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.

Challenges: GRAMPAL has been one of the research topics of the LLI-UAM (http://www.lllf.uam.es) directed by Antonio Moreno

Publications: (3) 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.

MAVIR 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 considered 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-syntactic 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.

Challenges: The development of this methodology has proved necessary to improve other processes at LLI-UAM.

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.

MAVIR contact: Antonio Moreno-Sandoval <antonio.msandoval@uam.es>

Information Retrieval System Based on Conceptual Clustering

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.

MAVIR contact: Juan M. Cigarrán <juanci@lsi.uned.es>, Julio Gonzalo <julio@lsi.uned.es>

ARIES: A Lexical Base and Platform

Authors: José M. Goñi-Menoyo, José C. González-Cristóbal (Universidad Politécnica de Madrid), Antonio Moreno Sandoval (Universidad 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.

MAVIR contact: José Miguel Goñi-Menoyo <josemiguel.goni@upm.es>

TRIELIB: A software library for building huge lexical databases access systems

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.

MAVIR contact: José Miguel Goñi-Menoyo <josemiguel.goni@upm.es>

SQUASH: A Question Answering System for Spanish

Authors: C de Pablo-Sánchez, P Martínez-Fernández, JL Martínez-Fernandez, MT Vicente-Díez (Univ. Carlos III de Madrid), A Moreno Sandoval, A García-Ledesma (Univ. Autónoma de Madrid).

References : The "Laboratorio de Bases de Datos Avanzadas" (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 an extensive activity in several Natural Language Processing and Information Retrieval projects.

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 *"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 to 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 STILUSis used for Language Analysis.

Modules: Modules for pre-processing information include language analysis and indexing libraries. Modules for online querying perform question classification, question analysis, query generation, sentence retrieval, answer extraction and answerranking.

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 is the result of the joint collaboration of a multidisciplinary team of researchers from Universidad Carlos III de Madrid, Universidad Politécnica de Madrid and Universidad Autónoma de Madrid. It has been developed for more than 4 years, from the experience gained in several 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.

MAVIR contact: Paloma Martínez Fernández <pmf@inf.uc3m.es>

INTERACTOR (Natural Interaction Platform)

Authors: Paloma Martínez, head of the Advanced Databases Laboratory (Labda). F. Javier Calle, Dolores Cuadra, Sennior Researchers. David del Valle, Jessica Rivero, Junior Researchers.

References: Natural Interaction subgroup at Advanced Databases Laboratory (Labda): http://basesdatos.uc3m.es/index.php?id=202&L=0

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) 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 spatio-temporal database. The rest of the Interactor system (other agents) is also implemented in Java. All the mentioned databases are set on the Oracle™ 10g DBMS.

Technical Requirements: Interactor is currently running on a Sun Fire X2200 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, (2) Interactor components, (3) Dialogue management agents, and (4) Situation agents. For new interaction domains implementation, (5) the Cognos toolkit (for implementing already formalised knowledge; currently being improved),and (6) the Tracer (tool for consulting the interaction traces).

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. © 2008 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.
- Rivero, J., Cuadra, D., Valle, D., Calle, F.J. *Incorporating Circumstantial Knowledge Influence over Natural Interaction*. Agent-Oriented Sw Engineering Challenges for Ubiquitous and Pervasive Computing W., 2007.

MAVIR contact: F. Javier Calle Gómez <fcalle@inf.uc3m.es>