Data-driven computational linguistics at FaMAF-UNC, Argentina

Laura Alonso i Alemany and **Gabriel Infante-Lopez**
Grupo de Procesamiento de Lenguaje Natural
Sección de Ciencias de la Computación
Facultad de Matemática, Astronomía y Física
Universidad Nacional de Córdoba
Córdoba, Argentina
{gabriel|alemany}@famaf.unc.edu.ar

Abstract

This paper provides a survey of some ongoing research projects in computational linguistics within the group of Natural Language Processing at the University of Córdoba, Argentina. We outline our future plans and spotlight some opportunities for collaboration.

1 Introduction

In this paper we present our group, describe its members, research agenda, interests and possible collaboration opportunities. The research agenda of the NLP group contains diverse lines of work. As a group, we have a special interest in producing language technologies for our languages, at a level comparable in performance with the state-of-the-art technology for English. We are developing such technology by deeply understanding its underling models and either adapting them to our languages or by creating new ones.

In this paper we present only those related to Natural Language Parsing and data-driven characterisation of linguistic phenomena. For both lines we provide a small survey of our results so far, we describe our current research questions and we spotlight possible opportunities of collaboration.

The paper is organized as follows. The following Section describes the group, its composition, projects and goals. Section 3 briefly introduces the research agenda related to natural language parsing and structure finding. Section 4 sketches the work on data-driven characterisation of linguistic phenomena in three main parts: semi-structured text mining, characterisation of verbal behaviour and mining of relations in biomedical text. Finally, Section 5 presents outlines our overall vision for collaboration with other researchers in the Americas.

2 Description of the group

The NLP Group[1] is part of the Computer Science section at the Facultad de Matemática, Astronomía y Física, at the Universidad Nacional de Córdoba. The group was started in 2005, with two full time researchers who had just got their doctorate degree in Amsterdam and Barcelona. Then, in 2009 and 2010 three more full-time researchers joined the group, coming from the Universities of Geneva and Nancy.

As of 2010, the group has 5 faculty researchers, 4 PhD students and several undergraduate students. The computer science section has around 20 members – including the NLP group, faculty members and PhD students.

The faculty researchers are, by alphabetical order:

- Laura Alonso Alemany, working in text mining and data-driven systematization of language.
- Carlos Areces, investigating different reasoning tasks and their applications in natural language processing.
- Luciana Benotti, investigates the addition of pragmatic abilities into dialogue systems.
- Paula Estrella, working in Machine Translation.
- Gabriel Infante-Lopez, working on Natural Language Parsing and Structure Finding.

[1] http://www.cs.famaf.unc.edu.ar/~pln/

Proceedings of the NAACL HLT 2010 Young Investigators Workshop on Computational Approaches to Languages of the Americas, pages 8–14, Los Angeles, California, June 2010.

One of the main aims of the group has been education, both at undergraduate and graduate levels. Computer Science is an under-developed area in Argentina, and Natural Language Processing even more so. When the group was created, there were very few NLP researchers in the country, and they worked in isolation, with little connection to other researchers from neighbouring countries. One of the strategic goals of our University and of the NLP group itself were to create a critical mass of researchers in NLP. To that aim, we worked on incorporating researchers to our group and establishing relations with other groups. Researchers were incorporated via special programmes from both the Faculty and the Argentinean Government to increase the number of doctors in Computer Science in the scientific system in Argentina.

Most of our efforts in the first years went to raise awareness about the area and provide foundational and advanced courses. This policy lead to a significant number of graduation theses[2] and to the incorporation of various PhD students to our group.

We taught several undergraduate and graduate courses on various NLP topics at our own University, at the University of Río Cuarto, at the University of Buenos Aires and at the Universidad de la República (Uruguay), as well as crash courses at the Society for Operative Investigations (SADIO) and at the Conferencia Latinoamericana de Informática (CLEI 2008). We also gave several talks at various universities in the country, and participated in local events, like JALIMI'05 (*Jornadas Argentinas de Lingüística Informática: Modelización e Ingeniería*) or the Argentinean Symposium on Artificial Intelligence.

Since the beginning of its activities, the group has received funding for two major basic research projects, funded by the Argentinean Agency for the Development of Science and Technology. A third such project is pending approval.

