources and Evaluation, LREC 2008*. ISBN: 2-9517408-4-0

- Carrera Ventura,J.T. (2007), *Análisis de técnicas de adquisición automática de restricciones selectivas*. GRIAL- Research Report 3/2007 Departament de Lingüística General. Universitat de Barcelona

- Castellón, I, L. Alonso, N.T. Tincheva (2008). "A procedure to automatically enrich verbal lexica with subcategorization frames". Lawrence Mandow (ed.), *Inteligencia Artificial*. Malaga (España):, 12:37, p. 45-53. ISSN: 1137-3601

- Castellón, I., A. Fernández, G. Vázquez (2005). "La semántica oracional del español: perspectiva desde el léxico". G. Wotjak, J. Cantero (ed.), *Entre semántica léxica, teoría del léxico y sintaxis*. Frankfurt:Leipzig. Peter Lang, Europaishcher Verlag der Wissenschaften, p. 113-122. ISBN: 3-631-53207-5. ISSN: 1436-1914

- Castellón, I., L. Alonso and N. Tincheva (2007). "A procedure to automatically enrich verbal lexica with subcategorization frames", *Argentinean Symposium on Artificial Intelligence, ASAI'07.*

- Castellón,I i A. Fernández (eds.) (2007), *Perspectivas de análisis de la unidad verbal. Seres.*. Barcelona: Publicacions i Edicions de la Universitat de Barcelona. ISBN: 978-84-475-3177-6

- Coll, M. (2006). "Diferentes niveles de aspecto". WORKSHOP SERES. Universitat de Barcelona

- Coll-Florit, M. (2007). "Aktionsart y polisemia verbal", *Actas del III Congreso Internacional de Lingüística Hispánica (CILHIS)*

- Coll-Florit, M., S.Climent, I.Castellón (2007). "Aspecto léxico y desambiguación semántica. El caso de los estados". Ricardo Mairal Usón (ed.), *Aprendizaje de lenguas, uso del lenguaje y modelación cognitiva: perspectivas aplicadas entre disciplinas*. Madrid:. ISBN: 978-84-611-6897-2

- Fernández, A., G. Vázquez, I. Castellón (2004). "Sensem: base de datos verbal del español". G. de Ita, O. Fuentes, M. Osorio (ed.), *IX Ibero-American Workshop on Artificial Intelligence, IBERAMIA*. Puebla de los Ángeles, Mexico:, p. 155-163. ISBN: 968-863-786-6

- Fernández, A., G. Vázquez, I. Castellón (2006). "SenSem: a Databank for Spanish Verbs", *Proceedings of the X Ibero-American Workshop on Artificial Intelligence, IBERAMIA.*. Ribeirão Preto, Brasil

- Vázquez, G., L. Alonso, J.A. Capilla, I. Castellón, A. Fernández (2006). "SenSem: sentidos verbales, semántica oracional y anotación de corpus", *Procesamiento del Lenguaje Natural*, 37, p. 113-120. ISSN: 1135-5948

Contact: Ana Maria Fernández Montraveta <ana.fernandez@uab.es>

SINAI SA Corpus

Authors: Rushdi Saleh, M., Martín-Valdivia, M. T., Montejo-Ráez, A., & Ureña-López, L. A.

References: http://sinai.ujaen.es/wp-content/uploads/2013/11/SINAI-SA-corpus.zip

Description: Este corpus ha sido preparado por el grupo SINAI en Diciembre de 2008. SINAI SA (Análisis de Sentimientos) fue creado rastreando la página web de Amazon. Casi 2000 comentarios se extrajeron de diferentes cámaras.

Estructura: El corpus de SINAI contiene 5 directorios y cada uno representa el número de estrellas por comentario. (ej. el directorio 1 contiene los valorados con una estrella). Cada directorio contiene un fichero en texto plano por documento/comentario.

La cantidad de comentarios se detalla a continuación:

- 1 estrella: 78 comentarios
- 2 estrellas: 67 comentarios
- 3 estrellas: 97 comentarios
- 4 estrellas: 411 comentarios
- 5 estrellas: 1,290 comentarios

Total: 1,943 comentarios

Functionality: -

Technology: -

Technical Requirements: -

Modules: -

Innovation: -

Development: -

Publications:

- Rushdi Saleh, M., Martín-Valdivia, M. T., Montejo-Ráez, A., & Ureña-López, L. A. (2011). Experiments with SVM to classify opinions in different domains.Expert Systems with Applications, 38(12), 14799-14804.

Contact: Maite Martín Valdivia <maiteujaen.es>

Single-label hep-ex Clustering Corpus

Authors: Alfonso Ureña López and Arturo Montejo Ráez (Universidad Jaén). Pre-processed by David Pinto; Héctor Jiménez (Universidad Autónoma Metropolitana, México).

References: http://www.dsic.upv.es/grupos/nle/downloads.html

Description: This corpus is a pre-processed version of the collection of scientific abstracts compiled by the University of Jaén, Spain named hep-ex [1].

Functionality: The aim of the pre-processed version of this corpus is to support experiments of supervised and unsupervised classifiers with narrow domain short texts.

Technology: The corpus (raw text) and the gold standard are provided.

Technical Requirements: No special requirements are needed in order to use the corpus.

Modules: -

Innovation: A relatively big pre-processed collection which may be used to experiment with different clustering methods on the narrow domain short-text clustering task.

Development: Developed as part of David Pinto Ph.D. and the MiDES CICYT TIN2006-15265-C06-04 research project.

Publications:

- Arturo Montejo-Ráez, Luis Alfonso Ureña-López, Ralf Steinberger; Text Categorization using bibliographic records: beyond document content. Procesamiento del Lenguaje Natural, nº 35, Septiembre 2005.

- David Pinto, Alfons Juan, Paolo Rosso: A Comparative Study of Clustering Algorithms on Narrow-Domain Abstracts. Procesamiento del Lenguaje Natural 37(1): 43-49, 2006.

- Héctor Jiménez-Salazar, David Pinto, Paolo Rosso: Uso del Punto de Transición en la Selección de Términos Índice para Agrupamiento de Textos Cortos, Procesamiento del Lenguaje Natural 35(1): 114-118, 2005.

- David Pinto, José-Miguel Benedí, Paolo Rosso: Clustering Narrow-Domain Short Texts by Using the Kullback-Leibler Distance. CICLing 2007. Lecture Notes in Computer Science 4394, Springer-Verlag: 611-622, 2007.

- David Pinto, Héctor Jiménez-Salazar, Paolo Rosso: Clustering Abstracts of Scientific Texts Using the Transition Point Technique. CICLing 2006. Lecture Notes in Computer Science 3878, Springer-Verlag:536-546, 2006.

Contact: David Eduardo Pinto Avendaño <dpintocs.buap.mx>

Social-ODP-2k9

Authors: Arkaitz Zubiaga

References: http://nlp.uned.es/social-tagging/socialodp2k9/

Description: Social-ODP-2k9 is a dataset created during December 2008 and January 2009 with data retrieved from the social bookmarking sites Delicious and StumbleUpon, the Open Directory Project and the Web. It is made up by 12,616 unique web documents, along with their corresponding social annotations, and classification data according to the ODP.

Functionality: Web page classification, analysis of social annotations, etc.

Technology: Data stored in XML format.

Technical Requirements:

Modules:

Innovation: To the best of our knowledge, this is the only dataset including social tagging data along with crawled documents and category data.

Development: The generation of this dataset was partially funded by the Regional Government of Madrid under the Research Network MAVIR (S-0505/TIC-0267), the Regional Ministry of Education of the Community of Madrid, by the Spanish Ministry of Science and the Innovation project QEAVis-Catiex (TIN2007-67581-C02-01).

Publications:

- Arkaitz Zubiaga, Raquel Martínez, and Víctor Fresno. Getting the Most Out of Social Annotations for Web Page Classification. Proceedings of DocEng 2009, the 9th ACM Symposium on Document Engineering, pp. 74-83, Munich, Germany. 2009.

Contact: Arkaitz Zubiaga azubiaga@lsi.uned.es

SoCo corpus

Authors: Enrique Flores, Paolo Rosso, Lidia Moreno, Esaú Villatoro

References: http://users.dsic.upv.es/grupos/nle/soco/corpus.html

Description: Este corpus pertenece a la competición internación SOCO en detección de reutilización de código fuente que se celebra en el forum internación FIRE2014. Consiste en códigos fuente escritos en C y Java con casos reales de reutilización de código fuente.

Functionality: -

Technology: -

Technical Requirements: -

Modules: -

Innovation: -

Development: -

Publications: -

Contact: -

Spanish QC

Authors: Miguel Ángel García Cumbreras

References: http://sinai.ujaen.es/wp-content/uploads/2013/11/Clasificacion-QA-6305.label_.txt

Description: Este recurso son 6305 preguntas en español etiquetadas para clasificación de Búsqueda de Respuestas, siguiendo la taxonomía definida en el artículo "X. Li and D. Roth. Learning Question Classifiers", y que tiene las siguientes categorías generales y detalladas:

- ABBR: abbreviation, expansion
- DESC: definition, description, manner, reason
- ENTY: animal, body, color, creation, currency, disease/medical, event, food, instrument, language, letter, other, plant, product, religion, sport, substance, symbol, technique, term, vehicle, word

- HUM: description, group, individual, title
- LOC: city, country, mountain, other, state
- NUM code, count, date, distance, money, order, other, percent, period, speed, temperature, size, weight

Partiendo de un conjunto de preguntas etiquetadas para inglés se ha generado este recurso con preguntas diversas en español etiquetadas y revisadas por 3 personas.

Functionality: -

Technology: -

Technical Requirements: -

Modules: -

Innovation: -

Development: -

Publications: García-Cumbreras, M. A., Ureña-López, L. A. & Martínez-Santiago, F. (2006). BRUJA: Question Classification for Spanish. Using Machine Translation and an English Classifier. EACL 2006 Workshop on Multilingual Question Answering – MLQA06.

Contact: Miguel Ángel García Cumbreras <magc@ujaen.es>

Spanish WordNet 3.0

Authors: Ana Fernández, Glòria Vázquez, Irene Castellón

References: Grup de Recerca Interuniversitari en Aplicacions Lingüístiques (GRIAL): http://grial.uab.es

Description: An open-source lexical and semantic resource for Spanish that has been created from the latest version of the English WordNet (3.0) and connected with it through the ID and the alignment of words contained in glosses and examples. It contains especially nouns and verbs. These words are annotated with morphosyntactic and semantic information.

Functionality: The query interface (http://grial.uab.es/recursos/wordnet30) displays all the information of the resource by selecting the corresponding option in the menu: definitions in Spanish and English, morphosyntactic and semantic annotation, Spanish examples and their translation into English. Moreover, the user can make queries about the variants; all queries can be done for both English and Spanish Wordnet"

Technology: The data are stored in a Mysql database and the interface has been developed in php.

Technical Requirements: -

Modules: -

Innovation: An open-source lexical and semantic resource for Spanish that contains an annotated corpus at semantic level (disambiguation of senses). Until this moment there was not an open-source Spanish version of Wordnet .

Development: This resource has been developed with the project "Traducción y anotación de las glosas de la Ontología WordNet 3.0 para la lengua española" Ministerio de Educación y Ciencia. HUM2006-27968-E.

Publications:

- Fernández, A., G. Vázquez (2008). "La construcción del WordNet 3.0 en español", Actas del III Congreso Internacional de Lexicografía
- Fernández-Montraveta, A., G. Vázquez, C. Fellbaum (2008). "The Spanish Version of WordNet 3.0". Storrer, A. et al. (ed.), Text Resources and Lexical Knowledge. Berlin:Mouton de Gruyter, p. 175-182. ISBN: 978-3-11-020735-4

Contact: Ana Maria Fernández Montraveta <ana.fernandez@uab.es>

Taxonomy-Based Opinion Dataset

Authors: Fermín L. Cruz Mata, JosA. Troyano Jiménez, Pablo Montoya.

References: http://www.lsi.us.es/~fermin/TBOD.tar.gz

Description: This dataset contains annotated reviews for three different domains: cars, headphones and hotels. Opinions are annotated at the feature level, with the following fields:

Required:

- polarity: positive (+) or negative (-).
- feature: a feature from the feature taxonomy.
- opWords: opinion words. The minimum set of words from the sentence from which you can decide the polarity of this opinion.

Optional:

- featWords: feature words. A set of words from the sentence naming the feature.
- potency: potency words. A set of words from the sentence which affect the strength of the opinion.

The feature taxonomy for each domain is defined in an xml file (featureTaxonomy.xml). For each feature, a set of feature words is included.

Development: Partially funded by Ministerio de Eduación y Ciencia (HUM2007-6607-C04-04).

Publications:

- F. L. Cruz, J. A. Troyano, F. Enríquez, J. Ortega, and C. G.Vallejo. Knowledge-rich approach to feature-based opinion extraction from product reviews In Proceedings of the 2nd International Workshop on Search and Mining User-Generated Contents, pages 130. ACM, 2010.

- Fermín L. Cruz, José A. Troyano, Fernando Enríquez, F. Javier Ortega, Carlos G. Vallejo: 'Long autonomy or long delay?' The importance of domain in opinion mining. Expert Syst. Appl. 40(8): 3174-3184 (2013)

Contact: Fermín L. Cruz , University of Seville, fcruz@us.es

The Arabic Wikipedia XML corpus

Authors: David Pinto(R30 version) and Ludovic Denoyer and Patrick Gallinari

References: http://www.dsic.upv.es/grupos/nle/downloads.html

Description: The 30 most frequent categories of the Arabic Wikipedia XML corpus gathered by Ludovic Denoyer and Patrick Gallinari were selected in order to provide a testbed for the single-label categorization task in the Arabic language.

Functionality: The aim of this corpus is to support experiments of supervised and unsupervised classifiers with Arabic-witten texts. The gold standard is provided, as well as the tokenized and untokenized versions of this corpus.

Technology: The corpus (raw text of tokenized and untokenized versions) and the gold standard are provided.

Technical Requirements: No special requirements are needed in order to use the corpus.

Innovation: This is an attempt to provide easy access to pre-processed texts in order to be used in the Arabic categorization task.

Development: Developed as part of David Pinto Ph.D. and the MiDES CICYT TIN2006-15265-C06-04 research project.

Publications:

- David Pinto: On Clustering of Narrow Domain Short-Text Corpora. PhD Thesis, Universidad Politécnica de Valencia, Spain, July 2008.
- David Pinto, Paolo Rosso, Yassine Benajiba, Anas Ahachad, Héctor Jiménez-Salazar: Word Sense Induction in the Arabic Language: A Self-Term Expansion Based Approach. The Egyptian Society of Language Engineering (ESOLE): 235-245, 2007.

Contact: David Eduardo Pinto Avendaño <dpinto@cs.buap.mx>

The DrugNer corpus

Authors: Isabel Segura-Bedmar, Paloma Martínez, Maria Segura-Bedmar

References: Advanced Databases Group (Labda) (http://labda.inf.uc3m.es/) of Universidad Carlos III de Madrid is a research group with an extensive activity in several Natural Language Processing, Information Retrieval and Information Extraction projects.

Description: DrugNer Corpus: a corpus annotated with generic drug names and other biomedical conceptsby ISABEL SEGURA-BEDMAR is licensed under a Creative Commons Reconocimiento-No comercialCompartir bajo la misma licencia 3.0 Unported License. http://labda.inf.uc3m.es/DrugDDI/.

Functionality:

Technology: DrugNer is an XML database. An XML database allows data to be stored in XML format. This data can then be queried, exported and serialized into the desired format. An XML database defines a logic model from an XML document and stores and retrieves information according to this model. An XML database does not use SQL like query language. The XML database supports at least one form of querying syntax. Minimally, just about all of them support XPath for performing queries against documents or collections of documents. XPath provides a simple pathing system that allows users to identify nodes that match a particular set of criteria. In addition, it supports XSLT as a method of transforming documents or query-results retrieved from the database. XSLT provides a declarative language written using an XML grammar. The main systematic criteria and methodology for ordering the data are the attributes ID for each element in XML format. Also, every element or attribute can be queried by some queryng language.

Technical Requirements

Modules:

Innovation: This is the first biomedical corpus annotated with drug and their pharmacological families. DrugNer, could encourage research on automatic extraction information of drug interactions, adverse drug events and other drug information, occurring in biomedical and pharmacological texts.

Development: This corpus was part of the thesis "Application of information extraction techniques to pharmacological domain: extracting drug-drug interactions" Isabel Segura-Bedmar, Advisor: Paloma Martínez. Recently, this thesis has been granted with the Extraordinary PhD award 2011. This work has been partially supported by the Spanish research projects: MA2VICMR consortium (S2009/TIC-1542, www.mavir.net), a network of excellence funded by the Madrid Regional Government and TIN2007-67407-C03-01 (BRAVO: Advanced Multimodal and Multilingual Question Answering).

Publications:

- Isabel Segura-Bedmar, Paloma Martínez, María Segura-Bedmar, (2008). Drug Name Recognition and classification in biomedical texts. , September, 2008, Drug Discovery Today, Elsevier Science, ISSN: 1359-6446, Volumen: 13, Número: 17-18, Páginas: 816-823, url.
- Isabel Segura-Bedmar, Paloma Martínez, Doaa Samy, (2008). Detección de fármacos genéricos en textos biomédicos., March, 2008, Revista Española para el procesamiento del lenguaje natural (SEPLN), ISSN: 1135-5948, Páginas: 27-34, pdf.
- Isabel Segura-Bedmar, Paloma Martínez, Doaa Samy, (2008). A preliminary approach to recognize generic drug names by combining UMLS resources and USAN naming conventions., Ohio, USA, June, 2008, Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing (BioNLP), Association for Computational Linguistics, ISBN: 978-1-932432-, Páginas: 100-10

Contact: Isabel Segura-Bedmar (isegura@inf.uc3m.es), Paloma Martínez (pmf@inf.uc3m.es)

The KnCr clustering corpus

Authors: David Pinto

References: http://www.dsic.upv.es/grupos/nle/downloads.html

Description: This is a new narrow-domain short text corpus in the medicine domain which was constructed by downloading the last sample of documents provided in MEDLINE and selecting only those which are related with the "Cancer" domain.

Functionality: The aim of this corpus is to support experiments of supervised and unsupervised classifiers with narrow domain short texts, especifically in the medicine field, with documents related with the “cancer” topic.

Technology: The corpus (raw text) and the gold standard are provided.

Technical Requirements: No special requirements are needed in order to use the corpus.

Innovation: To our knowledge, no other corpus of cancer domain has been constructed in order to be used in the categorization task.

Development: Developed as part of David Pinto Ph.D. and the MiDES CICYT TIN2006-15265-C06-04 research project.

Publications:

- David Pinto, Paolo Rosso: KnCr: A Short-Text Narrow-Domain Sub-Corpus of Medline. TLH 2006. Advances in Computer Science: 266-269, 2006
- David Pinto, Alfons Juan, Paolo Rosso: A Comparative Study of Clustering Algorithms on Narrow-Domain Abstracts. Procesamiento del Lenguaje Natural 37(1): 43-49, 2006
- David Pinto, José-Miguel Benedí, Paolo Rosso: Clustering Narrow-Domain Short Texts by Using the Kullback-Leibler Distance. CICLing 2007. Lecture Notes in Computer Science 4394, Springer-Verlag: 611-622, 2007
- David Pinto: On Clustering of Narrow Domain Short-Text Corpora. PhD Thesis, Universidad Politécnica de Valencia, Spain, July 2008.

Contact: David Eduardo Pinto Avendaño <dpinto@cs.buap.mx>

Twitter Hash tags Corpus

Authors: Antonio Reyes / NLE Lab. ELiRF

References: http://users.dsic.upv.es/grupos/nle/?file=kop4.php

Description: Corpus containing 50,000 textes extracted from Twitter. Each text contains an hash tag depending on the topic: **#humor**, **#irony**, #politics, #technology, #education

Functionality: It allows carrying out experiments on **Automatic Humour Recognition / Irony Detection.**

Innovation: No public corpora are available for irony detection

Development: MICINN research project TEXT-ENTERPRISE 2.0 TIN2009-13391-C04-03 (Plan I+D+i). Developed as part of the Ph.D. Thesis of Antonio Reyes (writing-up phase).

