Natural Language Science and Technology at the University of Lisbon, Department of Informatics: the NLX Group

António Branco

University of Lisbon, Department of Informatics
NLX - Natural Language and Speech Group
Antonio.Branco@di.fc.ul.pt

Abstract

This is a brief presentation of the NLX-Natural Language and Speech Group and its past and ongoing activities and results as of July 2009.
Index Terms: natural language processing, human language technology, computational linguistics, laboratory, Portuguese, University of Lisbon, NLX Group

1. Introduction

NLX Group is the Natural Language and Speech Group of the Department of Informatics of the University of Lisbon, Faculty of Sciences. This is a presentation of this group and its major past and ongoing results and activities as of July 2009.

For further and updated information, check its website at http://nlx.di.fc.ul.pt.

2. Research and Development

2.1. Mission

At the NLX Group, we aim at pursuing research and development (R&D) activities in the field of artificial intelligence and cognitive science, with special focus on speech and natural language interaction.

2.2. Sponsors

Our activities are undertaken with the support of a number of sponsors under individual fellowships granted to its members and under contracts for R&D projects granted in open competitive calls assigned by independent experts.

The sponsors contributing with the largest funding volume have been FCT-Fundação para a Ciência e Tecnologia of the Portuguese MCTES-Ministério da Ciência e Tecnologia, e Ensino Superior and the 6th and 7th Framework Programme of the European Commission.

The list of sponsors also include the Luso-American Foundation, the British Council Portugal, and the former GRICES-Gabinete de Relações Internacionais da Ciência e do Ensino Superior from MCTES.

2.3. Team

At present, our team comprises 13 elements. Besides 2 faculty, it includes 5 PhD students, 1 MA student and other 4 research assistants. In the past, it counted on the collaboration of 16 former members who have contributed to the group activities. It was founded and is directed by the author of this paper.

2.4. Projects

We have been participating in and coordinating a number national and international successfully completed R&D projects:

- LT4eL
  Language Technology for E-learning
- QueXting (coord.)
  Question Answering in the Portuguese Web
- GramaXing (coord.)
  Computational Grammar for Deep Linguistic Processing of Portuguese
- PALPORT
  Fine-grained Psycholinguistic Assessment of Aphasia and Other Language Impairments
- TagShare (coord.)
  Tagging and Shallow Processing Tools and Resources
- LTRC
  Language Typology Resource Center
- NeXing (coord.)
  Natural Negation Modeling and Processing

At present, a major project being conducted is

- SemanticShare (coord.)
  Resources and Tools for Semantic Processing

The major goal of this project is the construction of an annotated corpus of Portuguese, part of it aligned with similar corpora for others languages, and associated processing tools.

The texts included in this corpus are annotated with manually certified grammatical representations by human experts. These deep linguistic representations are informed by advanced linguistic analysis. They encompass different layers of linguistic information and are accessed under different views. These views include corpora of the last and next generations: PropBank: With phrases labeled with semantic functions and roles; LogicalFormBank: With sentence-level semantic representations.

Results from past projects will be presented in the next sections by way of the presentation of some of the most prominent online services, tools and resources developed in their scope.

2.5. Networking and cooperation

All our projects were deployed by consortia of several groups working in partnership. We have thus entertained a range of cooperative ties with a large number of other prominent national and international institutions, including colleagues from Brazil.

2.6. European research infrastructure

We are participating in the preparatory project for a European research infrastructure for human languages:

- CLARIN
  Common Language Resources and Technology Infrastructure

2.7. Applications and online services

Our R&D projects were or are supported by public funding. In order to showcase the results we obtained in a way easy to grasp by laymen, and to pay back to the community its support, we have been ensuring the following online services:

- **XisQuê: Question Answering**
  This is a real-time open-domain factoid question answering online service (beta version) based on the Portuguese web [3, 4].
  http://xisque.di.fc.lu.pt
- **LX-Center: Linguistic Processing**
  This is a web center of online linguistic services aimed at both demonstrating a range of language technology tools and at fostering the education, research and development in natural language science and technology [2].
  http://lxcenter.di.fc.lu.pt

2.8. Language processing tools

In the course of our R&D activities, as instrumental assets for the execution of our projects, we developed or are developing a range of language technology tools and resources.