We have a special interest in establishing working relations and strengthening the synergies with the research community in NLP, both within South America and the rest of the world. We have had scientific and teaching exchanges with the NLP group in Montevideo, Uruguay. From that collaboration, the Microbio project emerged[3], bringing together researchers on NLP from Chile, Brazil, Uruguay, France and Argentina. This project was funded by each country's scientific institutions (MinCyT, in the case of Argentina) within STIC-AmSud[4], a scientific-technological cooperation programme aimed to promote and strengthen South America regional capacities and their cooperation with France in the area of Information Technologies and Communication. Within this project, we hosted the kick-off workshop on February 2008, with attendants representing all groups in the project.

We have also had billateral international cooperation in some smaller projects. Together with the CNR-INRIA in Rennes, France, we have worked in a project concerning the smallest grammar problem. We tackle the same problem, finding small grammars in two different domains: ADN sequences and Natural Language sentences. In collaboration with several universities in Spain (UB, UOC, UPC, EHU/UPV), we have taken part in the major basic research programme KNOW[5], aiming to aggregate meaning, knowledge and reasoning to current information technologies. This project has now received funding to carry on a continuating project[6].

Moreover, we are putting forward some proposals for further international collaboration. Following the path opened by the Microbio project, we are working on a proposal to the Ecos Sud programme for joint collaboration with research teams in France[7].

We are also working in strengthening relations within Argentinean NLP groups. To that aim, we are collaborating with the NLP group at the University of Buenos Aires in the organization of the School on Computational Linguistics ELiC[8], with several grants for students sponsored by NAACL. We are also putting forward a proposal for a workshop on

[2] http://cs.famaf.unc.edu.ar/~pln/Investigacion/tesis_grado/tesis_grado.html

[3] http://www.microbioamsud.net/

[4] http://www.sticamsud.org/

[5] KNOW project: http://ixa.si.ehu.es/know.

[6] Representation of Semantic Knowledge, TIN2009-14715-C04-03 (Plan Nacional de I+D+i 2008-2011).

[7] ECOS-SUD programme: http://www.mincyt.gov.ar/coopinter_archivos/bilateral/francia.htm.

[8] ELiC school on Computational Linguistics: http://www.glyc.dc.uba.ar/elic2010/.

NLP to be co-located with the IBERAMIA conference on Artificial Intelligence, to be held at Bahía Blanca on November 2010.

3 Natural Language Parsing and Structure Finding

3.1 Unsupervised Parsing

Unsupervised parsing of Natural Language Syntax is a key technology for the development of language technology. It is specially important for languages that have either small treebanks or none at all. Clearly, there is a big difference between producing or using a treebank for evaluation and producing or using them for training. In the former case, the size of the treebank can be significantly smaller. In our group, we have investigated different approaches to unsupervised learning of natural language. and we are currently following two different lines, one that aims at characterizing the potential of a grammar formalism to learn a given treebank structure and a second that uses only regular automata to learn syntax.

Characterization of Structures In (Luque and Infante-Lopez, 2009) we present a rather unusual result for language learning. We show an upper bound for the performance of a class of languages when a grammar from that class is used to parse the sentences in any given treebank. The class of languages we studied is the defined by Unambiguous Non-Terminally Separated (UNTS) grammars (Clark, 2006). UNTS grammars are interesting because, first, they have nice learnability properties like PAC learnability (Clark, 2006), and, second, they are used as the background formalism that won the Omphalos competition (Clark, 2007). Our strategy consists on characterizing all possible ways of parsing all the sentences in a treebank using UNTS grammars, then, we find the one that is closest to the treebank. We show that, in contrast to the results obtained for learning formal languages, UNTS are not capable of producing structures that score as state-of-the-art models on the treebanks we experimented with.

Our results are for a particular, very specific type of grammar. We are currently exploring how to widen our technique to provide upper bounds to a more general class of languages. Our technique does not state how to actually produce a grammar that performs as well as the upper bound, but it can be useful for determining how to transform the training material to make upper bounds go up. In particular we have defined a generalization of UNTS grammars, called k-l-UNTS grammars, that transform a word w in the training material in a 3-uple $\langle \alpha, w, \beta \rangle$ where α contains the k previous symbols to w and β contains the l symbols following w. Intuitively, k-l-UNTS augments each word with a variable length context. It turns out that the resulting class of languages is more general than UNTS grammars: they are PAC learnable, they can be learned with the same learning algorithm as UNTS and, moreover, their upper bound for performance is much higher than for UNTS. Still, it might be the case that the existing algorithm for finding UNTS is not the right one for learning the structure of a treebank, it might be the case that strings in the PTB have not been produced by a k-l-UNTS grammar. We are currently investigating how to produce an algorithm that fits better the structure given in a treebank.