Contact: Paolo Rosso (prosso@dsic.upv.es)

Volem

Authors: Ana Fernández, Glòria Vázquez, Irene Castellón

References: Grup de Recerca Interuniversitari en Aplicacions Lingüístiques (GRIAL): http://grial.uab.es

Description: VOLEM (Verbs: Multilingual Lexical Organization) is a lexical multilingual data base of a subset of Spanish, Catalan and French verbs. In this multilingual resource, subcategorization frames together with their semantics are specified for approximately 100 verbs. It also provides information regarding semantic roles, selectional restrictions and examples of use. The resource was developed by researchers of 4 groups: GRIAL (Grup de Recerca Interuniversitari en Aplicacions Lingüístiques); Informatique Linguistique et Programation en Logique, (Université Paul Sabatier); IRIT (Institut de Recherche en Informatique; IXA, Universitad del País Vasco – Euskal Herriko Unibersitatea, Filosofia eta Hezkuntza Zientzien Fakultatea; TALP, Universitat Politècnica de Catalunya, Llenguatges i Sistemes Informàtics

Functionality: The query interface (http://grial.uab.es/multi/) allows the user to carry out simple and advanced searches with filters in order to obtain verb classes. Filtering criteria include: semantic role, type of construction, subcategorization pattern and semantic class. Queries can be made for just one language or for more than one language simultaneously (multilingual query). The menu interface allows users to select or deselect the search criteria they want to appear on the list. The interface also allows users to view the translation of the different verb senses in four languages (Catalan, Spanish, Basque and French).

Technology: The information is stored in a Mysql database and the interface has been developed in php.

Technical Requirements: -

Modules: -

Innovation: In this resource, the set of linguistic labels has been standardized so that the results obtained can be compared objectively. It is worth noting the inclusion of two minority languages, Basque and Catalan, for which the number of available resources is quite limited.

We would also like to point out the inclusion of information about selection preferences, a feature of particular interest in the field of NLP. Finally, one of the most relevant aspects of this project is the multilingual interface; users can work in several different languages to obtain information about verb behavior.

Development: Generalitat de Catalunya, Xarxes Temàtiques Regionals dels Pirineus (ABM/acs/XTI-CTP 2000-1, ABM/acs/XI2003-12)

Publications:

- Fernández, A., P. Saint-Dizier, G. Vázquez, F. Benamara , M. Kamel (2002). "The VOLEM Project: a Framework for the Construction of Advanced Multilingual Lexicons", *Proceedings of the Language Engineering Conference*, p. 89-98. ISBN: 0-7695-1885-0/02

Contact: Ana Maria Fernández Montraveta <ana.fernandez@uab.es>

Wiki10+

Authors: Arkaitz Zubiaga

References: http://nlp.uned.es/social-tagging/wiki10+/

Description: Wiki10+ is a dataset created during April 2009 with data retrieved from the social bookmarking site Delicious and Wikipedia. It is made up by 20,764 articles of the English Wikipedia, with their corresponding social tags. All of them had been annotated by at least 10 users on Delicious.

Functionality: Tag-based navigation, site-specific tagging, etc.

Technology: Data stored in XML format.

Technical Requirements:

Modules:

Innovation: To the best of our knowledge, this is the first dataset including social tagging data for a specific domain.

Development: The generation of this dataset was partially funded by the Regional Government of Madrid under the Research Network MAVIR (S-0505/TIC-0267), the Regional Ministry of Education of the Community of Madrid, by the Spanish Ministry of Science and the Innovation project QEAVis-Catiex (TIN2007-67581-C02-01).

Publications:

- Arkaitz Zubiaga. Enhancing Navigation on Wikipedia with Social Tags. Wikimania 2009. Buenos Aires, Argentina. 2009.

Contact: Arkaitz Zubiaga <azubiaga@lsi.uned.es>

2. Analizadores, Etiquetadores, Clasificadores

BIOS

Authors: Mihai Surdeanu

References: http://www.surdeanu.name/mihai/bios/

Description: Suite of Syntactico-Semantic Analyzers. Includes a named-entity recognizer, a syntactic chunker, a POS tagger, and a "smart" tokenizer. All processors are learned using the MiLL machine learning library (see below).

Functionality: -

Technology: Java

Technical Requirements: MiLL machine learning library, TnT tagger, YamChA.

Modules: Smart tokenizer that recognizes abbreviations, SGML tags etc.; Part-of-speech (POS) tagger. The POS tagger is implemented as a a wrapper around the TNT tagger by Thorsten Brants; Syntactic chunking using the labels promoted by the CoNLL chunking evaluations; Named-Entity Recognition and Classification (NERC) for the CoNLL entity types plus an additional 11 numerical entity types.

Innovation: -

Development: -

Publications: -

Contact: Mihai Surdeanu <mihai@surdeanu.name>

CIAOSENSO

Authors: Davide Buscaldi (Ph.D. student) and Paolo Rosso

References: http://www.dsic.upv.es/grupos/nle/resources/cdd-2.0.1.tar.gz

Description: This is a Word Sense Disambiguation (WSD) tool based on Conceptual Density.

Functionality: This tool can be used to disambiguate English text.

Technology: It has been written in C++.

Technical Requirements: any Linux, GCC4.0, WordNet.

Modules: -

Innovation: This tool participated to Senseval-3 and Semeval, obtaining good results among the unsupervised systems.

Development: Developed as part of the CIAOSENSO MCYT HI 2002-0140 research project.

Publications:

- Paolo Rosso, Francesco Masulli, Davide Buscaldi, Ferran Pla and Antonio Molina, Automatic Noun Sense Disambiguation., in: Computational Linguistics and Intelligent Text Processing, 4th International Conference, LNCS 2588, pages 273-276, Springer, 2003.
- Paolo Rosso, Francesco Masulli and Davide Buscaldi, Word Sense Disambiguation using Conceptual Distance, Frequency and Gloss, in: Int. Conf. on Natural Language Processing and Engineering Knowledge, pages 120-125, IEEE, Beijing, China, 2003.
- Davide Buscaldi, Paolo Rosso and Francesco Masulli, The upv-unige-CIAOSENSO WSD System, in: Senseval-3 workshop, ACL 2004, pages 77-82, 2004.

Contact: Davide Buscaldi <dbuscaldi@dsic.upv.es>

COMPAS (COMpiler for PArsing Schemata)

Authors: Carlos Gómez-Rodríguez, Miguel A. Alonso

References: http://www.grupolys.org/software/COMPAS/

Description: COMPAS (COMpiler for PArsing Schemata) is a system that can be used to automatically compile formal specifications of parsing algorithms (in the form of parsing schemata) to efficient Java implementations of the corresponding parsers.

Functionality: COMPAS allows compiling arbitrary user-defined parsing schemata into efficient implementations.

The distribution provided the following predefined parsing schemata for context-free grammars and tree-adjoining grammars:

- CYKRecognizer: A simple CYK context-free grammar recognizer.
- CYKVariant: A different way of expressing CYK, by expressing rules as items.
- CYKAnyGrammarRecognizer: The same as CYKRecognizer, but featuring a grammar class option so that if the input grammar is not in Chomsky Normal Form (CNF), the parser automatically transforms it to CNF when read.
- CYKWithTree: Simple CYK parser generating parse trees.
- SimpleEarleyRecognizer: An Earley recognizer, as it appears in the Parsing Schemata book by Sikkel (97).
- OptimizedEarleyRecognizer: An “unrolled” version of the Earley recognizer so that indexing is a bit faster.
- OptimizedEarleyWithTree: A simple Earley parser which generates parse trees.
- LCStandard.sch: A Left-Corner recognizer, this is the “LC” schema described by Sikkel (97). In this implementation, items are used to represent left-corner relationships.

- LCSimplifiedItems.sch: An optimized Left-Corner recognizer, this is the “sLC” schema described by Sikkel (97).
- LCWithPredicates.sch: An alternative implementation of Left-Corner where predicates, instead of items, are used to represent left-corner relationships.
- TopDown.sch: A simple, inefficient top-down parser.
- EarleyNVPforXTAG.sch: A tree-adjoining grammar parser for use with the XTAG English Grammar, including feature structure unification and several XTAG-specific features.

COMPAS can be used to compile error-repair parsing schemata. The following one is provided in the distribution:

Lyon.sch: Lyon’s error-correcting parser, using a priority queue as agenda.

Technology: The COMPAS system is written in Java

Technical Requirements: The COMPAS system is runnable in any system for which a Java Virtual Machine (JVM) is available, including Windows and Linux.

You need to have the following software installed in order to use the system::

- A Java Runtime Environment (JRE), version 1.4 or higher. Download it here.
- The Apache Ant Build System is not strictly necessary, but highly recommended in order to be able to easily compile the code generated by the system.

Once you have this software, download the system and refer to “readme.txt” for detailed usage instructions. If you understand Spanish, you can also download an user manual in this language.

Modules: -

Innovation: The first compiler for arbitrary parsing schemata.

It defined the notion of error-repair parsing schemata and provided a mechanism for efficnet compilation of these kind of schemata.

Development: -

Publications:

- Carlos Gómez-Rodríguez, Parsing Schemata for Practical Text Analysis, Imperial College Press, London, 2010. ISBN 978-1-84816-560-1.
- Carlos Gómez-Rodríguez, Miguel A. Alonso and Manuel Vilares, Error-repair parsing schemata, Theoretical Computer Science, 411(7-9):1121-1139, 2010. ISSN 0304-3975. DOI 10.1016/j.tcs.2009.12.007
- Carlos Gómez-Rodríguez, Jesús Vilares and Miguel A. Alonso, A compiler for parsing schemata, Software: Practice and Experience, 39(5):441-470, 2009. ISSN 0038-0644. DOI 10.1002/spe.904

Contact: carlos.gomez@udc.es, miguel.alonso@udc.es

Dependency Grammar for Catalan

Authors: Jordi Carrera, Marina Lloberas, Nevena Tincova, Irene Castellón

References: Grup de Recerca Interuniversitari d'Aplicacions Lingüístiques (in collaboration with TALP – UPC): http://grial.uab.es, http://garraf.epsevg.upc.es/freeling

Description: Catalan dependency grammar consists of a set of 2,914 rules, of which 2,565 complete the parse tree by creating dependencies and the remaining 349 label these dependencies. Catalan grammar treats dependency recursion and dependency relations between phrases, clauses headed by conjunctions or relative pronouns, non-finite clauses and punctuation marks.

Functionality: Dependency parsing

Technology: The grammar is included in Freeling http://www.lsi.upc.edu/~nlp/freeling

Technical Requirements: It has to be downloaded with Freeling.

Modules: -

Innovation: Wide-coverage deep parsing grammar for Catalan

Development: This grammar has been developed in the KNOW project . Ministerio de Educación y Ciencia (TIN2006-1549-C03-02)

Publications:

- J.Carrera, I. Castellón, M.Lloberes, Ll.Padró, N. Tincova (2008). "Dependency Grammars in Freeling", *Procesamiento del Lenguaje Natural*, 2008:41, p. 13-20. ISSN: 1135-5948

Contact: Irene Castellón <icastellon@ub.edu>

Dependency Grammar for English

Authors: Jordi Carrera, Marina Lloberas, Nevena Tincova, Irene Castellón

References: Grup de Recerca Interuniversitari d'Aplicacions Lingüístiques (in collaboration with TALP – UPC): http://grial.uab.es

Description: Dependency rules for the English grammar amount to circa 1606. Rules have been provided for all major kinds of clauses: declaratives, imperatives, interrogatives, completives, relatives, adverbial and existential. Analogously, separate verb phrase rules have been provided for intransitive, transitive and ditransitive sentences, including specific sets of rules for dealing with completive sentences.

Functionality: Dependency parsing of English

Technology: The grammar is included in Freeling (http://www.lsi.upc.edu/~nlp/freeling/)

Technical Requirements: It has to be downloaded with Freeling.

Modules: -

Innovation: Wide-coverage deep parsing grammar for English

Development: This grammar has been developed in the KNOW project . Ministerio de Educación y Ciencia (TIN2006-1549-C03-02)

Publications:

- J.Carrera, I. Castellón, M.Lloberes, Ll.Padró, N. Tincova (2008). "Dependency Grammars in Freeling", Procesamiento del Lenguaje Natural, 2008:41, p. 13-20. ISSN: 1135-5948

Contact: Irene Castellón <icastellon@ub.edu>

Dependency Grammar for Spanish

Authors: Jordi Carrera, Marina Lloberas, Nevena Tincova, Irene Castellón

References: Grup de Recerca Interuniversitari d'Aplicacions Lingüístiques (in collaboration with TALP – UPC): http://grial.uab.es

Description: Spanish dependency grammar consists of a set of 4349 rules, of which 3777 complete the parse tree by creating dependencies and the remaining 572 label these dependencies. These rules act on a number of categories, such as noun, verb and prepositional phrases, pronouns, coordination, passive voice, punctuation and subordination.

Functionality: Dependency parsing of Spanish

Technology: The grammar is included in Freeling (http://www.lsi.upc.edu/~nlp/freeling/)

Technical Requirements: It has to be downloaded with Freeling.

Modules: -

Innovation: Wide-coverage deep parsing grammar for Spanish

Development: This grammar has been developed in the KNOW project . Ministerio de Educación y Ciencia (TIN2006-1549-C03-02)

Publications:

- J.Carrera, I. Castellón, M.Lloberes, Ll.Padró, N. Tincova (2008). "Dependency Grammars in Freeling", Procesamiento del Lenguaje Natural, 2008:41, p. 13-20. ISSN: 1135-5948

Contact: Irene Castellón <icastellon@ub.edu>

Eihera

Authors: IXA group

References: http://ixa2.si.ehu.es/demo/entitateak.jsp

Description: Eihera is a system for Named Entity recognition and classification in written Basque. The system is designed in four steps: first, the development of a recognizer based on linguistic information represented on finite-state-transducers; second, the generation of semi-automatically annotated corpora from the result of these transducers; third, the achievement of the best possible recognizer by training different ML techniques on these corpora; and finally, the combination of the different recognizers obtained.

Functionality: Included in Zatiak

Technology: Finite-state and Machine learning.

Technical Requirements: -

Modules: Recognition by rules, recognition by ML, classification by rules, classification by ML. Eustagger is a previous step.

Innovation: It is the first NERC system for Basque.

Development: Different projects funded by the Basque government and the Spanish R&D agency.

Publications:

- Alegria I., Arregi O., Ezeiza N., Fernandez I., Urizar R. Design and Development of a Named Entity Recognizer for an Agglutinative Language. First International Joint Conference on NLP (IJCNLP-04). Workshop on Named Entity Recognition. 2004.
- Alegria I., Ezeiza N., Fernandez I., Urizar R. Named Entity Recognition and Classification for texts in Basque. II Jornadas de Tratamiento y Recuperación de Información, JOTRI, Madrid. 2003. ISBN 84-89315-33-7. 2003.

Contact: Iñaki Alegria <i.alegria@ehu.es>

Eustagger

Authors: IXA group

References: http://ixa2.si.ehu.es/demo/analisimorf.jsp

Description: Eustagger is a robust and wide-coverage morphological analyser and a Part-of-Speech tagger for Basque.

The analyser is based on the two-level formalism and has been designed in an incremental way with three main modules: the standard analyser, the analyser of linguistic variants, and the analyser without lexicon which can recognize word-forms without having their lemmas in the lexicon. Using lexical transducers for our analyser we have improved both the performance of the different components of the system and the description itself. Provides possible lemmas, PoS and other morphological information for a token. It also recognizes date/time expressions, numbers.

In the tagger combination of stochastic and rule-based disambiguation methods are applied to Basque language. The methods we have used in disambiguation are Constraint Grammar formalism and an HMM based tagger. CG rules are applied using all the morphological features and this process decreases morphological ambiguity of texts. Finally, we use the stochastic tool to select just one from the possible

remaining tags. Using only the stochastic method the error rate is about 14%, but the accuracy may be increased by about 2% enriching the lexicon with the unknown words. When both methods are combined, the error rate of the whole process is 3.5%.

Functionality: Tokenization, morphological analysis, lemmatization and tagging for Basque. There is a web service.

Technology: C++ using FSM technology from Xerox and CG library from Connexor

Technical Requirements: -

Modules: 4 main modules: tokenizer, morphological analyzer, rule-based disambiguation and HMM based disambiguation.

Innovation: Is the analyzer/tagger for Basque.

Development: Different projects funded by the Basque government and the Spanish R&D agency.

Publications:

- Alegria I., Artola X., Sarasola K., Urkia M. 1996. Automatic morphological analysis of Basque Literary & Linguistic Computing Vol. 11, No. 4, 193-203. Oxford University Press. Oxford.
- Alegria I., Aranzabe M., Ezeiza A., Ezeiza N., Urizar R. 2002 Robustness and customisation in an analyser/lemmatiser for Basque. LREC-2002 Customizing knowledge in NLP applications Workshop.
- Aduriz I., Alegria I., Arriola J.M., Urizar R. 1998. Combining Stochastic and Rule-Based Methods for Disambiguation in Agglutinative Languages COLING-ACL'98, Montreal (Canada). August 10-14, 1998.

Contact: Iñaki Alegria <i.alegria@ehu.es>

FreeLing

Authors: TALP

References: http://www.lsi.upc.edu/~nlp/freeling

Description: The FreeLing package is a library providing language analysis services. FreeLing is designed to be used as an external library from any application requiring this kind of services. This language analysis tool suite is released under the GNU General Public License of the Free Software Foundation.

Functionality: The main services offered by FreeLing library are: Text tokenization; Sentence splitting; Morphological analysis; Suffix treatment, retokenization of clitic pronouns; Flexible multiword recognition; Contraction splitting; Probabilistic prediction of unkown word categories; Named entity detection; Recognition of dates, numbers, ratios, currency, and physical magnitudes (speed, weight, temperature, density, etc.); PoS tagging; Chart-based shallow parsing; Named entity classification; WordNet based sense annotation; and Rule-based dependency parsing .

Most of these services are provided for all currently supported languages: Spanish, Catalan, Galician, Italian, and English.

Technology: C++

Technical Requirements: A typical Linux box with usual development tools: bash, make, and a C++ compiler with basic STL support. Enough hard disk space (about 120Mb) Some external libraries are required to compile FreeLing:

- libpcre (version 4.3 or higher): Perl C Regular Expressions. Included in most usual Linux distributions. You'll need binary and development packages.
- libdb (version 4.1.25 or higher): Berkeley DB. Included in all usual Linux distributions.
- libcfg+ (version 0.6.1 or higher): Configuration file and command-line options management. May not be in your linux distribution.
- Omlet & Fries (libomlet v.0.97 or later, libfries v.0.95 or later): Machine Learning utility libraries, used by Named Entity Classifier. Installation scrips are not very clever yet, so these libraries are required even if you do not plan to use the NEC ability of FreeLing. Available from http://www.lsi.upc.edu/~nlp/omlet+fries

Modules: The main processing classes in the library are: 1) tokenizer: Receives plain text and returns a list of word objects; 2) splitter: Receives a list of word objects and returns a list of sentence objects; 3) maco: Receives a list of sentence objects and morphologically annotates each word object in the given sentences. Includes specific submodules (e.g, detection of date, number, multiwords, etc.) which can be activated at will; 4) tagger: Receives a list of sentence objects and disambiguates the PoS of each word object in the given sentences; 5) parser: Receives a list of sentence objects and associates to each of them a parse_tree object; 6) dependency: Receives a list of parsed sentence objects associates to each of them a dep_tree object.

Innovation: -

Development: FreeLing was originally written by people in TALP Research Center at Universitat Politècnica de Catalunya. Spanish and Catalan linguistic data were originally developed by people in CLiC, Centre de Llenguatge i Computació at Universitat de Barcelona. Many people further contributed to it by reporting problems, suggesting various improvements, submitting actual code or extending linguistic databases (see web page).

Publications:

- Jordi Atserias and Bernardino Casas and Elisabet Comelles and Meritxell Gonzàlez and Lluís Padró and Muntsa Padró. FreeLing 1.3: Syntactic and semantic services in an open-source NLP library Proceedings of the fifth international conference on Language Resources and Evaluation (LREC 2006), ELRA. Genoa, Italy. May, 2006.
- Jordi Atserias and Elisabet Comelles and Aingeru Mayor. TXALA un analizador libre de dependencias para el castellano. Procesamiento del Lenguaje Natural, n. 35, pg. 455--456. September, 2005.
- Jordi Atserias and Josep Carmona and Irene Castellón and Sergi Cervell and Montserrat Civit and Lluís Màrquez and Ma Antònia Martí and Lluís Padró and Roberto Placer and Horacio Rodríguez and Mariona Taulé and Jordi Turmo. Morphosyntactic Analysis and Parsing of Unrestricted Spanish Text Proceedings of the 1st International Conference on Language Resources and Evaluation, LREC, pg. 1267--1274. Granada, Spain. May, 1998.

- Josep Carmona and Sergi Cervell and Lluís Màrquez and Ma Antònia Martí and Lluís Padró and Roberto Placer and Horacio Rodríguez and Mariona Taulé and Jordi Turmo. An Environment for Morphosyntactic Processing of Unrestricted Spanish Text Proceedings of the 1st International Conference on Language Resources and Evaluation, LREC, pg. 915--922. Granada, Spain. May, 1998.