In terms of language technology, we developed a complete pipeline of shallow processing tools that handle from the basic task of sentence splitting to the named entity recognition task. This pipeline include state of the art tools for [6, 7]:

- Sentence splitting,
- Tokenization
- Nominal lemmatization
- Nominal morphological analysis
- Nominal inflection
- Verbal lemmatization
- Verbal morphological analysis
- Verbal conjugation
- POS-tagging
- Named entity recognition

On a par with these tools for language processing, we developed also some auxiliary tools that are instrumental to explore the language resources developed (to be described in the next sections):

- Annotated corpus concordancing
- Treebank browsing and concordancing
- Aligned wordnet browsing.

As for the deep processing, we are authoring:

LXGram: Computational grammar
This is a large-scale, multi-purpose precision grammar for deep linguistic processing of Portuguese [5].
For detailed information, including distribution:
http://nlxgroup.di.fc.ul.pt/lxgram.

This grammar is being developed in the international consortium Delph-in (http://www.delph-in.net).

2.9. Language resources

In terms of resources, we have been responsible for key resources for the Portuguese language, from which we highlight here, among others, the pioneering effort devoted to:

- **MWNPT-Portuguese MultiWordnet**
  This is a lexical ontology with over 17,200 manually validated concepts/synsets, linked under the semantic relations of hyponymy and hypernymy. These concepts are made of over 21,000 word senses/word forms and 16,000 lemmas from both European and American variants of Portuguese. They are aligned with the translationally equivalent concepts of the English Princeton WordNet and, transitively, of the MultiWordNets of Italian, Spanish, Hebrew, Romanian and Latin.
  For detailed information, including distribution:
  http://mwnpt.di.fc.ul.pt
- **CINTIL-International Corpus of Portuguese**
  This is a high quality, linguistically interpreted, 1 Million token corpus accurately hand tagged with respect to POS, lemmata, inflection, multi-word proper names and adverbial and closed classes. This annotated corpus was developed in close cooperation with CLUL-Centro de Linguística da Universidade de Lisboa [1].
  For detailed information, including distribution:
  http://cintil.ul.pt

2.10. Events

We have organized several national and international scientific meetings. Among these, it is worth pointing out

- **DAARC - Discourse Anaphora and Anaphor Resolution Colloquia**
  We have been responsible for the organization of successive editions of these conferences, which are the international reference forum on anaphora.

3. Cooperation and Innovation

The previous sections briefly described the expertise we have been raising and the portfolio of key technological assets we developed for the computational processing of Portuguese. We are working to further expand and exploit these assets. We are looking for new and renewed partnerships aiming at establishing both further successful R&D projects and innovative and profitable entrepreneurial initiatives.

4. References

[1] F. Barreto, A. Branco, E. Ferreira, A. Mendes, M. F. Nascimento, F. Nunes and J. Silva, 2006. Open Resources and Tools for the Shallow Processing of Portuguese, *LREC2006*

[2] A. Branco, F. Costa, E. Ferreira, P. Martins, F. Nunes, J. Silva and S. Silveira. 2009. LX-Center: A center for linguistic services, ACL-IJCNLP2009.

[3] A. Branco, L. Rodrigues, J. Silva and S. Silveira, 2008, Real-Time Open-Domain QA on the Portuguese Web, LNAI 5290.

[4] A. Branco, L. Rodrigues, J. Silva and S. Silveira, 2008, XisQuê: An Online QA Service for Portuguese, LNAI 5190.

[5] A. Branco and F. Costa, 2008, A Computational Grammar for Deep Linguistic Processing of Portuguese: LXGram, version A.4.1, Technical Report, University of Lisbon.

[6] A. Branco, F. Costa, P. Martins, F. Nunes, J. Silva and S. Silveira. 2008. "LXService: Web Services of Language Technology for Portuguese". *LREC2008*.

[7] A. Branco and J. Silva, 2006,"LX-Suite: Shallow Processing Tools for Portuguese, EACL2006.