Learning Structure Using Probabilistic Automata DMV+CCM (Klein and Manning, 2004; Klein and Manning, 2002) is a probabilistic model for unsupervised parsing, that can be successfully trained with the EM algorithm to achieve state of the art performance. It is the combination of the Constituent-Context Model, that models unlabeled constituent parsing, and the Dependency Model with Valence, that models projective dependency parsing. On the other hand, CCM encodes the probability that a given string of POS tags is a constituent. DMV is more of our interest in this work, because it encodes a top-down generative process where the heads generate their dependents to both directions until there is a decision to stop, in a way that resembles successful supervised dependency models such as in (Collins, 1999). The generation of dependents of a head on a specific direction can be seen as an implicit probabilistic regular language generated by a probabilistic deterministic finite automaton.

Under this perspective, the DMV model is in fact an algorithm for learning several automata at the same time. All automata have in common that they have the same number of states and the same number of arcs between states, which is given by the definition of the DMV model. Automata differ in that they have different probabilities assigned to the transitions. The simple observation that DMV actually suppose a fixed structure for the automata it induces might explain its poor performance with freer order languages like Spanish. Using our own implementation (see (Luque, 2009)) we have empirically tested that DMV+CMV works well in languages with strict word order, like English, but for other languages with freer word order, like Spanish, DMV+CMV performance decreases dramatically. In order to improve DMV+CCM performance for this type of languages, the structure of the automaton might be modified, but since the DMV model has an *ad hoc* learning algorithm, a new parametric learning algorithm has to be defined. We are currently investigating different automaton structures for different languages and we are also investigating not only the induction of the parameters for fixed structure, but also inducing the structure of the automata itself.

3.2 Smallest Grammar and Compression for Natural Language

The smallest grammar problem has been widely studied in the literature. The aim of the problem is to find the smallest (smallest in the sense of number of symbols that occur in the grammar) context free grammar that produces only one given string. The smallest grammar can be thought as a relaxation of the definition of Kolmogorov Complexity where the complexity is given by a context free grammar instead of a Turing machine. It is believed that the smallest grammar can be used both for computing optimal compression codes and for finding meaningful patterns in strings.

Moreover, since the procedure for finding the smallest grammar is in fact a procedure that assigns a tree structure to a string, the smallest grammar problem is, in fact, a particular case of unsupervised parsing that has a very particular objective function to be optimized.

Since the search space is exponentially big, all existing algorithms are in fact heuristics that look for a small grammar. In (Carrascosa et al., 2010) we presented two algorithms that outperform all existing heuristics. We have produce and algorithm that produces 10% smaller grammars for natural language strings and 1.5% smaller grammars for DNA sequences.

Even more, we show evidence that it is possible to find grammars that share approximately the same small score but that have very little structure in common. Moreover, the structure that is found by the smallest grammar algorithm for the sentences in PTB have little in common with the structure that the PTB defines for those sentences.

Currently, we are trying to find answers to two different questions. First, is there a small piece of structure that is common to all grammars having comparable sizes? and second, can the grammars that are found by our algorithms be used for improving compression algorithms?

4 Data-driven characterisation of linguistic phenomena

4.1 Semi-structured text mining

One of our lines of research is to apply standard text mining techniques to unstructured text, mostly user generated content like that found in blogs, social networks, short messaging services or advertisements. Our main corpus of study is constituted by classified advertisements from a local newspaper, but one of our lines of work within this project is to assess the portability of methods and techniques to different genres.