- Xavier Carreras and Lluís Padró. A Flexible Distributed Architecture for Natural Language Analyzers Proceedings of the 3rd International Conference on Language Resources and Evaluation, LREC,

- Las Palmas de Gran Canaria, Spain. 2002.

- Xavier Carreras and Isaac Chao and Lluís Padró and Muntsa Padró. FreeLing: An Open-Source Suite of Language Analyzers Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04), 2004.

Contact: Lluís Padró <padro@lsi.upc.edu>

HMM PoS ACOPOST

Authors: Sergio Ferrández y Jesús Peral.

References: GPLSI, Departamento de Lenguajes y Sistemas Infomáticos, Universidad de Alicante: http://gplsi.dlsi.ua.es/mwgplsi/index.php/Portada

Description: PoS Tagger que está basado en el algoritmo de HMM implementado an la herramienta ACOPOST. Está entrenado para el español con el corpus CLIP-TALP.

Functionality: Esta herrramienta, que funciona para el español, requiere como entrada un fichero de texto en lenguaje natural y el fichero de salida. El fichero de salida quedará anotado con las etiquetas PoS correspondientes.

Technology: -

Technical Requirements: -

Modules: -

Innovation: -

Development: -

Publications:

- Sergio Ferrández, Jesús Peral: Investigating the Best Configuration of HMM Spanish PoS Tagger when Minimum Amount of Training Data Is Available. NLDB 2005: 341-344

Contact: Sergio Ferrández <sferrandez@dlsi.ua.es >

IXAti

Authors: IXA group. Propiedad intelectual: SS-78-08

References: http://ixa2.si.ehu.es/demo/zatiak.jsp

Description: Zatiak performs shallow syntactic analysis of a sentence. This program reads an input text and, after morphological processing, identifies pieces of text (chunks). Each chunk is marked with its type: nominal phrase (NP or PP) or verb chain, together with its associated information: grammatical case, number, definiteness and syntactic functions, among others.

Functionality: Chunking for Basque. There is a web service.

Technology: -

Technical Requirements: -

Modules: Modules for entities (Eihera) and postpositions are included. Eustagger is a previous step.

Innovation: It is the first chunker for Basque.

Development: Different projects funded by the Basque government and the Spanish R&D agency.

Publications:

- Aldezabal I., Gojenola K., Sarasola K. A Bootstrapping Approach to Parser Development International Workshop on Parsing Technologies (IWPT2000). Trento. 2000
- Aranzabe M., Arriola J.M., Díaz de Ilarraza A. Towards a Dependency Parser of Basque. Proceedings of the Coling 2004 Workshop on Recent Advances in Dependency Grammar. Geneva, Switzerland. 2004

Contact: Iñaki Alegria <i.alegria@ehu.es>

Jointparser

Authors: Xavier Lluís

References: http://www.lsi.upc.edu/~xlluis/jointparser

Description: Jointparser is a data-driven parser that jointly performs both syntactic dependency parsing and shallow semantic parsing. The system is based on an extension of the Eisner algorithm and uses an online averaged preceptron as a learning method. Shallow semantic parsing is performed for nominal and verbal predicates. The system was presented in the context of the CoNLL-2008 shared task.

Functionality: Noun Phrase and Verbal Phrase identification, joint sintactic and semantic analysis (on-line for english sentences)

Technology: C++, web interface

Technical Requirements: Included svmlight (svmlight.joachims.org).

Modules: -

Innovation: It was one of the two novel joint syntactic-semantic parsers presented at the CoNLL-2008 shared task. It achieved a reasonable performance given it is a built-from-scratch system.

Development: Xavier Lluís master's thesis (UPC 8/9/2008).

Publications:

- Xavier Lluís and Lluís Márquez, A Joint Model for Parsing Syntactic and Semantic Dependencies, Proceedings of CoNLL-2008, 2008.
- Xavier Lluís, advisor: Lluís Márquez, Joint Learning of Syntactic and Semantic Dependencies, Master's thesis, Technical University of Catalonia, 2008.

Contact: Xavier Lluís <xlluis@lsi.upc.edu>

LangIdent

Authors: Muntsa Padró

References: http://www.lsi.upc.edu/~nlp/tools/lang_ident.tar.gz

Description: LangIdent is a Markov-Model based language identifier under GPL license.

Functionality: -

Technology: C++

Technical Requirements: -

Modules: Compiling the program creates a module "idioma.o" you can link to your main program. Two executables are also created: a sample main program that ilustrates the usage of the language identifier and the program to train new models.

Innovation: Comparing the Markov-Model technique to other techniques, it has the best performance. Nevertheless, the usage of this technique for language identification is not new.

Development: In the framework of the ALIADO project.

Publications:

- Muntsa Padró and Lluís Padró. Comparing Methods for Language Identification. Procesamiento del Lenguaje Natural, n. 33, pg. 155--162. September, 2004.

Contact: Muntsa Padró <mpadro@lsi.upc.edu>

Mendekotasunak

Authors: IXA group

References :

Description: A dependency parser we establish the dependency-based grammatical relations (subject, object, modifier, etc.) between the components within the clause in order to obtain a dependency syntactic tree. Such a deep analysis is used to improve the output of the shallow parsing where syntactic structure ambiguity is not fully and explicitly resolved. Previous to the completion of the grammar for the dependency parsing, the design of the Dependency structure-based scheme is accomplished.

Functionality: Dependency parser for Basque

Technology: -

Technical Requirements: -

Modules: Ixati is a previous step.

Innovation: It is the first dependency parser for Basque.

Development: -

Publications:

- Aranzabe M., Arriola J.M., Díaz de Ilarraza. 2004. Towards a Dependency Parser of Basque Proceedings of the Coling 2004 Workshop on Recent Advances in Dependency Grammar. Geneva, Switzerland.

Contact: Iñaki Alegria <i.alegria@ehu.es>

MOSTAS

Authors: Ana Iglesias, Elena Castro, Rebeca Pérez, Leonardo Castaño and Paloma Martínez (Universidad Carlos III de Madrid)

References: Advanced Databases Group (Labda) of Universidad Carlos III de Madrid is a research group with an extensive activity in several Natural Language Processing and Information Retrieval projects: http://basesdatos.uc3m.es/index.php?id=202&L=0

Description: The MOSTAS system (MOrpho-Semantic Tagger, Anonymizer and SpellChecker for biomedical texts) preprocesses Clinical Reports in order to facilitate information retrieval tasks (clinical concepts, abbreviations, entities, etc.). MOSTAS system annotates clinical reports with morpho-semantic information, applies abbreviation and acronyms conversions and detects biomedical concepts using specialized biomedical resources (databases, thesaurus, a multilingual terminology server, etc.). Moreover, MOSTAS is able to anonymize and correct the clinical reports.

Functionality: MOSTAS preprocesses semi-structured information from clinical reports, tagging the clinical texts with specialized information, eliminating sensible information from patients and detecting if there are spellchecker errors.

Technology: The system is implemented in Java, suitable for being installed in Linux and Windows platforms.

Technical Requirements: Currently, STILUS[1] is used as morpho-semantic tagger. Biomedical resources are needed also, as the SNOMED[i] thesaurus, the list of abbreviations and acronyms provided by the Spanish Ministry of Health and Consume[ii] among others. Moreover, Java is necessary for its compilation and execution.

Modules: The system is divided mainly in five different modules. Three of them deal with the pre-processing phase of the clinical reports: the Morpho-semantic Analyzer, the Acronym/Abbreviation Finder and the Biomedical Concept Finder. The other two modules deal with a post-processing of the text: a domain-specific Spell-checker and Anonymizing module were sensible information of the patients is eliminated from the clinical texts.

Innovation: MOSTAS is a complete system that permits to tag clinical texts, anonymize them and detect and correct spellchecker errors. This system works with semi-structured texts in Spanish. Nowadays, most of the research in this area is done in English, so MOSTAS is the first complete system in Spanish. Moreover, MOSTAS implements a biomedical resource similar to English METAMAP for UML but for the Spanish SNOMED thesaurus.

Development: MOSTAS has been developed during the ISSE project[2] and the result of different PHD works in LABDA have been taken into account, as the SPINDEL system.

Publications:

- Ana Iglesias, Elena Castro, Rebeca Pérez, Leonardo Castaño, Paloma Martínez, José Manuel Gómez Pérez, Sandra Kohler y Ricardo Melero. MOSTAS: Un Etiquetador Morfo-Semántico, Anonimizador y Corrector de Historiales Clínicos. XXIV edición del Congreso Anual de la Sociedad Española para el Procesamiento del Lenguaje Natural 2008 (SEPLN´08). Vol. 41, Pp. 299-300.

Contact: Ana Mª Iglesias Maqueda <aiglesia@inf.uc3m.es>

NERUA

Authors: Oscar Ferrandez, Zornitsa Kozareva, Andres Montoyo y Rafael Muñoz

References: GPLSI, Departamento de Lenguajes y Sistemas Infomáticos, Universidad de Alicante: http://gplsi.dlsi.ua.es/mwgplsi/index.php/Portada

Description: NERUA es un sistema de reconocimiento de entidades para el español. Realiza el etiquetado de las entidades en cuatro categorías: PERSONA, LOCALIZACION, ORGANIZACION y MISCELÁNEA (aquellas que no corresponden a ninguna de las categorías anteriores). Para el reconocimiento y la clasificación, NERUA emplea tres algoritmos de aprendizaje automático: Hidden Markov Model, Máxima Entropía y Memory-based learner. Para su aprendizaje y evaluación se utilizaron los recursos proporcionados por la conferencia CoNLL-2002. Además, NERUA tiene la opción de utilizar una combinación de los clasificadores mediante una estrategia de votación simple.

---

1 A grammatical and style checker developed by Daedalus, http://www.daedalus.es

2 FIT-350300-2007-75 (Semantic Interoperability in Electronic Health Care)

Functionality: Etiquetado de entidades nombradas del tipo PERS PERSONA, LOCALIZACION, ORGANIZACION y MISCELÁNEA (aquellas que no corresponden a ninguna de las categorías anteriores) en texto plano.

Technology: Herramienta desarrollada en el lenguaje de programación C++.

Technical Requirements: Este reconocedor de entidades utiliza los siguientes recursos externos: ACOPST (http://acopost.sourceforge.net), TiMBL (http://ilk.uvt.nl/timbl) y MaxEnt, desarrollada por Armando Suárez <armando@dlsi.ua.es>, miembro del GPLSI.

Modules: Consta de dos módulos: uno para la detección y otro para la clasificación de entidades. Development: El desarrollo de la herramienta fue parcialmente financiada bajo los proyectos de investigación nacionales CICyT número TIC2003-07158-C04-01 y PROFIT número FIT-340100-2004-14 y por la Generalitat Valenciana bajo los proyectos GV04B-276 y GV04B-268.

Publications:

- Kozareva, Z; Ferrández O.; Montoyo, A.; Muñoz R.; Suarez A.; Gómez J.; Combining Data-driven systems for improving Named Entity Recognition Año: 2007 Volumen: 61 Número: 3 Páginas: 449-466 Publicación Data & Knowledge Engineering

- Ferrández, O.; Kazareva, Z.;Montoyo, A.; Muñoz, R ; NERUA: sistema de detección y clasificación de entidades utilizando aprendizaje automático Año: 2005 Número: 35 Revista:1135-5948. Procesamiento del Lenguaje Natural

Contact: Oscar Ferrández <ofe@dlsi.ua.es >

SemRol

Authors: Paloma Moreda y Manuel Palomar

References: GPLSI, Departamento de Lenguajes y Sistemas Informáticos, Universidad de Alicante: http://gplsi.dlsi.ua.es/mwgplsi/index.php/Portada

Description: Herramienta basada en corpus para la anotación con roles semánticos de los constituyentes de una oración respecto al verbo. Conjunto de roles: PropBank. Corpus entrenamiento: PropBank.

Functionality: La anotación requiere como entrada un fichero en inglés en lenguaje natural con cada oración que se desee anotar en una línea diferente. Como resultado del proceso se devuelve la anotación de roles en formato Start*End, además de la información proporcionada por otras herramientas utilizadas en el proceso de análisis: morfológico, sintáctico y entidades nombradas.

Technology: Herramienta desarrollada en C y ejecutable desde línea de comandos.

Technical Requirements: Requiere un analizador morfológico (Freeling), sintáctico (D. Roth), reconocedor de entidades (LinPipe) y algoritmos de aprendizaje TiMBL y Máxima entropía.

Modules: Arquitectura estructurada en torno a dos grandes módulos: i) módulo de procesamiento offline: lleva a cabo el proceso de ajuste y selección de la información necesaria para el proceso de anotación; ii) módulo de procesamiento online: realiza la propuesta de anotación.

Innovation: A diferencia de otras herramientas de anotación de roles semánticos, SemRol puede afrontar la anotación desde diferentes perspectivas dependiendo de lo que se desee anotar. Así, es posible llevar a cabo una anotación por sentidos, más adecuada para argumentos numerados; una anotación única, adecuada para adjuntos; una anotación específica para tipos de roles concretos, en el caso de querer anotar sólo lugar, tiempo, etc.; o bien una anotación completa.

Development: La herramienta forma parte de los trabajos realizados dentro de la Tesis doctoral de Paloma Moreda y de los trabajos desarrollados en el proyecto TEST-MESS (TIN2006-15265- C06-01).

Publications:

- Moreda Pozo, P. Los roles semánticos en la tecnología del lenguaje humano: Anotación y Aplicación. Tesis Doctoral. Director: Dr. Manuel Palomar. Julio, 2008.
- Moreda, P., Llorens, H., Saquete E., Palomar M.The influence of Semantic Roles in QA: A comparative analysis. Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN), num. 41, pp. 5562, ISSN 11355948, 2008, Spain.

Contact: Paloma Moreda <paloma@dlsi.ua.es >

SRG: Spanish Resource Grammar

Authors: Montserrat Marimon Felipe, GRIAL, Universitat de Barcelona.

References: http://grial.uab.es/srg

Description: open-source, multi-purpose, large-coverage HPSG grammar for Spanish.

Functionality: The SRG records analyses in three different formats: 1) a derivation tree composed of identifiers of lexical items and constructions used to construct the analysis, 2) a traditional phrase structure tree labeled with atomic labels. e.g. S, NP, VP, etc., and 3) a Minimal Recursion Semantic Representation meaning representation which subsumes the tectogrammatical (functor-argument) structure.

Technology: The SRG is implemented within the Linguistic Knowledge Builder (LKB) system (http://wiki.delph-in.net/moin/LkbTop).

Technical Requirements: The LKB system runs in linux and windows.

Modules: The SRG includes: 1) about 200 phrase structure rules and 50 lexical rules, and 2) a lexicon of about 50,000 entries.

Innovation: The SRG is open-source, multi-purpose, large-coverage and precise (HPSG-based).

Development: The development of the SRG was funded by the Juan de la Cierva program (MEC, Spain) within the TEXTERM-2 project (BFF2003-2111) at the IULA of the UPF. The Current research, which is funded by the Ramon y Cajal program (MICINN, Spain), takes place in the following areas: 1) treebanking , 2) disambiguation , and 3) evaluation .

Publications:

- Montserrat Marimon, Núria Bel, Sergio Espeja and Natalia Seghezzi. The Spanish Resource Grammar: pre-processing strategy and lexical acquisition, in T. Baldwin et al. (eds.) *Proceedings of*

the Workshop on Deep Linguistic Processing, Association for Computational Linguistics (ACL-DLP-2007).

- Montserrat Marimon, Núria Bel and Natalia Seghezzi. Test Suite Construction for a Spanish Grammar, in Tracy Holloway King and Emily M. Bender (eds.) *Proceedings of the Grammar Engineering Across Frameworks (GEAF-2007) Workshop "CSLI Studies in Computational Linguistics ONLINE"*, pp. 250-264. ISSN 1557-5772.

- Montserrat Marimon, Natalia Seghezzi and Núria Bel. An Open-source Lexicon for Spanish, *Procesamiento del Lenguaje Natural*, n. 39, pp. 131-137. Septiembre, 2007. ISSN 1135-5948.

Contact: Montserrat Marimon <montserrat.marimon@gmail.com>

SUPAR

Authors: Antonio Ferrández Rodríguez.

References: GPLSI, Departamento de Lenguajes y Sistemas Informáticos, Universidad de Alicante: http://gplsi.dlsi.ua.es/mwgplsi/index.php/Portada

Description: Es un Sistema de Procesamiento de Lenguaje Natural orientado al análisis sintáctico (completo o parcial) y a la resolución de la anáfora. Éste sistema puede incorporar cualquier etiquetador léxico (POS tagger).

Functionality: El sistema funciona tanto para el español como para el inglés. Para realizar un análisis morfosintáctico, recibe como entrada un fichero de texto en lenguaje natural y el idioma. Como salida se almacena en el fichero de salida el texto analizado morfosintácticamente y con resolución de anáfora pronominal, la cual se etiqueta.

Technology: Sistema desarrollado en Prolog con versiones ejecutables tanto en Windows como en Linux. Dispone de un interfaz gráfico para su funcionamiento interactivo con el usuario, y también se ha desarrollado una versión ejecutable desde la línea de comando.

Technical Requirements: esta herramienta hace uso del postagger Maco para el español y del Tree Tagger para el inglés.

Modules: Necesita de un POS tagger que etiquete léxicamente el texto.

Innovation: Amplia cobertura sintáctica para el español e inglés, permitiendo un análisis sintáctico

equivalente entre varios idiomas (bloques sintácticos similares: SN, SP, Aposiciones, Núcleos

Verbales, etc.). Rapidez de ejecución que permite su uso en grandes corpus, como los empleados en

las competiciones de RI y QA del CLEF y TREC.

Development: Fue resultado de la Tesis Doctoral de Antonio Ferrández Rodríguez, y el cual

evolucionó durante las competiciones CLEF y TREC.

Publications:

- Antonio Ferrández Rodríguez, Manuel Palomar, Lidia Moreno: An Empirical Approach to Spanish Anaphora Resolution. Machine Translation 14(34): 191216 (1999)
- Antonio Ferrández, Manuel Palomar, Lidia Moreno: Anaphor Resolution in Unrestricted Texts with Partial Parsing. COLINGACL 1998: 385391

Contact: Antonio Ferrández Rodríguez <antonio@dlsi.ua.es >

SVM Model for Arabic NER

Authors: Yassine Benajiba (Ph.D. student) and Paolo Rosso

References: http://www.dsic.upv.es/grupos/nle/

Description: A Named Entity Recognition model which is trained using an SVM-based approach over a 125,000 Arabic tokens training file.

Functionality: The model allows the user to extract the named entities with an opn-domain text and classify them into 4 different categories, namely: person, location, organization and miscellaneous. In order to enhance the performance, the model was trained over ATB segmented data which helps to decrease the sparseness in Arabic data.

Technology: The model is trained using Support Vector Machines approach with the Yamcha Toolkit (http://chasen.org/~taku/software/yamcha/).

Technical Requirements: The input file should be ATB segmented and transliterated to Romanized characters. Also it requires Yamcha to be installed in the machine.

Modules: One module which consists of basic decoding on the data provided by the user.

Innovation: To our knowledge, no Arabic NER systems are freely available for the research community. The model has been tested and the results have been presented at EMNLP and ACIT conferences.

Development: Developed as part of Yassine Benajiba's AECI Ph.D. and the MiDES CICYT TIN2006-15265-C06-04 research project, co-funded by the AECI-PCI A01031707 project.

Publications:

- Benajiba Y., Diab M., Rosso P. Arabic Named Entity Recognition using Optimized Feature Sets. In: Proc. Int. Conf. on Empirical Methods in Natural Language Processing, EMNLP-2008, Waikiki, Honolulu, U.S.A., October, 2008
- Benajiba Y., Diab M. Rosso P. Arabic Named Entity Recognition: An SVM-based approach. In: Proc. Int. Arab Conf. on Information Technology, ACIT-2008, Hammamet, Tunisia, December, 2008

Contact: Yassine Benajiba <benajibayassine@gmail.com>

SVMTool

Authors: Jesús Giménez and Lluís Màrquez

References: http://www.lsi.upc.edu/~nlp/SVMTool

Description: A simple, flexible, and effective generator of sequential taggers based on Support Vector Machines. We have appied the SVMTool to the problem of part-of-speech tagging. By means of a rigorous experimental evaluation, we conclude that the proposed SVM-based tagger is robust and flexible for feature modelling (including lexicalization), trains efficiently with almost no parameters to tune, and is able to tag thousands of words per second, which makes it really practical for real NLP applications. Regarding accuracy, the SVM-based tagger significantly outperforms the TnT tagger exactly under the same conditions, and achieves a very competitive accuracy of 97.2\% for English on the Wall Street Journal corpus, which is comparable to the best taggers reported up to date. It has been also successfully applied to Spanish and Catalan exhibiting a similar performance, and to other tagging problems such as base phrase chunking.

Functionality: -

Technology: Perl / C++

Technical Requirements: -

Modules: SVM Learning, SVM Tagger

Innovation: Fast and accurate (state-of-the-art) part-of-speech tagging.

Development: -

Publications:

- Jesús Giménez and Lluís Màrquez. Fast and Accurate Part-of-Speech Tagging: The SVM Approach Revisited. In Proceedings of the International Conference RANLP - 2003 (Recent Advances in Natural Language Processing), pages 158 - 165. September, 10-12, 2003. Borovets, Bulgary. (ISBN 954-90906-6-3). Selected as a chapter in RANLP 2003 volume in CILT series (Current Issues in Linguistic Theory). John Benjamins Publishers, Amsterdam.

- Jesús Giménez and Lluís Màrquez. SVMTool: A general POS tagger generator based on Support Vector Machines. In Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04), vol. I, pages 43 - 46. Lisbon, Portugal, 2004. (ISBN 2-9517408-1-6). Departament Research Report (LSI-04-34-R), Technical University of Catalonia, 2004.

Contact: Jesús Giménez <jgimenez@lsi.upc.edu>

SwiRL

Authors: Mihai Surdeanu

References: http://www.surdeanu.name/mihai/swirl/

Description: SwiRL is a Semantic Role Labeling (SRL) system for English constructed on top of the full syntactic analysis of text. Achieved state-of-the-art performance in the CoNLL 2005 SRL evaluation.

Functionality: -

Technology: C++

Technical Requirements: Charniak Parser (included), Xavi Carreras' AdaBoost library (included).

Modules: Semantic Role Labeler

Innovation: It has state-of-the-art performance: currently its F1 on the WSJ corpus is 77+, and on the Brown corpus it is 66+ points. SwiRL ranks fifth among the systems that participated at the CoNLL 2005 shared task evaluation, but all the systems that scored higher were actually combinations of several individual models.

Development: -

Publications:

- Mihai Surdeanu and Jordi Turmo, "Semantic Role Labeling Using Complete Syntactic Analysis", CoNLL Shared Task 2005.

Contact: Mihai Surdeanu <mihai@surdeanu.name>

The DrugDDI Extractor System

Authors: Isabel Segura-Bedmar, Paloma Martínez, César de Pablo-Sánchez, Daniel Sánchez

References: Advanced Databases Group (Labda) (http://labda.inf.uc3m.es/) of Universidad Carlos III de Madrid is a research group with an extensive activity in several Natural Language Processing, Information Retrieval and Information Extraction projects.

Description: We have developed a system that combines IE techniques. In particular, we have proposed two different approximations for the extraction of DDIs from texts. The first approximation proposes a hybrid approach, which combines shallow parsing and pattern matching to extract relations between drugs from biomedical texts. A pharmacist defined a set of lexical patterns (12) to capture the various language constructions used to express DDIs in pharmacological texts. The second approximation is based on a supervised machine learning approach, in particular, a kernel-based approach that uses Support Vector Machines (SVM). While the first approximation based on pattern matching achieves low performance (Precision=48.7\%, Recall=25.7\%, F-measure=33.6\%), the approach based on kernel-methods achieves better performance, especially better recall (Precision=55.1\%, Recall=82.3\%, F-measure=66.0\%). A web tool can be found at http://163.117.129.57:8080/ddiextractorweb).

Functionality: Búsqueda de documentos en la colección MedLine 2010 y procesamiento de textos para el reconocimiento de fármacos y sus interacciones.

Technology: DrugDDIExtractor combines several IE techniques (described in the previous paragraph) and has been implemented using Java. The web tool is written in JSP. Also, this tool integrates Apache Lucene for supportin the search of documents from Medline 2010 collection.

Technical Requirements:

Modules:

Innovation: This is the first system to extract drug names and drug-drug interactions from biomedical texts.

Development: This system was part of the thesis “Application of information extraction techniques to pharmacological domain: extracting drug-drug interactions” Isabel Segura-Bedmar, Advisor: Paloma Martínez. Recently, this thesis has been granted with the Extraordinary PhD award 2011. This work has been

partially supported by the Spanish research projects: MA2VICMR consortium (S2009/TIC-1542, www.mavir.net), a network of excellence funded by the Madrid Regional Government and TIN2007-67407-C03-01 (BRAVO: Advanced Multimodal and Multilingual Question Answering).

Publications:

- Isabel Segura-Bedmar, Paloma Martínez, César de Pablo-Sánchez, (2011). A linguistic rule-based approach to extract drug-drug interactions from pharmacological documents, January, 2011, BMC BioInformatics, ISSN: 1471-2105, Volumen: In Press.
- Isabel Segura-Bedmar, Paloma Martínez, César de Pablo-Sánchez, (2010). Extracting drug-drug interactions from biomedical texts., May, 2010, BMC BioInformatics, ISSN: 1471-2105, Volumen: 11, Número: Suppl 5.
- Isabel Segura-Bedmar, Paloma Martínez, César de Pablo-Sánchez, (2010). Combining syntactic information and domain-specific lexical patterns to extract Drug-Drug Interactions from biomedical texts., Toronto, Canada, October, 2010, ACM Fourth International Workshop on Data and Text Mining in Bioinformatics (DTMBIO 10)., ACM.
- Isabel Segura-Bedmar, Paloma Martínez, María Segura-Bedmar, (2008). Drug Name Recognition and classification in biomedical texts. , September, 2008, Drug Discovery Today, Elsevier Science, ISSN: 1359-6446, Volumen: 13, Número: 17-18, Páginas: 816-823, url.

Contact: Isabel Segura-Bedmar (isegura@inf.uc3m.es), Paloma Martínez (pmf@inf.uc3m.es)

TIPSem

Authors: Hetor Llorens, Estela Saquete, Borja Navarro-Colorado.

References: GPLSI. Universitat d'Alacant. Demo: http://gplsi.dlsi.ua.es/demos/TIMEE/

Description: TIPSem is a temporal information processing tool for English and Spanish. This is based on semantics (lexical and compositional), in addition to morphosyntax.

Functionality: Given a raw text, this outputs its TimeML annotation (http://timeml.org). TIPSem performs the following processes:

- temporal expression recognition, classification and normalization
- event recognition and classification
- temporal relation identification and categorization

Technology: The tool is programmed in Java but is currently offered as an online application in the following URL: http://gplsi.dlsi.ua.es/demos/TIMEE/

Technical Requirements: A standard web browser and Internet connection.

Technical Requirements: A standard web browser and Internet connection.

Modules:

Innovation: TIPSem makes a extensive usage of lexical semantics (WordNet and EuroWordNet) and compositional semantics (semantic roles) to perform the temporal information processing and achieves a competitive performance in comparison with the state-of-art approaches. It is the only available multilingual tool (including a complete temporal information processing in Spanish).

Development: This was part of the development of a thesis supported by the Spanish Government, projects TIN-2006-15265-C06-01, where Hector Llorens is funded under a FPI grant (BES-2007-16256).

Publications:

- Llorens, H., Saquete, E., & Navarro-Colorado, B. (2010b). TIPSem (English and Spanish): Evaluating CRFs and Semantic Roles in TempEval-2. In Proceedings of the 5th International Workshop on Semantic Evaluation, pp. 28491. ACL.
- Llorens, H., Saquete, E., & Navarro-Colorado, B. (2010a). TimeML Events Recognition and Classification: Learning CRF Models with Semantic Roles. In Proceedings of the 23rd International Conference on Com-putational Linguistics (Coling 2010), pp. 72533.

Contact: Hector Llorens (hllorens@dlsi.ua.es), DLSI, Universitat d'Alacant

WaCOS: Watermarking Corpora Online System

Authors: David Pinto

References: http://www.dsic.upv.es/grupos/nle, http://nlp.dsic.upv.es:8080/watermarker,

http://nlp.cs.buap.mx/watermarker

Description: The Watermarking Corpora On-line System (WaCOS) is made up of a set of measures for the assessment of text corpora.

Functionality: WaCOS allows linguistics and computational linguistics researchers to study the following corpus features: *domain broadness*, *shortness*, *class imbalance, stylometry* and *structure*. WaCOS provides a friendly interface in order to easily evaluate corpora.

Technology: WaCOS front-end has been programmed with PHP. It integrates a set of modules written in different programming languages (C, C++, Java, AWK). Among the several components of this system, it uses n-gram language modelling, Zipf distribution of frequencies, density-based measures, internal clustering validity measures, etc in order to assess the relative hardness of a given corpus.

Technical Requirements: The end user is only required of an Internet browser in order to access the on-line system.

Innovation: A freely available web-based tool which may be used to study peculiarities of textual corpus features.

Development: Developed as part of David Pinto's Ph.D. and the MiDES CICYT TIN2006-15265-C06-04 research project.

Publications:

- David Pinto: On Clustering of Narrow Domain Short-Text Corpora. PhD Thesis, Universidad Politécnica de Valencia, Spain, July 2008.

- Diego Ingaramo, David Pinto, Paolo Rosso, Marcelo Errecalde: Evaluation of Internal Validity Measures in Short-Text Corpora. CICLing 2008. Lecture Notes in Computer Science 4919, Springer-Verlag: 555-567, 2008

- Rafael Guzman, Manuel Montes, Paolo Rosso, Luis Villaseñor-Pineda and David Pinto: Semi-supervised Approach for WSD using the Web as Corpus. CICLing 2009. Lecture Notes in Computer Science, Springer-Verlag, 2009

Contact: David Eduardo Pinto Avendaño <dpinto@cs.buap.mx>

WSD-IXA

Authors: IXA group and Asier Gabiola

References: http://ixa3.si.ehu.es/wsd-demo

Description: The WSD system is based on the well known Support Vectors Machine (SVM) Algorithm. This system has been trained on EuSemCor corpus (the unique basque corpus semantically tagged). Due to corpus's reduced size, the WSD system has been trained for 402 polysemous nouns.

Functionality: Perl CGI script runs the input raw text over Eustagger basque lemmatizer in order to extract features. Then, the feature-vector is classified by the WSD (SVM) system. Finally, the CGI manage classifier and lemmatizer output in order to show in a proper format.

Technology: C, C++, Perl

Technical Requirements: -

Modules: Perl CGI script, EusSemcor data base (MySql), Eustagger, SVM-light

Innovation: First online WSD system for Basque

Development: -

Publications:

- Agirre E., and Martinez D.2004. The Basque Country University system: English and Basque tasks..Proceedings of the 3rd ACL workshop on the Evaluation of Systems for the Semantic Martinez D. 2005

- Supervised Word Sense Disambiguation: Facing Current Challenges. SEPLN Journal. Vol. 34. pgs 125-126.

Contact: Iñaki Alegria <i.alegria@ehu.es>

WSD-UA

Authors: Andrés Montoyo y Manuel Palomar

References: GPLSI, Departamento de Lenguajes y Sistemas Informáticos, Universidad de Alicante: http://gplsi.dlsi.ua.es/mwgplsi/index.php/Portada

Demo disponible en: http://www.dlsi.ua.es/projectes/srim/desambiguacion.html

Description: Esta aplicación utiliza el método de marcas de especificidad para el tratamiento de la desambiguación de textos. Asigna sentidos a las palabras según el diccionario electrónico EuroWordNet, tanto en español como en inglés.

Functionality: Esta herramienta funciona tanto para el español como para el inglés. Como entrada, el sistema recibe el fichero a procesar, el fichero dónde se guardará la salida, el idioma y el sufijo.

Technical Requirements: -

Innovation: -

Development: Esta herramienta fue desarrollada dentro del proyecto PROFIT "Construcción de un sistema de recuperación de información multilingüe en la web" (FIT-150500-2002-416), subvencionado por el antiguo Ministerio de Ciencia y Tecnología. Fue creada en la Universidad de Alicante por Andrés Montoyo con la colaboración de May Calle y Sonia Vázquez.

Publications:

- Andrés Montoyo y Manuel Palomar. Word Sense Disambiguation with Specification Marks in Unrestricted Texts. Páginas 103–107 de: DEXA Workshop. IEEE Computer Society. 2002.
- Andrés Montoyo, Armando Suárez, Manuel Palomar: Combining Supervised-Unsupervised Methods for Word Sense Disambiguation. CICLing 2002: 156-164

Contact: Andrés Montoyo: montoyo@dlsi.ua.es

XIADA (Etiquetador/lematizador del gallego actual)

Authors: Guillermo Rojo, Manuel Vilares Ferro, Eva Domínguez Noya, María Sol López Martínez, Francisco García Gondar, Fco. Mario Barcala Rodríguez, Miguel Angel Molinero Álvarez, Jorge Graña Gil y Miguel Ángel Alonso Pardo

References: http://corpus.cirp.es/xiada/

Description: Etiquetador y lematizador de textos escritos en lengua gallega

Functionality: En el año 2003 se terminó una primera versión operativa del etiquetador que trabajaba con archivos de texto que se adecuaban a la normativa. Además, para obtener esta versión fue necesario el desarrollo del juego de etiquetas apropiado (cuenta con alrededor de 400 etiquetas diferentes), de un lexicón formado por aproximadamente 31.200 léemelas y 630.000 elementos gramaticales y de un subcorpus anotado de entrenamiento de unas 100.000 formas ortográficas.

En 2005 se externalizaron las reglas de funcionamiento, facilitando así la actualización y/o modificación de las mismas por parte del equipo de desarrollo.

Durante 2006 se adaptó el etiquetador para que pudiera trabajar con archivos codificados en XML y, por lo tanto, con los archivos de la nueva codificación de los documentos del CORGA. También en este año se

desarrolló un sistema genérico de resolución de ambigüedades segmentales, se amplió considerablemente el lexicón que utiliza, que además incluye muchas formas no normativas para que puedan ser reconocidas.

En 2007 se publicó el etiquetario utilizado por el proyecto.

En 2009 se hace pública una demostración del funcionamiento del etiquetador y se libera la primera versión del léxico (2.2) y del corpus de entrenamiento que utiliza (2.3). Este último incluye 309.505 elementos gramaticales.

En 2010 se publica la versión 2.4, que incluye la liberación de una nueva versión del léxico, con 718.189 entradas y 53.888 lemas(427 lemasmás que en la versión anterior); la del corpus de entrenamiento, con 426.051 elementos gramaticales y, por último, la de la demostración del etiquetador entrenado con estos nuevos recursos. También se actualiza en la web el etiquetario que emplea el etiquetador (383 etiquetas) y se recopilan ejemplos de uso de cada etiqueta.

En 2013 se publica la versión 2.5. El léxico posee 730.256 entradas, añadiéndose respecto a la versión anterior 4.463 nuevos lemas. El corpus de entrenamiento se completa con texto extraído de colecciones de relato corto y pasa a constar de 594.993 elementos gramaticales.

Technical Requirements: -

Modules: -

Innovation: Primer etiqeutador/lematizador públicamente disponible para gallego.

Development: -

Publications:

- Practical application of one-pass Viterbi algorithm in tokenization and part-of-speech tagging, Miguel A. Molinero, Fco. Mario Barcala, Juan Otero, Jorge Graña. Proc. of International Conference RANLP 2007, Recent Advances in Natural Language Processing, Borovets, Bulgaria, 2007, pp. 35-40.

- XML rules for enclitic segmentation, Fco. Mario Barcala, Miguel A. Molinero, Eva Domínguez. Computer Aided Systems Theory – EUROCAST 2007, Revised Selected Papers, Lecture Notes in Computer Science, 4739 Springer-Verlag, Berlin-Heidelberg-New York, 2007, pp. 273-281.

- Automatic Spelling Correction in Galician. Manuel Vilares, Juan Otero, Fco. Mario Barcala, Eva Domínguez. José Luis Vicedo, Patricio Martínez-Barco, Rafael Muñoz and Maximiliano Saiz Noeda (eds.), Advances in Natural Language Processing, volume 3230 of Lecture Notes in Artificial Intelligence, Springer-Verlag, Berlin-Heidelberg-New York, 2004, pp. 51-57.

- Formal Methods of Tokenization for Part-of-Speech Tagging. Jorge Graña, Fco. Mario Barcala, Jesús Vilares. Alexander Gelbukh (ed.), Computational Linguistics and Intelligent Text Processing, volume 2276 of Lecture Notes in Computer Science, Springer-Verlag, Berlin-Heidelberg-New York, 2002, pp. 240-249.

- A Common Solution for Tokenization and Part-of-Speech Tagging: One-Pass Viterbi Algorithm vs. Iterative Approaches, Jorge Graña, Miguel A. Alonso, Manuel Vilares. Petr Sojka, Ivan Kopecek and Karel Pala (eds.), Text, Speech and Dialogue, volume 2448 of Lecture Notes in Artificial Intelligence, Springer-Verlag, Berlin-Heidelberg-New York, 2002, pp. 3-10.

Contact: Andrés Montoyo <montoyo@dlsi.ua.es>

3. Sistemas para tareas específicas

3.1 Asistentes y Sistemas de Diálogo

Asistente Virtual Semántico

Autores: Juan Llorens, Jose Antonio Moreiro, Jose Miguel Fuentes, Jorge Morato Lara, Sonia Sánchez Cuadrado, Anabel Fraga, Valentín Moreno, George Andreadakis, Vicente Palacios, Mónica Marrero, Karina Robles

Referencias: http://www.asistentevirtual.es/Asistente_Virtual_Semantico.html

Descripción: Se ha desarrollado un asistente virtual basado en semántica y procesamiento del lenguaje natural para proporcionar atención de un sitio Web. Sus principales características son el uso del contexto, anticipación de consultas frecuentes y solución a consultas fuera del dominio de interrogación.

Funcionalidad: El asistente virtual desarrollado, es un agente conversacional representado por un avatar configurable, que puede adaptarse a diferentes áreas temáticas. Existen tres tipos de consultas a las que puede responder: cultura general, preguntas temáticas de respuesta estructurada, preguntas temáticas de respuesta textual. Además es capaz de utilizar el contexto de preguntas previas y analizar documentos para identificar posibles respuestas.

Tecnología: se ha desarrollado en la plataforma DotNet, utilizando como gestor de bases de datos SQL Server.

Requisitos técnicos: 2GB de memoria; 100 MB en disco; windows server, IIS y SQLServer

Módulos: servicio AVS (Asistente virtual semántico), AVS web service, audio server, cliente (Windows u otros)

Innovación: Mecanismos para su adaptación a intranets: 1) Uso de ontologías para expansión de las consultas ; 2) Uso de datos localizados en bases de datos relacionales como respuestas a preguntas , y 3) Uso de listados de preguntas-respuesta (FAQ's)

Development: desarrollo propio

Publicaciones:

- Sonia Sánchez-Cuadrado, Jorge Morato, Vicente Palacios-Madrid, Juan Llorens-Morillo y José Antonio Moreiro-González. De repente, ¿todos hablamos de ontologías? El profesional de la Información, Noviembre-diciembre 2007, vol. 16, núm. 6 .
- Sánchez Cuadrado, Sonia; Morato, Jorge; Moreiro, José Antonio; Marrero, Mónica. Definición de una metodología para la construcción de sistemas de organización del conocimiento a partir de un corpus documental en lenguaje natural. Revista de la SEPLN, 39 septiembre 2007, p. 213-220. XXIII Congreso de la Sociedad Española para el Procesamiento del Lenguaje Natural. Universidad de Sevilla 10, 11, 12 de septiembre de 2007

Flexible Dialogue System

Authors: Marta Gatius, Meritxell González.

References: http://www.lsi.upc.es/~nlp/disi

Description: This is a mixed-initiative Dialogue System designed to guide the user when accessing the web. It has been designed to access different types of web services and information. The main goal in the system design have been its adaptability to the different types of web contents and different types of users and different languages. The Dialogue System consists of four main modules: the Natural Language Analyzer, the Dialogue Manager, the Language Generator and the Services' Access. The Dialogue System is focused in the development of a platform for accessing Web Services. The communication channel can be either textual or by voice, using natural language. The aim of our Dialogue System is to facilitate the integration of new applications into the platform. Then, the development of application specific resources should be easy and able to do by non skilleds.

Functionality: Textual or spontaneous speech access to Web services (currently, Spanish on-line access to cultural agenda and large objects collection services).

Technology: The Dialogue System is developed in J2EE, nevertheless some modules additionally uses other technologies. The Natural Language Analyzer has been developed in Prolog, more suitable for this task. All the resources are written in XML files. Some of them use standards, such as VoiceXML, SRGSXML and OWL. Standards are used when possible. Additionally, we have developed specific XML Schemas for writing the rest of the resources, for example the communication between the Prolog subsytem and the Natural Language Analyzer module, the linguistic resources for the Language Generator, and the Dialogue Plans.