The goals we pursue are:

creating corpora and related resources, and making them publicly available. A corpus of newspaper advertisements and a corpus of short text messages are underway.

normalization of text bringing ortographic variants of a word (mostly abbreviations) to a canonical form. To do that, we apply machine learning techniques to learn the parameters for edit distances, as in (Gómez-Ballester et al., 1997; Ristad and Yanilos, 1998; Bilenko and Mooney, 2003; McCallum et al., 2005; Oncina and Sebban, 2006). We build upon previous work on normalization by (Choudhury et al., 2007; Okazaki et al., 2008; Cook and Stevenson, 2009; Stevenson et al., 2009). Preliminary results show a significant improvement of learned distances over standard distances.

syntactic analysis applying a robust shallow parsing approach aimed to identify entities and their modifiers.

ontology induction from very restricted domains, to aid generalization in the step of information extraction. We will be following the approach presented in (Michelson and Knoblock, 2009).

information extraction inducing templates from corpus using unsupervised and semi-supervised techniques, and using induced templates to extract information to populate a relational database, as in (Michelson and Knoblock, 2006).

data mining applying traditional knowledge discovery techniques on a relational database populated by the information extraction techniques used in the previous item.

This line of research has been funded for three years (2009-2012) by the Argentinean Ministry for Science and Technology, within the PAE project, as a PICT project (PAE-PICT-2007-02290).

This project opens many opportunities for collaboration. The resulting corpora will be of use for linguistic studies. The results of learning edit distances to find abbreviations can also be used by linguists as an input to study the regularities found in this kind of genres, as proposed in (Alonso Alemany, 2010).

We think that some joint work on learning string edit distances would be very well integrated within this project. We are also very interested in collaborations with researchers who have some experience in NLP in similar genres, like short text messages or abbreviations in medical papers.

Finally, interactions with data mining communities, both academic and industrial, would surely be very enriching for this project.

4.2 Characterisation of verbal behaviour

One of our research interests is the empirical characterization of the subcategorization of lexical items, with a special interest on verbs. This line of work has been pursued mainly within the KNOW project, in collaboration with the UB-GRIAL group[9].

Besides the theoretical interest of describing the behaviour of verbs based on corpus evidence, this line has an applied aim, namely, enriching syntactic analyzers with subcategorization information, to help resolving structural ambiguities by using lexical information. We have focused on the behaviour of Spanish verbs, and implemented some of our findings as a lexicalized enhancement of the dependency grammars used by Freeling[10]. An evaluation of the impact of this information on parsing accuracy is underway.

We have applied clustering techniques to obtain a corpus-based characterization of the subcategorization behaviour of verbs (Alonso Alemany et al., 2007; Castellón et al., 2007). We explored the behaviour of the 250 most frequent verbs of Spanish on the SenSem corpus (Castellón et al., 2006), manually annotated with the analysis of verbs at various linguistic levels (sense, aspect, voice, type of construction, arguments, role, function, etc.). Applying clustering techniques to the instances of verbs in these corpus, we obtained coarse-grained classes of verbs with the same subcategorization. A classifier was learned from considering clustered instances as classes. With this classifier, verbs in unseen sentences were assigned a subcategorization behaviour.

Also with the aim of associating subcategorization information to verbs using evidence found in corpora, we developed IRASubcat (Altamirano, 2009). IRASubcat[11]. is a highly flexible system designed to gather information about the behaviour of verbs from corpora annotated at any level, and in any language. It identifies patterns of linguistic constituents that co-occur with verbs, detects optional constituents and performs hypothesis testing of the co-occurrence of verbs and patterns.

We have also been working on connecting predicates in FrameNet and SenSem, using WordNet synsets as an interlingua (Alonso Alemany et al., SEPLN). We have found many dissimilarities between FrameNet and SenSem, but have been able to connect some of their predicates and enrich these resources with information from each other.

We are currently investigating the impact of different kinds of information on the resolution of pp-attachment ambiguities in Spanish, using the ANCORA corpus (Taulé et al., 2006). We are exploring

[9] `http://grial.uab.es/`

[10] `http://www.lsi.upc.edu/~nlp/freeling/`

[11] `http://www.irasubcat.com.ar/`

the utility of various WordNet-related information, like features extracted from the Top Concept Ontology, in combination with corpus-based information, like frequencies of occurrence and co-occurrence of words in corpus.