Technical Requirements: Basic System: The basic Dialogue System uses Textual Web Chat Client, runnin on Tomcat Server. The rest of the System runs JVM plus Swi-Prolog Server. More specific technical requirements depends on the Application integrated in the System. Our current Services are also Java Applications using XML files containing the database.

Voice-added System: The Voice subsystem is telephone based, and needs specific hardware and software. The HOPS System uses Loquendo(TM) VoiceXML platform plus a telephone card. Our current system doesn't allow voice access to the platform, nevertheless the linguistic resources are developed taking into account both channels.

Modules: For dialogue management we developed an independent module, the Dialogue Manager (DM). It enables mixed-initiative conversations and it is mode, language and domain independent. It follows the issue-based dialogue management model described in (Larsson, 2002), which describes dialogues in terms of issues that are raised and that have to be solved. The Dialogue Manager uses plans for guiding the conversation through the Operations of the Service. We have developed general recipes for creating plans and resources for new applications. Those recipes are classified into three types: SEARCH, QUERY and TRANSACTION.

The Dialogue System incorporates a Natural Language Generator (NLG) component to generate the system's messages. In order to obtain the most appropriate system prompts for a specific service, the generator component uses a syntactic-semantic taxonomy which relates the specific domain concepts to the linguistic structures needed for their expression. The system messages are not generated at run-time but when adapting the DS to a new web service. Several of the messages generated contain variables that would be instantiated at run-time, considering the dialogue context. When the DS is adapted to a new sevice, the service task parameters have to be classified according to the syntactic-semantic taxonomy and linked to the corresponding lexical entries. Then, the system prompts are generated automatically. However, the resulting sentences can be supervised and selected manually.At run-time the NLG would complete these messages considering dialogue context. Additionally, when system intervention includes several communication acts, the messages expressing this acts are combined dynamically (following a fixed order): greet moves first, then confirmation moves, then answer moves, then asking moves, and finally quit moves.

Textual inputs are processed by a left-corner parser performing syntactic and semantic analysis in parallel has been adapted for a practical dialogue system for multiple applications. The parser uses domain-restricted grammars and lexicons obtained from ontologies representing application specific knowledge. It also uses the knowledge of the dialogue context to select the grammar rules related to the dialogue focus.

The current working System uses XML based data sources for the simulating the Services Access. The ongoing work includes the use of OWL and RDFS for enabling the online access to Web Services as well as the Standarization of the resources.

An ontology-based representation is used for sharing data among modules, as well as for specific modules resources. This representation facilitates the integration of new applications into the system.

Innovation: One of the main goals of this project has been the improvement of existing commercial Dialogue Systems by combining practical results from various research areas: Speech, Natural Language, Dialogue Management and Ontologies.

Most recent works on spoken DSs in the commercial area are based on exploiting the VoiceXML language. But, it also presents several limitations, for example, it presents some dialogue management drawbacks, as the limited support to the user-initiative interactions.

Our work has been focused on the study of how Dialogue Systems, and in particular VoiceXML platforms, could be improved by using language, dialogue and ontology techniques. We also have work on techniques for adapting the interaction to the user expertise. We studied how it improves the communication process. Finally, we are currently working on improved plans and recipes representation. The aim of our current direction is to enable more sophisticated dialogue management, without losing the simplicity in the specific application resources.

Development: The Dialogue System was developed in the context of the HOPS project (IST-2002-507967).

Publications:

- Meritxell González. Dialogue Management for Multilingual communication through different channels. DEA and PT, LSI Department, UPC. Barcelona, July 2007

- Marta Gatius and Meritxell González. Discourse Management in Voice Systems for Accessing Web Services. Workshop on the Semantics and Pragmatics of Dialogue. DECALOG 2007. Rovereto, May 2007.

- Marta Gatius, Meritxell González and Elisabet Comelles. An Information State-Based Dialogue Manager for Making Voice Web Smarter. 16th International Worl Wide Web Conference. WWW2007. Banff, May 2007.

- Marta Gatius, Meritxell González, Eli Comelles and Leonardo Lesmo. Natural Language Processing and Dialogue Management Development. European Project HOPS (IST-2002-507967). Deliverable D4.3. April 2007.

- Marta Gatius and Meritxell González. A multilingual Dialogue System for Accessing the WEB. 3rd International Conference on Web Information Systems and Technologies. WEBIST 2007. Barcelona, March 2007.

- Pablo Hernández, Jordi Sánchez, Ángel López, Sheyla Militello, Marta Gatius, Meritxell González, Eli Comelles, Leonardo Lesmo, Xavier Noria, Robert Salla, Carlos de la Morena, Jose Antonio Fernández, Alberto Deiro. HOPS Architecture Specifications. European Project HOPS (IST-2002-507967). Deliverable D4.1. May 2006.

- Marta Gatius and Meritxell González. Integrating Semantic Web and Language Technologies to Improve the Online Public Administrations Services. 15th International Worl Wide Web Conference. WWW2006. May 23-26, 2006, Edinburgh, Scotland. ACM 1-59593-323-9/06/0005.

- Marta Gatius and Meritxell González. Using Application-Specific Ontologies to Improve Performance in a Bottom-up Parser. Workshop Knowledge and Reasoning for Language Processing, KRAQ'06 . 11th Conference of the European Chapter of the Association for the Computacional Linguistics. EACL'06. Trento, Italy, April 2006. Association for the Computacional Linguistics, ISBN, 1-932432-59-0, pp. 12-19.

- Marta Gatius and Meritxell González. Obtaining Linguistic Resources for Dialogue Systems from Application Specifications and Domain Ontologies. 10th International Conference on Speech and Computer, SPECOM 2005. University of Patras, October 2005

- Meritxell González, Marta Gatius. Un sistema de diálogo multicanal para acceder a la información y servicios de las administraciones públicas. XXI Congreso de la SEPLN, I Congreso Español de Informática, CEDI. Granada , Septiembre 2005

- Marta Gatius, Meritxell González. The project HOPS: Enabling an Intelligent Natural Language Based Hub for the Deployment of Advanced Semantically Enriched Multi-channel Mass-scale Online Public Services. XXI Congreso de la SEPLN, I Congreso Español de Informática. Granada , Septiembre 2005

- Marta Gatius, Meritxell González. Un sistema de diálogo multilingüe dirigido por la semántica. Revista de la SEPLN, Vol.34. Junio 2005.

- Marta Gatius, Meritxell González, Leonardo Lesmo, Pietro Torasso. Natural Language Processing Technologies. European Project HOPS (IST-2002-507967). Deliverable D3.2. February 2005.

- Marta Gatius, Meritxell González. Using Ontologies for Improving the Communication Process in a Dialogue System. Proceedings of the Sixth International Workshop on Computational Semantics, IWCS-6. Tilburg, 2005

- Marta Gatius, Meritxell González. Utilización de ontologías en el desarrollo de sistemas de diálogo. III Jornadas en Tecnología del Habla. Valencia, Noviembre 2004

- Marta Gatius, Meritxell González. Ontology-driven VoiceXML Dialogues Generation. Berliner XML-Tage 2004. Humboldt University Berlin, October 2004

Contact: Marta Gatius <gatius@lsi.upc.edu>, Meritxell González <mgonzalez@lsi.upc.edu>

INTERACTOR (Natural Interaction Platform)

Authors: Paloma Martínez, head of the Advanced Databases Laboratory (Labda). F. Javier Calle, Dolores Cuadra, Senior Researchers. David del Valle, Jessica Rivero, Junior Researchers.

References: Advanced Databases Group (Labda): http://labda.inf.uc3m.es/doku.php?id=es:labda_lineas#

Description: Interactor is an interaction platform based on Natural Interaction (human-like) techniques. It enables to implement a corpus-based Task Oriented Interaction Domain with little effort, and thus it assumes an application and provides access to it through Natural Interaction.

Functionality: Once implanted onto a set of tasks, Interactor receives user interventions (represented through semantic structures) and handles the interaction in a human-like way. When required, it invokes the

execution of some task(s), which are applications or drivers accessible from the server, and feeds the interaction back with its (their) results. Finally, it constructs system's interventions (represented through semantic structures) and provides them. Besides, its Situation Model gathers knowledge on spatio-temporal features and objects within the interaction domain, and it is able to receive and process information about the user's situation.

Technology: Interactor dialogue management is composed of four agents implemented in Java, running in Ecosystem, which provide basic services of agency registration, agent communication, and brokering, among others. The situation component is implemented trough two agents (also in Java) and based on a spatiotemporal database. The rest of the Interactor system (other agents) is also implemented in Java. All the mentioned databases are set on the Oracle TM 11g DBMS.

Technical Requirements: Interactor is currently running on a Sun Fire X4500 server, which also has the DBMS server, under Windows XP. However, due to the flexibility of the system architecture, the DBMS can be set on another server (also with different OS), and even each individual agent might run in different computers (with access to the DBMS server). Thus, system efficiency can be boosted as required.

Modules: For interaction performance, the following modules are required (1) Ecosystem (Multi-agent platform); (2) Interactor components; (3) Dialogue management agents; (4) Situation agents; (5) User model agent; and (6) Ontology agent. For following the internal processes and reasoning of the different agents, it is also required the (7) Tracer, a tool for monitoring the interaction. Finally, implementing new interaction domains involves corpus analysis tasks for which another tool is utilized, (8) the Cognos Toolkit (including relaxed grammars supported NLP and pragmatic analysis and annotation). This Toolkit is available for free use at the following URL: http://labda.inf.uc3m.es/doku.php?id=es:labda_lineas:cognos

Innovation: Very few interaction systems count on a Situation Model (few prototypes, none commercial). This enriches interactive reasoning with the circumstantial aspect, apart from the situational services it can provide. In addition, this Situation Model is empowered by spatio-temporal database technology, ensuring versatility, scalability and efficiency.

Development: Interactor and the Threads Model were the result of a PhD dissertation in 2005, and developed from the experience gained in several funded European and National research projects.

Publications:

- Calle, J., Martínez, P., Valle, D., Cuadra, D. Towards the Achievement of Natural Interaction. Engineering the User Interface: from Research to Practice. © 2009 Springer (ISBN: 978-1-84800-135-0).

- Calle, J., García-Serrano, A., Martínez, P. Intentional Processing as a Key for Rational Behaviour through Natural Interaction. Interacting With Computers (ISSN: 0953-5438), Vol. 18/6, 1419—1446. © 2006 Elsevier.

- Cuadra, D., Calle, F.J., Rivero, J., Valle, D. Applying Spatio-Temporal Databases to Interaction Agents. International Symposium on Distributed Computing and Artificial Intelligence 2008 (DCAI'08). Volume: 50. Pag: 536-540. © 2009 Springer Berlin / Heidelberg (1615-3871).

Contact: F. J. Calle <fcalle@inf.uc3m.es>

3.2 Buscadores

Arabic JIRS

Authors: Yassine Benajiba (Ph.D. student), José Manuel Gomez (Universidad Alicante) and Paolo Rosso

References: http://www.dsic.upv.es/grupos/nle/

Description: Arabic JIRS is an adaptation of the JIRS system to the Arabic language. It is a passage retrieval system for Arabic texts which return a set of relevant passages to the user's query (which is written in Arabic).

Functionality: The Arabic JIRS, indexes Arabic documents and then provides an interface in order to extract passages which are relevant to the user's query.

Technology: JIRS is fully built on JAVA and adopts a server/client framework. It uses n-gram language modelling for indexation and an a distance density based approach to rank the retrieved passages from the most to the least relevant.

Technical Requirements: The machine should be able to run JAVA programs.

Modules: A server module which indexes the document and a client module to send the query.

Innovation: To our knowledge, no Arabic PR systems are freely available for the research community.

Development: Developed as part of Yassine Benajiba's AECI Ph.D. and the MiDES CICYT TIN2006-15265-C06-04 research project, co-funded by the A706706 project.

Publications:

- Benajiba Y., Rosso P., Gómez J.M. Adapting JIRS Passage Retrieval System to the Arabic. In: Proc. 8th Int. Conf. on Comput. Linguistics and Intelligent Text Processing, CICLing-2007, Springer-Verlag, LNCS(4394), pp. 530-541, 2007

Contact: Yassine Benajiba <benajibayassine@gmail.com>

FlickLing

Authors: Víctor Peinado, Javier Artiles, Fernando López-Ostenero, Julio Gonzalo

References: NLP Group at UNED: http://nlp.uned.es

Description: FlickLing is a multilingual search interface for Flickr designed and implemented for the CLEF 2008 interactive task.

Functionality: FlickLing consists of two interfaces which allow to perform monolingual and multilingual image retrieval over the Flickr database, retrieving results with images annotated in different languages. From a given query entered by the user, FlickLing performs automatic term-by-term translation into up to six languages, and provides assistance for interactive query refinement and translation. Since it has been design with the goal of collecting a large search log, it works as an online competitive game, where users have to find as many images as possible to obtain the highest individual and team scores.

Modules: FlickLing search engine consists of two basic components: 1) the graphical user interface which controls the search engine features (queries, ranking of results, etc.), functionalities related to cross-linguality (translations, suggestions, related terms, etc.) and the game-like features of the task (flow of images, users ranking, etc.); 2) a set of web services working behind which are in charge of accessing Flickr, managing the system's databases and generating user logs.

Technology: The graphical user interface is implemented in Java using Google Web Toolkit (GWT) and makes extensive use of the GWT-Ext library. The web services are implemented in Python using the TurboGears web application framework. Databases are stored in a MySQL server.

Technical Requirements: Most of the Flickling functionalities can be accessed through any modern web browser (Firefox 2 or greater).

Innovation: FlickLing is designed as an online competitive game. It allows to perform monolingual and multilingual searches in up to six different languages over the Flickr image collection. The systems captures users' interactions and generates extensive user logs, which are freely available at: http://nlp.uned.es/iCLEF/2008/iclef2008.logs.zip.

Publications:

- V. Peinado, J. Artiles, J. Gonzalo, E. Barker, F. López- Ostenero. FlickLing: a Multilingual Search Interface for Flickr., Working Notes for the CLEF 2008 Workshop. 2008.
- P. Clough, J. Gonzalo, J. Karlgren, E. Barker, J. Artiles, V. Peinado . Large-Scale Interactive Evaluation of Multilingual Information Access Systems - the iCLEF Flickr Challenge. Workshop on Novel Methodologies for Evaluation in Information Retrieval. 30th European Conference on Information Retrieval (ECIR 2008). 2008.

Contact: Víctor Peinado <victor@lsi.uned.es>, Julio Gonzalo <julio@lsi.uned.es>

FlickrBabel

Authors: Francisco Carrero, José Carlos Cortizo, Borja Monsalve (Wipley, Social Gaming Platform)

References: English version: http://www.flickrbabel.com/en,

Spanish version: http://www.flickrbabel.com/es/

Description: Sometimes, when you're searching through Flickr, it's better to search in multiple languages in order to get more results. Since Flickr is a site whose popularity has spread throughout the world, millions of users upload picture titles and tags in their native tongue. This site will allow you to insert a search query for Flickr, and it will search for you in multiple languages. This greatly broadens the scope of results you'll get. For instance, search for the word "car", and the site will search for it in many different languages, allowing you to see pictures of a car someone form Turkey might have uploaded, which you might otherwise not have been able to find. A simpler way to look at it is as a mash-up of Flickr search and Google's translation services.

Functionality: FlickrBabel is a images search engine that retrieves images from Flickr that matches the user's query. FlickrBabel first expands the query by adding the translation of the query to 5 languages (English, Spanish, German, French and Portuguese) and then sends the expanded query to Flickr and shows the images retrieved. FlickrBabel also shows a map for the geolocalized images and allows the users to restrict the results to a given location (city, area, country, etc.)

Technology: FlickrBabel is developed in PHP and makes use of Google Translation and Flickr APIs.

Technical Requirements: The system's needs are a Web Server with PHP support, constant Internet access and a developer's key to Flickr API.

Modules: There are 4 main modules: 1) query pre-processing and expansion, 2) search process, 3) visualization and 4) geolocation restrictions.

Innovation: There are some R&D projects related to multilingual multimedia retrieval but there are no user-friendly tools that implements the techniques related to those projects. FlickrBabel is a simple but user oriented solution.

Development: FlickrBabel is the first product of Social Gaming Platform, first created as a experiment to test Google Translation and Flickr APIs, but finally implemented as a mainstream Internet application.

Publications: There are no scientific references that directly explains FlickrBabel, but part of the work behind FlickrBabel has been applied in:

- Carrero, F., Cortizo Pérez, J. C., Gómez Hidalgo, J. M., Testing Concept Indexing in Crosslingual Medical Text Classification. Proceedings of the IEEE ICDIM 2008

Contact: José Carlos Cortizo Pérez <josecarlos.cortizo@wipley.com>

IR-n

Authors: Fernando Llopis, José Luis Vicedo, Antonio Ferrández

References: GPLSI, Departamento de Lenguajes y Sistemas Informáticos, Universidad de Alicante : http://gplsi.dlsi.ua.es/mwgplsi/index.php/Portada

Description: Los sistemas de recuperación de información se encargan de procesar una colección de textos y entre todos ellos seleccionar aquellos que contengan algún término relacionado con la pregunta y descartando los que no estén relacionados. El sistema IR-n es un sistema de recuperación de información basada en pasajes que utiliza un modelo probabilístico como motor de búsqueda y además utiliza un módulo de expansión de la pregunta que mejora los resultados obtenidos. Este sistema ha participado en concursos internacionales como el CLEF.

Functionality: Sistema de recuperación de información basado en pasajes.

Technology: Se basa en tecnología propia de los sistemas de RI, progamado en C++.

Technical Requirements: Plataforma Linux

Innovation: Lo que diferencia a este sistema respecto a otros sistemas de RI basados en pasajes es que IR-n utiliza la frase como unidad para la definición de pasajes lo que garantiza una estructura sintáctica completa en los pasajes utilizados.

Publications:

- Llopis F., A. Ferrández, J.L. Vicedo “Selección de pasajes para facilitar el proceso de búsqueda de respuestas”. Procesamiento Lenguaje Natural. Número 29, Pags. 273-280, Septiembre 2002

- Llopis, F., A. Ferrández, J.L. Vicedo, “Utilización de pasajes de tamaño variable para mejorar el proceso de recuperación de información”. Procesamiento Lenguaje Natural. Número 23, Pags. 47-53, Septiembre 1998.

- Llopis, F.: IR-n: Un Sistema de Recuperación de Información Basado en Pasajes. PhD thesis, University of Alicante (2003) .

- Vicedo, J.L, Llopis, F., Ferrández, A. "University of Alicante. Experiments at TREC-2002". Eleventh Text REtrieval Conference (TREC-11). Gaithersburg. Maryland (EEUU). November 2002. NIST Special Publication 500-250. Gaithersburg, US. Noviembre 2002.

Contact: Fernando Llopis <llopis@dlsi.ua.es >

JBrainDead

Authors: Juan M. Cigarrán, researcher, and Julio Gonzalo, head of the NLP Group at UNED.

References: Demo available: http://bender.lsi.uned.es:8080/ModuloWeb/jbraindead.html , NLP Group at UNED: http://nlp.uned.es

Description: An information retrieval system which performs the clustering of results by the automatic selection document descriptor, formal concept analysis and latent semantic indexing techniques.

Functionality: The IR system analyses the results and retrieves a concept lattice (showing the he most general information on the top and the most specific on the bottom) and allowing the user to browse across the relevant documents in a different fashion.

Technology: The system integrates an IR technology with information extraction and formal concept analysis techniques.

Technical Requirements: The system needs an IR module which retrieves a set of relevant documents from a query, expressing the user's information needs. The system is currently using the Yahoo! and Google API to retrieve documents from the Web, but this is an independent module.

Innovation: The added value of this technology, which is a result of a research project, is that a new form of exploring the results and browsing across relevant information, is proposed, allowing the non explicit relations discovery.

Publications:

- J. Cigarrán, A. Peñas, J. Gonzalo, F. Verdejo. 2005. "Automatic selection of noun phrases as document descriptors in an FCA-based Information Retrieval system". International Conference on Formal Concept Analysis (ICFCA 2005). Lecture Notes in Computer Science. Springer-Verlag, vol. 3403.

- J. Cigarrán, J. Gonzalo, A. Peñas, F. Verdejo (2004). "Browsing search results via Formal Concept Analysis: Automatic selection of Attributes". Concept Lattices Proceedings of the Second International Conference on Formal Concept Analysis (ICFCA 2004). Lecture Notes in Computer Science. Springer-Verlag.

Contact: Juan M. Cigarrán <juanci@lsi.uned.es>, Julio Gonzalo <julio@lsi.uned.es>

3.3 Sistemas de Búsqueda de Respuestas

Ihardetsi

Authors: IXA group

References: http://sisx04.si.ehu.es:8080/IhardetsiWebDemo/IhardetsiBezeroa.jsp

Description: Ihardetsi is a question answering system for Basque. It is a general platform which architecture pays special attention to: 1) the integration of the development and evaluation environments, and 2) the systematic use of XML declarative files to control the execution of the modules and the communication between them. It can work in a multilingual environment (see QA@CLEF 2008 evaluation forum) using MT systems.

Functionality: QA for Basque.

Technology: C++, XML.

Technical Requirements: -

Modules: The current version has three main modules, as it is very common in the question answering systems: Question analysis, Passage retrieval and Answer extraction.

Innovation: It is the first QA system for Basque.

Development: Different projects funded by the Basque government

Publications:

- Ansa O., Arregi X., Otegi A., Valverde A.. An XML Framework for a Basque Question Answering System. 7th International Conference on Flexible Query Answering Systems. Milano, Italia. (ISSN 0302-9743, ISBN 3-540-34638-4, pp 577-588). 2006.
- Ansa O., Arregi X., Otegi A., Soraluze A. Ihardetsi question answering system at QA@CLEF 2008. Working Notes of the Cross-Lingual Evaluation Forum, Aarhus, Denmark. ISBN 2-912335-43-4, ISSN 1818-8044. 2008.

Contact: Iñaki Alegria <i.alegria@ehu.es>

SQUASH

Authors: César de Pablo-Sánchez, Paloma Martínez-Fernández, José Luis Martínez-Fernandez, Maria Teresa Vicente-Díez (Universidad Carlos III de Madrid), Antonio Moreno Sandoval, Ana García-Ledesma (Universidad Autónoma de Madrid).

References: The Advanced Databases Group (Labda-UC3M) of Universidad Carlos III de Madrid, and the "Laboratorio de Lingüística Informática" of Universidad Autónoma de Madrid (LLI-UAM), are well-known research groups, with ample activity in several Natural Language Processing and Information Retrieval issues.

Description: SQUASH is a modular question answering system for the Spanish language. It enhances traditional search engine functionality by providing precise answers in real time to questions in natural

language like *"¿Cuándo se firmó el Tratado de Maastrich? (When was the Maastricht treaty signed?)"*. It reduces significantly the time a user must spend searching for precise information in textual databases.

Functionality: The system is composed of rules to select the type of information needed by a question and generate a suitable query for an information retrieval system. It also includes information extraction components to select and rank from documents appropriate sentences and answers.

Technology: The system integrates technology for question analysis, information extraction and information retrieval.

Technical Requirements: The system is implemented in Java and requires modules for Information Retrieval (IR) and Language Analysis. Several IR systems have been integrated (Lucene, Xapian and Google API). Currently Daedalus STILUS is used for Language Analysis.

Modules: Modules for preprocessing information includes language analysis and indexing libraries. Modules for online querying perform question classification, question analysis, query generation, sentence retrieval, answer extraction and answer ranking.

Innovation: SQUASH is the result of advances in natural language processing (technological push) and the need of fast semantic search engines to alleviate information overload (market pull). The system provides precise answer from Spanish in real time.

Development: SQUASH has been the result of the joint collaboration of a multidisciplinary team of researchers from the Universidad Carlos III de Madrid, Universidad Politécnica de Madrid and from the Universidad Autónoma de Madrid. It has been developed during more than 4 years, from the experience gained in several funded research projects. It has been independently evaluated in CLEF (Cross Lingual Evaluation Forum).

Publications:

- de Pablo-Sánchez, C., González-Ledesma, A., Martínez-Fernández, J., Guirao, J., Martínez, P. and Moreno, A. "MIRACLE's Cross-Lingual Question Answering Experiments with Spanish as a Target Language," *Accessing Multilingual Information Repositories* (), 2006, pp. 488--491.
- de Pablo-Sánchez, C., González-Ledesma, A., Moreno, A. and Vicente, M. T. "MIRACLE experiments in QA@CLEF 2006 in Spanish: main task, real-time QA and exploratory QA using Wikipedia (wiQA)," *Evaluation of Multilingual and Multi-modal Information Retrieval* (4730/2007), 2007, pp. 463-472.

Contact: Paloma Martínez Fernández <pmf@inf.uc3m.es>

3.4 Sistemas de Recuperación Automática

Detective Brooklynk: System for Automatic Recovery of Broken Web Links

Authors: Juan Martinez-Romo and Lourdes Araujo.

References: Demo available at http://bender.lsi.uned.es:8080/brooklynk/ NLP Group at UNED: http://nlp.uned.es

Description: Detective Brooklynk is an information retrieval system to provide a list of possible web pages to substitute the one pointed by a broken link.

Functionality: The system uses natural language techniques, such as named entity recognition, information extraction techniques and language models to extract information related to the considered broken link. This information is then used to make several queries, which are submitted to different search engines to retrieve documents related to the missing web page. In order to tune the results, the pages recovered in this way are ranked according to relevance measures obtained by applying information retrieval (IR) techniques, and finally this ordered list of results is presented to the user.

Technology: The system integrates an IR technology such as information extraction, ranking functions and natural language techniques.

Technical Requirements: The system needs an IR module which retrieves a set of relevant documents from a query constructed from the information extracted from the web pages. The system is currently using the Yahoo! BOSS API to retrieve documents from the Web, but this is an independent module.

Innovation: Most existing technologies to deal with the problem of broken links are based on the storage of information related to the site links in advance. Thus, Detective Brooklynk is a novel technology able to recommend, with high accuracy, candidate web pages to substitute a link without the need of information previously stored.

Development: Detective Brooklynk was developed as part of Juan Martinez-Romo's Ph.D. dissertation in 2010 and has been partially supported by the Spanish Ministry of Science and Innovation within the project QEAVis-Catiex (TIN2007-67581-C02-01) and the Regional Government of Madrid under the Research Networks MAVIR (S-0505/TIC-0267) and MA2VICMR (S2009/TIC-1542).

Publications:

- Juan Martinez-Romo and Lourdes Araujo: “Analyzing Information Retrieval Methods to Recover Broken Web Links”. ECIR 2010: LNCS 5993, pp. 26-37, Springer (2010).
- Juan Martinez-Romo and Lourdes Araujo "Retrieving Broken Web Links using an Approach based on Contextual Information". *ACM conference on Hypertext. Torino, Italy. June 29th - July 1th, 2009.*
- Juan Martinez-Romo and Lourdes Araujo: “Recommendation System for Automatic Recovery of Broken Web Links”. IBERAMIA 2008: LNCS 5290, pp. 302-311, Springer (2008).

Contact: Juan Martinez-Romo <juaner@lsi.uned.es>, Lourdes Araujo <lurdes@lsi.uned.es>

3.5 Sistemas de Traducción Automática

Matxin

Authors: IXA group

References: http://www.opentrad.org/index.php?idioma=eu, http://matxin.sourceforge.net

Description: Matxin is Transfer-based MT system from Spanish into Basque. It is an open, reusable and interoperable framework which can be improved in the next future combining it with the statistical model. The MT architecture reuses several open tools and it is based on an unique XML format for the flow between the different modules, which makes easer the interaction among different developers of tools and resources. Being Basque a resource-poor language this is a key feature in our aim for future improvements and extensions of the engine. The result is an open source software which can be downloaded from matxin.sourceforge.net, and we think it could be adapted to translating between other languages with few resources.

Functionality: MT form Spanish to Basque. English to Basque in progress

Technology: C++. It uses Freeling for analyzing Spanish or English and Lttoolbox-Apertium for access to the dictionaries. It reuses Basque morphology for morphological generation.

Technical Requirements: -

Modules: Lexical transfer, structural transfer, syntactical generation and morphological generation.

Innovation: It is the first MT system for Basque.

Development: OpenTrad and EurOpenTrad projects (Profit)

Publications:

- Alegria I., Díaz de Ilarraza A., Labaka G., Lersundi M., Mayor A., Sarasola K. Transfer-based MT from Spanish into Basque: reusability, standardization and open source. LNCS 4394. 374-384. Cicling 2007. ISBN-10: 3-540-70938-X ISSN 0302-9734. 2007.

Contact: Iñaki Alegria <i.alegria@ehu.es>

3.6 Sistemas de Resumen Automático
AutoPan

Authors: Maria Fuentes, Edgar Gonzàlez and Horacio Rodríguez.

References: http://www.lsi.upc.edu/~egonzalez/autopan.html

Description: AutoPan is a tool that helps in the evaluation of Automatic Summaries. In DUC 2001 to 2004, the manual evaluation was based on comparison with a single human-written model and a lot of the information of evaluated summaries (both human and automatic) was marked as "related to the topic, but not directly expressed in the model summary". The pyramid method (proposed by [Nenkova and Passoneau, 04]) addresses the problem by using multiple human summaries to create a gold-standard and by exploiting the frequency of information in the human summaries in order to assign importance to different facts. However, the method of pyramids for evaluation requires a human annotator to match fragments of text in the system summaries to the SCUs in the pyramids. We have tried to automate this part of the process.

The text in the SCU label and all its contributors is stemmed and stop words are removed, obtaining a set of stem vectors for each SCU. The system summary text is also stemmed and freed from stop words.

A search for non-overlaping windows of text which can match SCUs is carried. A window and an SCU can match if a fraction higher than a threshold (experimentally set to 0.90) of the stems in the label or some of the contributors of the SCU are present in the window, without regarding order. Each match is scored taking into account the score of the SCU as well as the number of matching stems. The solution which globally maximizes the sum of scores of all matches is found using dynamic programming techniques.

The constituent annotations automatically produced are scored using the same metrics as for manual annotations, and it is found that there is statistical evidence supporting the hypothesis that the scores obtained by automatic annotations are correlated to the ones obtained by manual ones for the same system and summary.

Functionality: Takes a pyramid file and a summary and produces the peer annotation file that afterwards can be evaluated using the software provided by DUC.

Technology: Perl 5.6.0 or greater

Technical Requirements: XML::Parser Perl Module; Expat Library

Modules: -

Innovation:

Development: AutoPan was developed for Maria Fuentes' Phd thesis within the framework of the CHIL projects.

Publications:

- Maria Fuentes, Edgar Gonzàlez, Daniel Ferrés, Horacio Rodríguez. QASUM-TALP at DUC 2005 Automatically Evaluated with the Pyramid based Metric AutoPan DUC 2005 Evaluation Campaign, 2005
- Maria Fuentes, Enrique Alfonseca, Horacio Rodríguez "Support Vector Machines for Query-focused Summarization trained and evaluated on Pyramid data". In Proceedings of the ACL 2007, Prague, Czech Republic, June 2007

Contact: Maria Fuentes <mfuentes@lsi.upc.edu>, Edgar Gonzàlez <egonzalez@lsi.upc.edu>

GPLSI COMPENDIUM

Authors: Elena Lloret y Manuel Palomar.

References: GPLSI, Departamento de Lenguajes y Sistemas Informáticos, Universidad de Alicante: http://gplsi.dlsi.ua.es/gplsi09/doku.php

Description: GPLSI COMPENDIUM es una herramienta de generación de resúmenes modular para el idioma inglés. Permite producir resúmenes de texto de forma automática, extrayendo las frases más relevantes de uno o varios documentos (resúmenes mono-documento y multi-documento), generando así un resumen informativo, cuyo tamaño podrá consistir en un número fijo de palabras o bien un porcentaje respecto al documento origen.

Functionality: Para la generación del resumen final, GPLSI COMPENDIUM acepta como entrada un documento o un conjunto de documentos relacionados y el tamaño que deseemos para el resumen final (en número de palabras o en porcentaje). Como salida devolverá el resumen correspondiente en el directorio “summaries” que se creará automáticamente para tal fin.

Technology: La herramienta se ha desarrollado en Java y Bash y se ejecuta desde la línea de comandos en Linux.

Technical Requirements: sistema operativo Linux que tenga instalado Freeling (http://nlp.lsi.upc.edu/freeling/)

Modules: GPLSI COMPENDIUM se basa en cinco etapas: i) análisis lingüístico; ii) detección de redundancia; iii) identificación del tópico; iv) detección de relevancia; y v) generación del resumen. En la etapa de detección de redundancia se utiliza la técnica de implicación textual; en la de identificación del tópico, utilizamos la frecuencia de las palabras, y finalmente en la etapa de detección de redundancia, nos

basamos en el principio de la cantidad de información. El sistema permite generar resúmenes, utilizando alguna o todas de las etapas anteriormente citadas. Mediante módulos adicionales específicos se puede generar otros tipos de resúmenes, como por ejemplo resúmenes orientados a un tema en concreto, resúmenes subjetivos u orientados a abstractos.

Innovation: La diferencia de esta herramienta con respecto a otras herramientas de generación automática de resúmenes que existen en la actualidad radica en que integra un novedoso método basado en el reconocimiento de la implicacioń textual (Ferrández, 2009[3]) para identificar y detectar información redundante. Además, la detección de relevancia de las oraciones de un documento se basa en un principio de origen lingüístico-cognitivo (Givón, 1990[4]).

Development: La herramienta forma parte de los trabajos realizados dentro de la Tesis doctoral de Elena Lloret y de los trabajos desarrollados en los proyectos TEXTMESS (TIN200615265-C0601), TEXT-MESS 2.0 (TIN2009-13391-C04-01) y PROMETEO (PROMETEO/2009/199).

Publications:

- Elena Lloret and Manuel Palomar: Challenging Issues of Automatic Summarization: Relevance Detection and Quality-based Evaluation. *International Journal of Informatica, 34 (2). ISSN 0350-5596,* 2010.
- Elena Lloret and Manuel Palomar:. A Gradual Combination of Features for Building Automatic Summarisation Systems. Proceedings of the 12th International Conference on Text, Speech and Dialogue, pp 16–23, 2009.

Contact: Elena Lloret (elloret@dlsi.ua.es)

LCsum

Authors: Maria Fuentes, Horacio Rodríguez and Edgar Gonzàlez.

References: http://nidhoggr.lsi.upc.edu/~demo/summary.html

Description: A summarizer for different tasks involving aspects related with the language, the media and the domain of the document to be summarized.

Functionality: Currently on-line summarization of generic or scientific textual or spontaneous speech documents in English.

Technology: The summarizer is developed in Perl, and a specific wrapper is developed when some component using different technology is included (Freeling, TNT, YAMCHA, ...).

Technical Requirements: The summarizer requires a linguistic preprocess. For Catalan and Spanish Freeling and euroWordnet are used, while TNT, WordNet and YAMCHA are used to process English spontaneous speech documents.

[3] Ferrández Escámez, Óscar. 2009. Textual Entailment Recognition and its Applicability in NLP Task. Ph.D. thesis, Universidad de Alicante.

[4] Givón, Talmy. 1990. Syntax: A functional-typological introduction, II. John Benjamins.

Modules: Linguistic preprocessing; Lexical chain based summarizer (includes: Discourse Marker annotator, Text Tiler, Lexical Chainer and chunkSum).

Innovation: -

Development: The FEMsum was developed for Maria Fuentes' Phd thesis within the framework of the HERMES, ALIADO and CHIL projects.

Publications:

- Maria Fuentes. "A Flexible Multitask Summarizer for Documents from Different Media, Domain, and Language". Ph.D. Thesis on Artificial Intelligence, Advisor Horacio Rodríguez. March, 2008.
- Maria Fuentes, Edgar Gonzàlez, Horacio Rodríguez, Jordi Turmo, Laura Alonso. "Summarizing Spontaneous Speech Using General Text Properties". In Proceedings of the Crossing Barriers in Text Summarization Research Workshop held in conjunction with RANLP, Borovets, Bulgary, September 2005.
- Maria Fuentes, Edgar Gonzàlez, and Horacio Rodríguez. "Resumidor de noticies en català del projecte Hermes". II Congrés d'Enginyeria en Llengua Catalana, Andorra, 2004.
- Laura Alonso, Maria Fuentes. "Integrating Cohesion and Coherence for Text Summarization". In Proceedings of the EACL'03 Student Session, Budapest, Hungary, April 2003.
- Maria Fuentes, Horacio Rodríguez. "Using cohesive properties of text for Automatic Summarization". In Proceedings of the Primeras Jornadas de Tratamiento y Recuperación de Información, Valencia, Spain, 2002.

Contact: Maria Fuentes <mfuentes@lsi.upc.edu>, Horacio Rodríguez <horacio@lsi.upc.edu>, Edgar Gonzàlez <egonzalez@lsi.upc.edu>

3.7 Recursos de morfología y léxico

CANeo TIP

Authors: Carreras-Riudavets, F.; Jiménez-Estupiñán, R.; Hernández-Figueroa, Z.; Rodríguez-Rodríguez, G. (ULPGC)

References: http://tip.dis.ulpgc.es/es/catalogador-de-neologismos

Description: Catalogador automático de neologismos sufijales y prefijales.

Functionality: El sistema CANeo TIP es capaz de detectar los prefijos y los sufijos de un posible neologismo y aplicar las diferentes reglas del español para conseguir un conjunto de palabras primitivas; es decir palabras de las que puede provenir. Para catalogar el neologismo es necesario primero catalogar las palabras primitivas. Para ello utilizamos el servicio de lematización. Realizando consultas a este servicio, obtendremos información acerca de la categoría gramatical de una determinada palabra primitiva, además de información muy valiosa para realizar la catalogación y valoración de los resultados. Utilizando la información estadística recopilada en el estudio, y partiendo de la categoría gramatical de la primitiva, se realiza una estimación a cerca de la categoría gramatical del neologismo.

Technology: Página web ASP.NET programada en C#

Technical Requirements: http://tip.dis.ulpgc.es/neologismo

Modules: -

Innovation: Esta aplicación se basa en el estudio de unas setenta mil palabras derivadas de palabras primitivas que reúne, entre otras cosas, información muy valiosa referente a la utilización de los afijos más productivos del idioma español, sus significados, información estadística de frecuencias de utilización, etc. De manera general, podemos definir la metodología de trabajo para localizar posibles reglas a aplicar, para obtener la palabra primitiva de la que proviene una palabra derivada, de la siguiente manera:

- Análisis sufijal: Se revisa el conjunto de reglas sufijales. Se buscarán reglas que puedan encajar con la terminación sufijal y se incluirán anotaciones acerca de las estadísticas de uso, significados, reglas de corte, etc.

- Análisis prefijal: Se revisa el conjunto de reglas prefijales. Se buscarán reglas que puedan encajar con la terminación prefijal y se incluirán anotaciones acerca de las estadísticas de uso, significados, reglas de corte, etc.

- Se revisarán sustituciones de raíces irregulares: Se anotarán las transformaciones de raíces irregulares en pares que incluyan la raíz de origen y la raíz transformada.

- Se revisan reglas ortográficas: Reglas tales como diptongos, hiatos y otras reglas propias del español.

- Análisis parasintético: Una parasíntesis es la formación de palabras por medio de una combinación de afijos, normalmente pares de prefijo-sufijo. Algunas de estas parasíntesis describen una relación habitual y deben ser tratadas de manera diferente. Del conjunto de sufijos y prefijos estudiados, se reunirán y anotarán estadísticas de uso de las relaciones que existan entre ellos.

- Tratamiento de tildes: Existe un conjunto de reglas de acentuación que también son consideradas en este trabajo.

Development: -

Publications: -

Contact: Francisco Javier Carreras Riudavets <tip@dis.ulpgc.es>

Conjugador TIP

Authors: Rodríguez-Rodríguez, G.; Carreras-Riudavets, F; Hernández-Figueroa, Z (ULPGC)

References: http://tip.dis.ulpgc.es/es/conjugacion-de-verbos

Description: Conjugador de verbos español

Functionality: El Conjugador TIP de verbos del español presenta por separado la conjugación de formas en negativo, la conjugación pronominal y la conjugación con el sujeto en femenino, con el fin de hacer más sencilla la información que muestra. Recoge las diferentes conjugaciones, aceptadas por la Asociación de Academias de la Lengua Española, correspondientes a distintas zonas geográficas: el uso de vos como segunda persona del singular (principalmente usada en Río de la Plata) y el uso de ustedes como segunda persona del plural (principalmente usada en las Islas Canarias), y el uso de formas de respeto usted y ustedes.

Se muestran notas con la información morfológica y ortográfica de todos los verbos irregulares y defectivos. A partir de estas notas se puede obtener el listado de verbos que las cumplen. Igualmente, se muestran otros verbos que siguen el mismo modelo de conjugación, pudiéndose obtener todos los verbos que se conjugan de esa forma. El Conjugadr TIP incluye: de cada forma verbal se muestra su frecuencia de aparición en el Corpus de Referencia del Español Actual (CREA) y las formas con pronombres enclíticos que aparecen en dicho Corpus.