The line of research of characterisation of verbal behaviour presents many points for collaboration. In collaboration with linguists, the tools and methods that we have explained here provide valuable information for the description and systematization of subcategorization of verbs and other lexical pieces. It would be very interesting to see whether these techniques, that have been successfully applied to Spanish, apply to other languages or with different resources. We are also interested in bringing together information from different resources or from different sources (corpora, dictionaries, task-specific lexica, etc.), in order to achieve richer resources. We also have an interest for the study of hypothesis testing as applied to corpus-based computational linguistics, to get some insight on the information that these techniques may provide to guide research and validate results.

4.3 Discovering relations between entitites

As a result of the Microbio project, we have developed a module to detect relations between entities in biomedical text (Bruno, 2009). This module has been trained with the GENIA corpus (Kim et al., 2008), obtaining good results (Alonso Alemany and Bruno, 2009). We have also explored different ways to overcome the data sparseness problem caused by the small amount of manually annotated examples that are available in the GENIA corpus. We have used the corpus as the initial seed of a bootstrapping procedure, generalized classes of relations via the GENIA ontology and generalized classes via clustering. Of these three procedures, only generalization via an ontology produced good results. However, we have hopes that a more insightful characterization of the examples and smarter learning techniques (semi-supervised, active learning) will improve the results for these other lines.

Since this area of NLP has ambitious goals, opportunities for collaboration are very diverse. In general, we would like to join efforts with other researchers to solve part of these complex problems, with a special focus in relations between entities and semi-supervised techniques.

5 Opportunities for Collaboration

We are looking for opportunities of collaboration with other groups in the Americas, producing a synergy between groups. We believe that we can articulate collaboration by identifying common interests and writing joint proposals. In Argentina there are some agreements for billateral or multi-lateral collaboration with other countries or specific institutions of research, which may provide a framework for starting collaborations.

We are looking for collaborations that promote the exchange of members of the group, specially graduate students. Our aim is to gain a level of collaboration strong enough that would consider, for example, co-supervision of PhD students. Ideally, co-supervised students would spend half of their time in each group, tackle a problem that is common for both groups and work together with two supervisors. The standard PhD scholarship in Argentina, provided by Conicet, allows such modality of doctorate studies, as long as financial support for travels and stays abroad is provided by the co-supervising programme. We believe that this kind of collaboration is one that builds very stable relations between groups, helps students learn different research idiosyncrasies and devotes specific resources to maintain the collaboration.

References

Laura Alonso Alemany and Santiago E. Bruno. 2009. Learning to learn biological relations from a small training set. In *CiCLing*, pages 418–429.

Laura Alonso Alemany, Irene Castellón, and Nevena Tinkova Tincheva. 2007. Obtaining coarse-grained classes of subcategorization patterns for spanish. In *RANLP'07*.

Laura Alonso Alemany, Irene Castellón, Egoitz Laparra, and German Rigau. SEPLN. Evaluación de métodos semi-automáticos para la conexión entre FrameNet y SenSem. In *2009*.

Laura Alonso Alemany. 2010. Learning parameters for an edit distance can learn us tendencies in user-generated content. Invited talk at *NLP in the Social Sciences*, Instituto de Altos Estudios en Psicologia y Ciencias Sociales, Buenos Aires, Argentina, May 2010.

I. Romina Altamirano. 2009. Irasubcat: Un sistema para adquisición automática de marcos de subcategorización de piezas léxicas a partir de corpus. Master's thesis, Facultad de Matemática, Astronomía y Física, Universidad Nacional de Córdoba, Argentina.

Mikhail Bilenko and Raymond J. Mooney. 2003. Adaptive duplicate detection using learnable string similarity measures. In *Proceedings of the ninth ACM SIGKDD*.

Santiago E. Bruno. 2009. Detección de relaciones entre entidades en textos de biomedicina. Master's thesis, Facultad de Matemática, Astronomía y Física, Universidad Nacional de Córdoba, Argentina.

Rafael Carrascosa, François Coste, Matthias Gallé, and Gabriel Infante-Lopez. 2010. Choosing Word Occurrences for the Smallest Grammar Problem. In *Proceedings of LATA 2010*. Springer.

Irene Castellón, Ana Fernández-Montraveta, Glòria Vázquez, Laura Alonso, and Joanan Capilla. 2006. The SENSEM corpus: a corpus annotated at the syntactic and semantic level. In *5th International Conference on Language Resources and Evaluation (LREC 2006)*.