Technology: Página web ASP.NET programada en C#

Technical Requirements: http://tip.dis.ulpgc.es/conjugar-verbo/

Modules: -

Innovation:

- Actualizado con la Nueva Gramática de la Lengua Española (2009)
- Conjugación dialectal
- Conjugación pronominal
- Conjugación negada
- Conjugación con sujetos femeninos
- Conjugación dependiendo del significado
- Formas con pronombres enclíticos que aparecen en el Corpus de Referencia del Española Actual (CREA)
- Notas morfológicas detalladas
- Notas ortográficas detalladas
- Frecuencia de aparición de las formas en el CREA
- Listados de verbos según modelo de conjugación
- Listados de verbos según cambios morfológicos u ortográficos

Development: -

Publications: -

Contact: Francisco Javier Carreras Riudavets <tip@dis.ulpgc.es>

LIBNAFDA (Library for the Efficient Handling of Large Dictionaries)

Authors: María Nieves Fernández Formoso, Fco. Mario Barcala Rodríguez y Jorge Graña Gil

References: http://libnadfa.sourceforge.net/index.html

Description: LIBNAFDA is a C library which allows to manage large dictonaries of many kinds efficiently and minimizing memory consumption. For this purpose we use numbered acyclic deterministic finite-state automata, and we understand as dictionary, in this context, any structure that can link dictionary entries (words) which their information.

Functionality: LIBNAFDA is a C library where you can manage, efficiently and minimizing memory consumption, large dictionaries of many kinds. For this purpose it uses numbered acyclic deterministic finite-state automata. In this work we assume that dictionary means any structure that would associate its entries (words) to any kind of information. For the development of this library we have followed two maxims:

1. To minimize memory consumption needed to store the dictionaries.

2. To minimize the running time for them to be used in systems that require a high performance.

The library consists of two parts: one part is used to build the compiler, which deals with the task of compiling or compressing the words dictionaries, and the other part is responsible for facilitating access to these compiled dictionaries.

The compiler needs a list of words and the information associated with them to generate the compressed dictionaries. From this information, it generates a compiled dictionary (compressed) in binary format, which can be accessed by any independent program throw the second part of the library.

The key features that differentiate this library from other existing proposals are:

1. You can store any type of information associated with words inside it. Actually, dictionaries store integers and/or floats, but since these integers can be interpreted as indexes to any structure external to the library, they may reference data of any type, including strings.

2. It enables the simultaneous management of multiple dictionaries at a time and with references between them. Because the integer data can be interpreted in different ways, a particular case is to consider them as indexes to other dictionaries, which allows several dictionaries to be connected.

Technology: LIBNAFDA is imlmeented in C.

For words storage the library uses a numbered acyclic deterministic finite-state automaton, which is built by the compiler using the automata building algorithm proposed by Jan Daciuk in his article: Incremental Construction of Minimal Acyclic Finite-State Automata. Therefore, the automaton is built in an incremental and minimal way, and the use of memory and word recognition speed are optimized.

Technical Requirements: To compile the library sources the following packages must be installed:

- libxml2 2.6 or higher
- libxml2-dev 2.6 or higher
- libglib 2.16 or higher

Modules: -

Innovation: We have combined the building principles of minimal automata proposed by Jan Daciuk and others in his paper Incremental Construction of Minimal Acyclic Finite-State Automata, with the concepts raised by Jorge Graña Gil and others to manage the information associated with words in Compilation

Methods of Minimal Acyclic Finite-State Automata for Large Dictionaries. The result is an useful library for environments which need a very efficient access to information associated with words.

Development: -

Publications:

- Jorge Graña, Fco. Mario Barcala, and Miguel A. Alonso, Compilation Methods of Minimal Acyclic Finite-State Automata for Large Dictionaries, in Bruce W. Watson and Derick Wood (eds.), Implementation and Application of Automata, volume 2494 of Lecture Notes in Computer Science, pp. 135-148, Springer-Verlag, Berlin-Heidelberg-New York, 2002. ISSN 0302-9743 / ISBN 3-540-00400-9.

- Alejandro Sobrino, Santiago Fernández, and Jorge Graña, Access to a large dictionary of Spanish synonyms: a tool for fuzzy information retrieval, in Enrique Herrera-Viedma, Gabriella Pasi and Fabio Crestani (eds.), Soft computing in web information retrieval: models and applications, volume 197 of Studies in Fuzziness and Soft Computing, pp. 299-316, Springer-Verlag, Berlin-Heidelberg-New York, 2006. ISSN 1434-9922 / ISBN 3-540-31588-8.

- Santiago Fernández, Jorge Graña, and Alejandro Sobrino, Introducing FDSA (Fuzzy Dictionary of Synonyms and Antonyms): Applications on Information Retrieval and Stand-Alone Use, Mathware & Soft Computing, 10(2-3):57-70, 2003. ISSN 1134-5632.

Contact: jorge.grana@udc.es

ML-SentiCon: A Layered, Multilingual Sentiment Lexicon

Authors: Fermín L. Cruz, José A. Troyano, Beatriz Pontes, F. Javier Ortega

References: http://www.lsi.us.es/~fermin/index.php?title=Datasets

Description: Se trata de varias listas de lemas positivos y negativos para inglés, español, catalán, gallego y vasco. Cada lema viene acompañado de una estimación numérica de su polaridad (entre -1.0 y 1.0) así como de un valor de desviación típica de dicha polaridad. Las listas están organizadas en varias capas, de manera que las primeras capas contienen estimaciones más precisas de los valores anteriores, aunque contienen menos elementos que las capas posteriores.

Además de las listas de lemas, el recurso también contiene un lexicón a nivel de synsets para el inglés, con el mismo formato de SentiWordNet (SWN). Este lexicón fue obtenido a partir de una versión mejorada del método original de SWN, y utilizado para la generación de las listas de lemas anteriores.

Functionality: -

Technology: -

Technical Requirements: -

Modules: -

Innovation: -

Development: -

Publications:

- Cruz, Fermín L., José A. Troyano, Beatriz Pontes, F. Javier Ortega. Building layered, multilingual sentiment lexicons at synset and lemma levels, Expert Systems with Applications, 2014.

Contact: Fermín L. Cruz: fcruz@us.es

Números TIP

Authors: Carreras-Riudavets, F; Rodríguez-Rodríguez, G; Hernández-Figueroa, Z; Arroyo-Herrero, L.; (ULPGC)

References: http://tip.dis.ulpgc.es/es/convertir-numeros-a-texto-letras

Description: Conversor de números a su texto correspondiente.

Functionality: Números TIP realiza la conversión de una cifra a su texto cardinal, ordinal, fraccionario o partitivo, multiplicativo, romano…, y ofrece información morfológica, ortográfica y gramatical de cada uno de los términos. Además, se incluyen ejemplos que ayudan a la comprensión y buen uso. Los contextos implementados son:

- Texto de un número cardinal: Los números cardinales expresan cantidad en relación con la serie de los números naturales. El rango de números cardinales admitido por Números TIP es: ±999 999 x 10120. En texto sería desde el menos novecientos noventa y nueve mil novecientos noventa y nueve vigintillones al novecientos noventa y nueve mil novecientos noventa y nueve vigintillones.

- Texto de un número ordinal: Los números ordinales expresan orden o sucesión e indican el lugar que ocupa el elemento en una serie ordenada. El rango de números ordinales admitido por Números TIP es: 1 … 999 999 x 10120. En texto sería desde el primero hasta el novecientos noventa y nueve mil novecientos noventa y nuevevigintillonésimo.

- Texto de un número fraccionario: Los números partitivos expresan división de un todo en partes y designan las fracciones iguales en que se ha dividido la unidad. El rango de números fraccionarios admitido por Números TIP es: 2 … 999 999 x 10120. En texto sería desde la mitad hasta el novecientosnoventainuevemilnovecientosnoventainuevevigintillonésimo.

- Texto de una fracción: Las fracciones expresan una cantidad dividida entre otra cantidad. El rango de fracciones admitido por Números TIP es: ±999 999 x 10120 / ±999 999 x 10120 .

- Texto de un número mulltiplicativo: Los multiplicativos expresan que el sustantivo al que se refieren se compone de tantas unidades o implica tantas repeticiones como el numeral indica. El rango de números multiplicativos admitido por Números TIP es: 2 … 999 999 x 10120, aunque a partir de trece se escribe con una expresión. En texto sería desde el doble hasta novecientos noventa y nueve mil novecientos noventa y nueve vigintillones de veces.

- Número romano: Los números romanos expresan los valores numéricos de nuestro sistema de cifras con un repertorio de signos distintos. El rango de números romanos admitido por Números TIP es: 1 … 3 999 999 999 999.

En texto sería desde el I hasta el MMMCMXCIXCMXCIXCMXCIXCMXCIX.

En caso de introducir un número romano incorrecto, Números TIP corregirá el error ofreciendo la numeración adecuada.

- Texto de un número de colectivo: Los números de grupos o colectivos expresan el número de elementos que lo componen. El rango de números de conjunto admitido por Números TIP es: 2 … 20 y las decenas hasta 100. En texto sería desde el par o la pareja hasta la centena.

- Texto de un número de sílabas: Número de sílabas de que se compone una palabra o métrica usada en los versos para clasificarlos. El rango de números de sílabas admitido por Números TIP es: 1 … 19. En texto sería desde el monosílabo hasta el eneadecasílabo.

- Nombre de polígonos: Un polígono es una figura plana limitada por varias líneas rectas. El rango del número de lados para definir un polígono es: 3 … 999 999 x 10120, aunque a partir de 15, excepto los múltiplos de diez y cien, se escribe con una expresión. En texto sería desde el triángulo hasta el polígono de novecientos noventa y nueve mil novecientos noventa y nueve vigintillones de lados.

- Nombre de poliedros: Un poliedro es un cuerpo limitado por superficies planas. El rango del número de caras para definir un poliedro es: 4 … 999 999 x 10120, aunque a partir de 15 se escribe con una expresión. En texto sería desde el tetrágono hasta el poliedro de novecientos noventa y nueve mil novecientos noventa y nueve vigintillones de caras.

- Edades: Tiempo vivido o duración aproximada de la existencia de algo o alguien. El rango para las edades admitido por Números TIP es: 2 … 100 y luego el 1000. En texto sería desde el dosañero o dosañal hasta el milenario.

- Nacido: Nacido expresa el número de seres nacidos en un mismo parto. El rango para los nacidos admitido por Números TIP es: 2 … 10. En texto sería desde el mellizo o gemelo hasta el decallizo.

Números TIP se ha desarrollado a partir de las siguientes fuentes normativas y descriptivas:

- OLE-RAE-2010. Ortografía de la lengua española, edición revisada por las Academias de la Lengua Española y publicada por la Real Academia Española (2010).

- NGLE-RAE-2009. Nueva gramática de la lengua española. Espasa Calpe (2009)

- DPD-RAE-2005. Diccionario panhispánico de dudas de la Real Academia Española – 1ª edición (2005)

- DOLE-1996. Diccionario de ortografía de la lengua española. Thomson Paraninfo (1996).

Technology: Página web ASP.NET programada en C#

Technical Requirements: http://tip.dis.ulpgc.es/numeros-texto/

Modules: -

Innovation: -

Development: -

Publications:

Contact: Francisco Javier Carreras Riudavets <tip@dis.ulpgc.es>

ParamText TIP

Authors: Carreras-Riudavets, F.; Santana-Herrera, J.C.; Hernández-Figueroa, Z.; Rodríguez-Rodríguez, G. (ULPGC)

References: http://tip.dis.ulpgc.es/paramtext

Description: -

Functionality: El ParamText TIP analiza un documento y extrae información estadística de interés. Los datos analizados se muestran en gráficas y en tablas, exportables a Microsoft excel, para su estudio y análisis posterior por el usuario.

El ParamText TIP analiza el contenido léxico de un texto, extrayendo el número de párrafos, oraciones, palabras y caracteres. Asímismo, se extrae para cada uno de estos grupos el número de oraciones, de palabras y de caracteres de cada párrafo, el número de palabras y de caracteres de cada oración y el número de caracteres de cada palabra. Ofrece información métrica como la frecuencia de aparición de las palabras en el texto, el centro de gravedad de los vocablos, la distribución de las palabras según su primera aparición y su frecuencia de uso en el español. Asimismo, se muestra en una tabla el vocabulario completo utlizado en el texto.

El ParamText TIP analiza morfológicamente el texto y extrae información relacionada con las categorías gramaticales de las palabras y su flexión morfológica. ParamText no realiza un análisis sintáctico de las frases y, por tanto, no desambigüa las múltiples opciones morfológicas que en ocasiones puede tener una palabra, sino que ofrece el reconocimiento morfológico de cada palabra independientemente de su función en la oración. El grupo Text & Information Processing está trabajando para extraer además la función gramatical que le corresponde a cada palabra en la oración.

Por otro lado, el ParamText TIP permite distinguir en sus análisis entre palabras con significado o sentido semántico y palabras vacías. El ParamText TIP aporta un conjunto de palabras vacías por defecto que el usuario puede modificar en cualquier momento según sus intereses.

Technology: Página web ASP.NET programada en C#

Technical Requirements: http://tip.dis.ulpgc.es/paramtext/

Modules: -

Innovation:

- Extrae información estadística de documentos
- Admite documentos de tipo word, pdf y txt
- Se muestran gráficas y tablas con los resultados
- Analiza morfológicamente las palabras del documento
- Todos los datos son exportables a excel
- Permite definir el conjunto de palabras vacías

Development: -

Publications:

Contact: Francisco Javier Carreras Riudavets <tip@dis.ulpgc.es>

Silabeador TIP

Authors: Hernández-Figueroa, Z; Rodríguez-Rodríguez, G; Carreras-Riudavets, F (ULPGC)

References: http://tip.dis.ulpgc.es/es/silabeador

Description: Silabeador de palabras español.

Functionality: Silabeador-TIP es una aplicación en línea que realiza la separación en sílabas de cualquier palabra española. Silabeador-TIP usa un separador de sílabas basado en criterios ortográficos que complementa con una herramientas de análisis morfológico, que soporta más de cuatro millones de palabras, y con una base de datos, que tiene más de 80 mil relaciones semántico-léxicas, para realizar separaciones de forma inteligente por componentes, averiguando, primero, si la palabra existe, para posteriormente identificar la posible existencia de afijos que condicionen la silabación. Además, resalta la sílaba tónica de las palabras y marca la existencia de diptongos, triptongos e hiatos.

Technology: Página web ASP.NET programada en C#

El código fuente en C++ del silabeador básico se puede descargar gratuitamente bajo licencia GNU General Public License

Technical Requirements: http://tip.dis.ulpgc.es/Silabas_Web/

Para el código descargable se necesita un compilador de C++

Modules: -

Innovation: Identificación de afijos que condicionan la silabación basándose en conocimiento morfológico y en relaciones semántico-léxicas

Doble acentuación de los adverbios terminados en -mente

Development: -

Publications:

Contact: Zenón Hernández Figueroa <tip@dis.ulpgc.es>

4. Librerías y software de propósito general (en ingeniería lingüística)

ARIES: A Lexical Base and Platform

Authors: José M. Goñi-Menoyo, José C. González-Cristóbal (Universidad Politécnica de Madrid), Antonio Moreno Sandoval (Univers idad Autónoma de Madrid).

References: The "Grupo de Sistemas Inteligentes" (GSI-UPM) of Universidad Politécnica de Madrid, and the "Laboratorio de Lingüística Informática" of Universidad Autónoma de Madrid (LLI-UAM) are well-known research groups with extensive activity in several Natural Language Processing projects.

Description: ARIES is a large lexical database for Spanish language that includes a formalism for lexical representation and a morphological model for inflectional morphology. This model is based on allomorphs, so rules for automatic allomorph generation from lexical roots are also provided.

Functionality: The database consists of the lexical database, declarative rules for inflectional morphology and for allomorph expansion, and related documentation .

Technology: TRIELIB is a declarative lexical database with no implementation associated. However, a version of the lexical database is translated to Prolog DCG clauses.

Technical Requirements: The data needs additional implementation before being processed.

Modules: Basic indexing management library and lexical information management library.

Innovation: ARIES was developed due to the dramatic lack of lexical resources for Spanish language in 1995.

Development: ARIES has been the result of the joint collaboration of a multidisciplinary team of researchers from the Universidad Politécnica de Madrid and from the Universidad Autónoma de Madrid. It has been developed during more than 5 years, from the experience gained in several funded research project.

Publications:

- Goñi-Menoyo, J.M.; González-Cristóbal, J.C.; and Moreno-Sandoval, A. (1995). Manual de Referencia de la Plataforma Léxica ARIES, versión 5.0. Universidad Politécnica de Madrid.
- Right properties / owner: Universidad Politécnica de Madrid and DAEDALUS-Data, Decisions and Language, S.A. The exploitation rights are currently transferred to DAEDALUS-Data, Decisions and Language, S.A.

Contact: José Miguel Goñi-Menoyo <josemiguel.goni@upm.es>

JDBIR library

Authors: José Manuel Gómez

References: GPLSI, Departamento de Lenguajes y Sistemas Informáticos, Universidad de Alicante: http://gplsi.dlsi.ua.es/mwgplsi/index.php/Portada. Disponbile en: http://sourceforge.net/projects/jirs/

Description: Simple gestor de base de datos para el desarrollo de sistemas de recuperación de información.

Functionality: Una librería JAVA para crear bases de datos especialmente adaptadas para tratar con ficheros invertidos.

Technology: Desarrollado integramente en Java.

Technical Requirements: Java

Innovation: Permite utilizar ficheros invertidos como si se trataran de campos convencionales de

una BD.

Development: Realizado durante la tesis de José M. Gómez en el marco del proyecto R2D2 (TIC200307158C04).

Publications:

- Gómez, J.M. "Recuperación de pasajes multilingüe para la búsqueda de respuestas", Tesis Doctoral. Director: Dr. Emilio Sanchis Arnal. Noviembre, 2007.

Contact: José Manuel Gómez <jmgomez@ua.es >

InTime Platform

Authors: José Manuel Gómez, Sergio Navarro, Patricio Martínez-Barco.

References: GPLSI, Departamento de Lenguajes y Sistemas Informáticos, Universidad de Alicante: http://intime.dlsi.ua.es

Description: Plataforma de integración basada en una arquitectura P2P, que permite compartir y acceder a distintos recursos de Procesamiento del Lenguaje Natural (PLN).

Functionality: La plataforma InTime permite conocer, acceder utilizar y compartir datos y herramientas de PLN. Consta de una arquitectura distribuida similar a las redes P2P basada en servicios Web. Los usuarios de InTime tienen la posibilidad de publicar sus propios recursos de PLN y compartirlos con el resto del mundo usando tecnologías de la Web 2.0.

Technology: Básicamente, el sistema está desarrollado en Java sobre Tomcat 5.5 , usando servicios Web, pero el Panel de Control Web usa Apache 2 y PHP5.

Technical Requirements: JRE 5 , Tomcat 5.5, Apache 2 y MySQL 5

Modules: Cliente, Servidor y Panel de Control Web

Innovation: El aspecto más importante de InTime es la capacidad de que todo sea transparente al usuario, y que éste no tenga que cambiar su metodología de trabajo.

Development: Desarrollado en el marco del proyecto Text-Mess (TIN2006-15265-C06-01).

Publications:

- Gómez, J.M. “InTiMe: Plataforma de Integración de Recursos de PLN”. Procesamiento del Lenguaje Natural (40), 2008, pp. 83-90. Spain.

Contact: José Manuel Gómez <jmgomez@ua.es >

IQmt

Authors: Jesús Giménez and Lluís Màrquez

References: http://www.lsi.upc.edu/~nlp/IQMT

Description: A a common workbench on which automatic MT evaluation metrics can be robustly used and combined for the purpose of MT system development. Current version includes a rich set of metrics operating at different linguistic levels (lexical, shallow syntactic, syntactic, and shallow semantic).

Functionality: -

Technology: Perl

Technical Requirements: The followinf requirements are optional: Individual metrics (BLEU, NIST, GTM, ROUGE, METEOR, ...); Linguistic processors (C&C Tools); SwiRL; BIOS; Charniak Parser; MINIPAR dependency parser; SVMTool.

Modules: Metrics: Lexical Similarity; Syntactic Similarity (Shallow parsing, Dependency parsing, Constituency parsing); Semantic Similarity (Named entities, Semantic roles, Discourse representations); QARLA Framework.

Innovation: IQmt allows for evaluation of translation quality aspects at different linguistic levels. Metric scores can also be combined into a single measure of quality, either through QARLA, of via arithmetic mean.

Development: -

Publications:

- Jesús Giménez and Enrique Amigó. IQMT: A Framework for Automatic Machine Translation Evaluation. Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'06). Genoa, Italy, 22-28 May. 2006. Departament Research Technical Report (LSI-07-29-R).

- Jesús Giménez and Lluís Màrquez. Linguistic Features for Automatic Evaluation of Heterogeneous MT Systems. Proceedings of WMT 2007 (ACL'07).

- Jesús Giménez and Lluís Màrquez. A Smorgasbord of Features for Automatic MT Evaluation. Proceedings of the 3rd ACL Workshop on Statistical Machine Translation (shared evaluation task).

Contact: Jesús Giménez <jgimenez@lsi.upc.edu>

JPM Framework

Authors: José Manuel Gómez

References: GPLSI, Departamento de Lenguajes y Sistemas Informáticos, Universidad de Alicante: http://gplsi.dlsi.ua.es/mwgplsi/index.php/Portada. Disponbile en: http://sourceforge.net/projects/jirs/

Description: Un framework para el desarrollo de aplicaciones de PLN.

Functionality: Permite generar aplicaciones a partir de módulos (procesos y métodos) definidos en un archivo de configuración.

Technology: Desarrollado íntegramente en Java.

Technical Requirements: Java

Innovation: Permite modificar tanto los parámetros como la funcionalidad y el modo de ejecución de las aplicaciones modificando únicamente el archivo de configuración.

Development: Proyecto iniciado con la tesis de José M. Gómez en el marco del proyecto R2D2 (TIC2003-07158-C04) y se ha usado en otros proyectos como el Text-Mess (TIN2006-15265-C06-01). Es el núcleo de la plataforma de integración de InTime así como de varios buscadores federados.

Publications:

- Gómez, J.M. "Recuperación de pasajes multilingüe para la búsqueda de respuestas", Tesis Doctoral. Director: Dr. Emilio Sanchis Arnal. Noviembre, 2007.

Contact: José Manuel Gómez <jmgomez@ua.es>

LabelTranslator

Authors: Mauricio Espinoza Mejía, Asunción Gómez-Pérez, Elena Montiel-Ponsoda, Guadalupe Aguado de Cea y Eduardo Mena.

References: The Ontology Engineering Group (http://www.oeg-upm.net/) at the Artificial Intelligence Laboratory in the Computer Science School at Universidad Politécnica de Madrid (UPM) carries out research on the Ontological Engineering field and the Semantic Web.

Description: Ontology localization consists in adapting an ontology to a specific natural language and culture community. LabelTranslator takes as input an ontology or set of ontologies whose labels are specified in a given natural language and obtains the most probable translation of each ontology component label in a target natural language.

Functionality: The main functionalities of our tool are: i) obtains a ranked set of translations for each ontology component label, ii) stores the additional linguistic information associated to each label, and iii) uses a synchronization mechanism to maintain ontologies and their associated linguistic information always up to date.

Technology: LabelTranslator is implemented in Java suitable for being installed on Linux or Windows platforms.

Technical Requirements: LabelTranslator needs NeOn ToolKit version 1.2 and onwards

Modules: i) *translation service* automatically obtains a set of applicable translations of an ontology component label, ii) *translation ranking service* sorts these translations according to the similarity with its

lexical and semantic context, by means of external lexical resources, and iii) linguistic model (LIR) stores the linguistic information associated to the translated label.

Innovation: The added value of this technology is the automatization of the localization process which reduces human efforts to localize manually the ontology.

Development: This work has been funded by the NeOn project (http://www.neon-project.org/).

Publications:

- Mauricio Espinoza, Asunción Gómez-Pérez and Eduardo Mena, "Enriching an Ontology with Multilingual Information", Proc. of 5th European Semantic Web Conference (ESWC'08), Tenerife (Spain), Springer Verlag LNCS, pp. 333-347, June 2008.
- Mauricio Espinoza, Asunción Gómez-Pérez and Eduardo Mena, "LabelTranslator - A Tool to Automatically Localize an Ontology", Proc. of 5th European Semantic Web Conference (ESWC'08), Tenerife (Spain), Springer Verlag LNCS, ISBN 978-3-540-68233-2, ISSN-0302-9743, pp. 792-796, June 2008. demo paper.

Contact: Asunción Gómez-Pérez <asun@fi.upm.es>

MiLL

Authors: Mihai Surdeanu

References: -

Description: Machine Learning Library. Includes SVM, Maximum Entropy and Perceptron classifiers under a unique and simple interface. All algorithms support multi-class problems.

Functionality: -

Technology: -

Technical Requirements: OpenNLP MaxEnt (included), libsvm.

Modules: Multiclass Perceptron; One-vs-rest Perceptron with uneven and dynamic margins; Maximum Entropy; SVM

Innovation: MiLL includes the novel Perceptron algorithm with dynamic uneven (?) margins Dr. Surdeanu designed for his ACE Information Extraction system (see publications).

Development: -

Publications:

- Mihai Surdeanu and Massimiliano Ciaramita, Robust Information Extraction with Perceptrons, Proceedings of the NIST 2007 Automatic Content Extraction Workshop (ACE07), March 2007.

Contact: Mihai Surdeanu <mihai@surdeanu.name>

OMLET & FRIES

Authors: Xavi Carreras and Lluís Padró

References: http://www.lsi.upc.edu/~nlp/omlet+fries

Description: Omlet is an open source library providing services oriented to easily develop machine-learning based applications and experiments. Fries is an open source library useful to convert natural language sentences to feature vectors suitable to be input to Machine Learning algorithms.

OMLET provides an extensible framework where new ML algorithms and techniques can be integrated, tested, and combined.

FRIES provides an expressive feature definition language that enables the extraction of advanced patterns from input data. FRIES is specially oriented to encode Natural Language sentences and corpora into feature vectors.

OMLET & FRIES are released under the GNU General Public License of the Free Software Foundation

Functionality: -

Technology: C++

Technical Requirements: -

Modules: -

Innovation: -

Development: Omlet was originally written by Xavier Carreras at TALP Research Center at Universitat Politecnica de Catalunya.

Fries is based on the code developed by Dan Roth's Cognitive Computation Group at University of Illinois at Urbana Champaign (UIUC), who we thank for allowing us to distribute our modified version under GPL. Fries generalizes some of the patterns supported by the original FEX to any regular expression, and also adds new feature-building functionalities such as consulting external files.

Publications:

- Documentation can be found at: http://www.lsi.upc.edu/~nlp/omlet+fries/index.php?option=com_content&task=view&id=29&Itemid=45

Contact: Lluís Padró <padro@lsi.upc.edu>

OMV: Ontology Metadata Vocabulary

Authors: Raúl Palma y Asunción Gómez-Pérez

References: Ontology Engineering Group (http://www.oeg-upm.net/), UPM.

Description: The Ontology Metadata Vocabulary OMV is a standard proposal for describing ontologies and related entities. The OMV metadata schema is formally represented as ontology and is designed modularly:

OMV distinguishes between the OMV Core and various OMV Extensions. The core captures information which is expected to be relevant to the majority of ontology reuse settings, while the extensions allow ontology developers and users to specify task/application-specific ontology-related information (e.g. mappings, ontology evaluation, ontology changes etc,).

Functionality: OMV provides a basis for an effective access and exchange of ontologies across the web. This standard proposal provides the basis for interoperability at tool level through a common interface to ontology registries/repositories (e.g. Oyster, Centrasite) and related applications like the semantic web gateway Watson.

Modules: OMV distinguishes between the OMV Core and various OMV Extensions. The core captures information which is expected to be relevant to the majority of ontology reuse settings, while the extensions allow ontology developers and users to specify task/application-specific ontology-related information (e.g. mappings, ontology evaluation, ontology changes etc,)

Innovation: OMV enables users from academia and industry to identify, find and apply - basically meaning to reuse - ontologies and related entities effectively and efficiently.

Development: This work has been funded by the Knowledge Web project (http://knowledgeweb.semanticweb.org/semanticportal/sewView/frames.html) and NeOn project (http://www.neon-project.org).

Publications:

- Hartmann, J.; Bontas E.; Palma R.; Gómez-Pérez, A.: "DEMO - Design Environment for Metadata Ontologies." In: Proceedings of the 3rd European Semantic Web Conference, ESWC 2006. Volume 4011. June 2006. Budva, Montenegro. Springer Berlin. 427-441.
- Hartmann, J.; Palma, R.; Sure, Y.; Haase, P.; Suarez-Figueroa, M. "OMV– Ontology Metadata Vocabulary". In: Proceedings of the International Workshop on Ontology Patterns for the Semantic Web, located at the conference International Semantic Web Conference, ISWC2005. November, 2005. Galway, Ireland.
- Palma, R., Hartmann, J.; Haase, P. "O M V - Ontology Metadata Vocabulary for the Semantic Web". Technical Report Version 2.4. 2008. http://omv.ontoware.org

Contact: Asunción Gómez-Pérez <asun@fi.upm.es>

Oyster: Distributed Ontology Registry

Authors: Raúl Palma y Asunción Gómez-Pérez

References: Ontology Engineering Group (http://www.oeg-upm.net/), UPM.

Description: Oyster is a distributed registry that exploits semantic web techniques in order to provide a solution for exchanging and re-using ontologies and related entities. It provides services for storage, cataloguing, discovery, management, and retrieval of ontology (and related entities) metadata definition and services to support the management and evolution of ontologies in distributed environments.

Functionality: With Oyster each peer has its own local registry of ontology metadata and also has access to the information of other registries, thus creating a virtual decentralized ontology metadata registry. The goal

is a decentralized knowledge sharing environment using Semantic Web technologies that allows developers to easily share ontologies and related metadata.

Innovation: Users can benefit from services like: Creating and importing metadata, Formulating advanced semantic searches, Management and synchronization of ontology change information and ontology development activities information.

Technology: The Oyster system (available at http://ontoware.org/projects/oyster2/) was designed using a service-oriented approach, and it provides a well defined API. Accessing the registry functionalities can be done using directly the API within any application, invoking the web service provided or using the included java-based GUI as a client for the distributed registry. Moreover, Oyster is available also as a Plug-in component within the NeOn toolkit (www.neon-toolkit.org).

Technical Requirements: Oyster implements OMV as the way to describe ontologies and related entities.

Development: This work has been funded by the Knowledge Web project (http://knowledgeweb.semanticweb.org/semanticportal/sewView/frames.html) and NeOn project (http://www.neon-project.org).

Publications:

- Palma, R.; Haase P.; Gómez-Pérez, A. "Oyster – Sharing and Re-using Ontologies in a Peer-to-Peer Community". Poster in the Proceedings of the 15th International World Wide Web Conference, WWW'06. May, 2006. Edinburgh, Scotland. ACM. 1009-1010
- Palma, R.; Haase, P. "Oyster - Sharing and Re-using Ontologies in a Peer-to-Peer Community". In: Semantic Web Challenge of Proceedings of the 4th International Semantic Web Conference, ISWC 2005. Volume 3729.

Contact: Asunción Gómez-Pérez <asun@fi.upm.es>

QARLA

Authors: Enrique Amigó, Julio Gonzalo

References: http://www.lsi.upc.edu/~nlp/IQMT

Description: A framework on which automatic evaluation metrics over reference outputs can be robustly combined for the purpose of machine translation, summarization or language generation system development. This tool allows to combine metrics operating at different linguistic levels (lexical, shallow syntactic, syntactic, and shallow semantic). The framework provides measures to quantify the power of metric combinations and the appropriateness of the testbed.

Functionality: -

Technology: Perl

Technical Requirements: It requires multiple references for each evaluation test case and the evaluation results from each individual metric and each reference.

Modules: -

Innovation: Metric scores can be combined into a single measure of quality, without depending on scales properties or weighting criteria.

Development: This work has been developed in the context of a PhD. thesis.

Publications:

- Enrique Amigó, Julio Gonzalo, Anselmo Peñas and Felisa Verdejo, QARLA: a framework for the evaluation of text summarization systems. Proceedings of the 43rd Meeting of the Association for Computational Linguistics (ACL 2005). 2005.
- Enrique Amigó, Julio Gonzalo, Anselmo Peñas and Felisa Verdejo. Evaluating DUC 2004 Tasks with the QARLA Framework. Proceedings of the ACL 2005 Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization. 2005.
- Enrique Amigó, Julio Gonzalo, Anselmo Peñas and Felisa Verdejo, Evaluación de resúmenes automáticos mediante QARLA. Procesamiento del Lenguaje Natural, Nº 35 pp. 59-66, 2005.
- Jesús Giménez and Enrique Amigó. IQMT: A Framework for Automatic Machine Translation Evaluation. Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'06). Genoa, Italy, 22-28 May. 2006. Departament Research Technical Report (LSI-07-29-R).

Contact: Enrique Amigó <enrique@lsi.uned.es>

TeCat

Authors: Arturo Montejo Ráez

References: http://sinai.ujaen.es/wp-content/uploads/2013/11/tecat-0.2.tar__0.gz

Description: TECAT representa la categorización de textos. Es una herramienta para la creación de etiquetas multi-clasificadores de texto automático. Con TECAT usted puede experimentar con diferentes colecciones y clasificadores con el fin de construir un multi-etiqueta.

Por favor, envía un correo a amontejo AT ujaen punto es notificando su uso.

Licencia: GPL

Functionality: -

Technology: -

Technical Requirements: -

Modules: -

Innovation: -

Development: -

Publications:

- Montejo-Ráez A., Ureña-López, L. A., Steinberger, R. Adaptive Selection of Base Classifiers in One-Against-All Learning for Large Multi-labeled Collections. Lecture Notes in Computer Science Volume 3230, 2004, pp 1-12.

Contact: Arturo Montejo Ráez <amontejo@ujaen.es>

TRIELIB

Authors: José M. Goñi-Menoyo, José C. González-Cristóbal (Universidad Politécnica de Madrid), Jorge Fombella-Mourelle, Julio Villena-Román (DAEDALUS-Data, Decisions, and Language, S.A.)

References: The "Grupo de Sistemas Inteligentes" (GSI) of Universidad Politécnica de Madrid is a well-known research group with extensive activity in several Natural Language Processing projects.

Description: TRIELIB is a trie-based software library aimed to the development of efficient implementations of lexical and morphological components for the management of huge lexicons. Its main feature is that access time for a lexical entry is independent of the lexical database size. In general, the library is an indexing system for huge databases, lexical or not. For instance, an indexing and retrieval system for information retrieval has beenbuilt on top of TRIELIB.

Functionality: TRIELIB is a management library for indexing textual entries and its associated information.

Technology: TRIELIB is implemented in standard C++ suitable for being installed on Linux or Windows platforms.

Technical Requirements: GNU C++ development platform.

Modules: Basic indexing management library and lexical information management library.

Innovation: TRIELIB has been used for building huge lexical access systems and for implementing a continuation-based morphological analyser based on allomorph concatenation model. It benefits from the lexical access system. In addition, it has been also used for implementing an indexing and retrieval system for information retrieval purposes.

Development: TRIELIB is the result of the joint collaboration of a team of researchers from Universidad Politécnica de Madrid and from the company DAEDALUS-Data, Decisions and Language, S.A., a spin-off of the GSI university research group. It has been developed during more than 10 years, as a result of the experience gained in several research project.

Publications:

- Goñi-Menoyo, J.M.; Fombella-Mourelle, J.; González-Cristóbal, J.C.; and Villena- Román, J. (2006). Biblioteca "TRIELIB". Guía de uso. Informe técnico. Universidad Politécnica de Madrid y DAEDALUS-Data, Decisions, and Language.

- Right properties / owner: Universidad Politécnica de Madrid and DAEDALUS-Data, Decisions and Language, S.A. The exploitation rights are currently transferred to DAEDALUS-Data, Decisions and Language, S.A.

Contact: José Miguel Goñi-Menoyo <josemiguel.goni@upm.es>

5. Otros servicios y *know-how*

Morphosyntactic Annotation in Spanish Service (PoS and lemmatization)

Authors: Antonio Moreno-Sandoval, Head of the Laboratorio de Lingüística Informática (Computational Linguistics Lab) and José María Guirao, Sennior Programmer References: The Laboratorio de Lingüística Informática of Universidad Autónoma de Madrid (LLI- UAM, http://www.lllf.uam.es) is a well-known research laboratory. This lab is linked by the "Bookmarks for Corpus-based Linguists" site by David Lee, as the main reference for Spanish corpora (http://personal.cityu.edu.hk/~davidlee/devotedtocorpora/corpora2.htm)

Description: GRAMPAL is a morphosyntactic tagger based on a large lexicon and with a disambiguation process based on statistical training. It can be adapted to any language register, ie. spontaneous speech, text corpora, or child language. The precision is over the 95% for any register, but the tagger reaches specially good results with spontaneous speech. The service offered combines the automatic tagging with manual revision of the annotation by linguist experts, providing a totally reliable annotation.

Functionality: From every given input text, GRAMPAL outputs the part-of-speech tagging and lemmatisation of every term. The system is trained both for spontaneous speech and written Spanish.

Technology: GRAMPAL is implemented in C++ in a linux platform. The technology is a hybrid system based on a large lexicon and in statistical disambiguation.

Technical Requirements: The service is obtained through an agreement between both parts. It works like a translation service, that is, the client sends the corpus and the annotated and verified version is returned. This service can be provided both for written and spoken resources. The output can be delivered in any format, ie., XML, plain text and any tagset.

Modules: Automatic tagging, and (2) Manual revision by expert linguists, controlled by devoted tool.

Innovation: GRAMPAL's main innovation was obtained when it was used in the tagging of the C- ORAL-ROM corpus, an EU-funded project of spontaneous speech resources. It must be pointed out that GRAMPAL has been specially adapted for spoken Spanish, what means a special training with spoken corpora for the disambiguation of PoS candidates.

Development: GRAMPAL was the result of a PhD dissertation in 1991, and it has been developed for more than 10 years long by a team of linguists and engineers, as a result of the experience gained in several funded research project.

Publications:

- Moreno, A. & Guirao, J.M. "Morpho-syntactic Tagging of the Spanish C- ORAL-ROM Corpus: Methodology, Tools and Evaluation.", in Spoken Language Corpus and Linguistic Informatics, John Benjamins, 2006.

- Guirao, J.M. y Moreno, A. A "toolbox" for tagging the Spanish C-ORAL-ROM corpus IV International Conference on Language Resources and Evaluation (LREC2004) Proceedings, 2004.

Contact: Antonio Moreno-Sandoval <antonio.msandoval@uam.es>

Systemized Process of Corpora Development

Authors: Marta Garrote y Antonio Moreno-Sandoval

Reference: Laboratorio de Lingüística Informática, UAM: http://www.lllf.uam.es

Description: Systemized process to collect both spontaneous speech and written corpora composed of the following stages (each stage is manually revised by more than one person):

1. Preliminary design considering participants, their socio-linguistic features (age, gender, demographics, linguistic origin, education, etc) and the communicative context. This information may be modified depending on the goals of the study. This design may also be modified according to the variables consi ered in the study.

2. Data collecting (recording, video captures, editing, etc.)

3. Orthographic transcription (both normative and real speech).

4. Prosodic annotation, marking pauses, vocal lengthening, overlaps, interruptions, intonation, etc.

5. Alignment of text-sound units in utterances.

6. Semi-automatic morpho-syntact c annotation (part-of-speech and lemmas).

7. Automatic phonological annotation.

Functionality: Besides the possible application of these data collections, this methodology allows automatic information processing and retrieval at each linguistic level, since all annotations are standardized using XML.

Technology: The complete process involves different technologies such as word sense disambiguation, part-of-speech tagging and lemmatization.

Technical Requirements: This is a service accessible after signing an agreement or contract with LLI-UAM.

Innovation: This service is presented as a result of different R&D projects. Each project focused on the development of one level of analysis, obtaining a complete toolkit. The added value is the systemized methodology that has been successfully proved in the elaboration of different customized corpora.

Development: The work has been mainly supported by public funding through research projects. The methodology was acquired during the C-ORAL-ROM corpus, a EU-funded project of the 5 FP.

Publications:

- Cresti, E. Moneglia, M .(eds). 2005. C-ORAL-ROM: Integrated Reference Corpora for Spoken Roman Languages. Amsterdam. John Benjamins.

- Garrote. M. CHIEDE: Corpus de habla infantil espontánea del español. PhD Dissertation. Universidad Autónoma de Madrid. 2008.

Contact: Antonio Moreno-Sandoval <antonio.msandoval@uam.es>

[i] http://www.snomed.org/

[ii] Provided by the Spanish Ministry of Health and Consume

http://www.msc.es/estadEstudios/estadisticas/docs/diccionarioSiglasMedicas.pdf