Irene Castellón, Laura Alonso Alemany, and Nevena Tinkova Tincheva. 2007. A procedure to automatically enrich verbal lexica with subcategorization frames. In *Proceedings of the Argentine Simposium on Artificial Intelligence, ASAI'07*.

Monojit Choudhury, Rahul Saraf, Vijit Jain, Animesh Mukherjee, Sudeshna Sarkar, and Anupam Basu. 2007. Investigation and modeling of the structure of texting language. *Int. J. Doc. Anal. Recognit.*, 10(3):157–174.

Alexander Clark. 2006. Pac-learning unambiguous nts languages. In *International Colloquium on Grammatical Inference*, pages 59–71.

Alexander Clark. 2007. Learning deterministic context free grammars: the omphalos competition. *Machine Learning*, 66(1):93–110.

M. Collins. 1999. *Head-Driven Statistical Models for Natural Language Parsing*. Ph.D. thesis, University of Pennsylvania, PA.

Paul Cook and Suzanne Stevenson. 2009. An unsupervised model for text message normalization. In *Workshop on Computational Approaches to Linguistic Creativity*. NAACL HLT 2009.

E. Gómez-Ballester, M. L. Micó-Andrés, J. Oncina, and M. L. Forcada-Zubizarreta. 1997. An empirical method to improve edit-distance parameters for a nearest-neighbor-based classification task. In *VII Spanish Symposium on Pattern Recognition and Image Analysis*, Barcelona, Spain.

Jin D. Kim, Tomoko Ohta, and Jun'ichi Tsujii. 2008. Corpus annotation for mining biomedical events from literature. *BMC Bioinformatics*, 9(1).

Dan Klein and Christopher D. Manning. 2002. A generative constituent-context model for improved grammar induction. In *ACL*, pages 128–135.

Dan Klein and Christopher D. Manning. 2004. Corpus-based induction of syntactic structure: Models of dependency and constituency. In *Proc. of ACL 42*.

Franco Luque and Gabriel Infante-Lopez. 2009. Upper bounds for unsupervised parsing with unambiguous non-terminally. In *International Workshop Computational Linguistic Aspects of Grammatical Inference. EACL*, Greece.

Franco M. Luque. 2009. Implementation of the DMV+CCM parser. `http://www.cs.famaf.unc.edu.ar/~francolq/en/proyectos/dmvccm`.

Andrew McCallum, Kedar Bellare, and Fernando Pereira. 2005. A conditional random field for discriminatively-trained finite-state string edit distance. In *Proceedings of the Proceedings of the Twenty-First Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-05)*, pages 388–395, Arlington, Virginia. AUAI Press.

Matthew Michelson and Craig A. Knoblock. 2006. Phoebus: a system for extracting and integrating data from unstructured and ungrammatical sources. In *AAAI'06: proceedings of the 21st national conference on Artificial intelligence*, pages 1947–1948. AAAI Press.

Matthew Michelson and Craig A. Knoblock. 2009. Exploiting background knowledge to build reference sets for information extraction. In *Proceedings of the 21st International Joint Conference on Artificial Intelligence (IJCAI-2009)*, Pasadena, CA.

Naoaki Okazaki, Sophia Ananiadou, and Jun'ichi Tsujii. 2008. A discriminative alignment model for abbreviation recognition. In *COLING '08: Proceedings of the 22nd International Conference on Computational Linguistics*, pages 657–664, Morristown, NJ, USA. Association for Computational Linguistics.

José Oncina and Marc Sebban. 2006. Learning stochastic edit distance: Application in handwritten character recognition. *Pattern Recognition*, 39(9):1575–1587.

E. S. Ristad and P. N. Yanilos. 1998. Learning string edit distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20:522–532.

Mark Stevenson, Yikun Guo, Abdulaziz Al Amri, and Robert Gaizauskas. 2009. Disambiguation of biomedical abbreviations. In *BioNLP '09: Proceedings of the Workshop on BioNLP*, pages 71–79, Morristown, NJ, USA. Association for Computational Linguistics.

M. Taulé, M.A. Martí, and M. Recasens. 2006. Ancora: Multilevel annotated corpora for catalan and spanish. In *LREC'06*.