LX-Center: a center of online linguistic services

António Branco, Francisco Costa, Eduardo Ferreira, Pedro Martins,
Filipe Nunes, João Silva and Sara Silveira
University of Lisbon
Department of Informatics
{antonio.branco, fcosta, eferreira, pedro.martins,
fnunes, jsilva, sara.silveira}@di.fc.ul.pt

Abstract

This is a paper supporting the demonstration of the LX-Center at ACL-IJCNLP-09.

LX-Center 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.

1 Introduction

This paper is aimed at supporting the demonstration of a web center of online linguistic services. These services demonstrate language technology tools for the Portuguese language and are made available to foster the education, research and development in natural language science and technology.

This paper adheres to the common format defined for demo proposals: the next Section 2 presents an extended abstract of the technical content to be demonstrated; Section 3 provides a script outline of the demo presentation; and the last Section 4 describes the hardware and internet requirements expected to be provided by the local organizer.

2 Extended abstract

The LX-Center is a web center of online linguistic services for the Portuguese language located at `http://lxcenter.di.fc.ul.pt`. This is a freely available center targeted at human users. It has a counterpart in terms of a webservice for software agents, the LXService, presented elsewhere (Branco et al., 2008).

2.1 LX-Center

The LX-Center encompasses linguistic services that are being developed, in all or part, and maintained at the University of Lisbon, Department of Informatics, by the NLX-Natural Language and Speech Group. At present, it makes available the following functionalities:

- Sentence splitting
- Tokenization
- Nominal lemmatization
- Nominal morphological analysis
- Nominal inflection
- Verbal lemmatization
- Verbal morphological analysis
- Verbal conjugation
- POS-tagging
- Named entity recognition
- Annotated corpus concordancing
- Aligned wordnet browsing

These functionalities are provided by one or more of the seven online services that integrate the LX-Center. For instance, the LX-Suite service accepts raw text and returns it sentence splitted, tokenized, POS tagged, lemmatized and morphologically analyzed (for both verbs and nominals). Some other services, in turn, may support only one of the functionalities above. For instance, the LX-NER service ensures only named entity recognition.

These are the services offered by the LX-Center:

- LX-Conjugator
- LX-Lemmatizer
- LX-Inflector
- LX-Suite
- LX-NER
- CINTIL concordancer
- MWN.PT browser

Proceedings of the ACL-IJCNLP 2009 Software Demonstrations, pages 5–8,
Suntec, Singapore, 3 August 2009.

The access to each one of these services is obtained by clicking on the corresponding button on the left menu of the LX-Center front page.

Each of the seven services integrating the LX-Center will be briefly presented in a different subsection below. Fully fledged descriptions are available at the corresponding web pages and in the white papers possibly referred to there.

2.2 LX-Conjugator

The LX-Conjugator is an online service for fully-fledged conjugation of Portuguese verbs. It takes an infinitive verb form and delivers all the corresponding conjugated forms. This service is supported by a tool based on general string replacement rules for word endings supplemented by a list of overriding exceptions. It handles both known verbs and unknown verbs, thus conjugating neologisms (with orthographic infinitival suffix).

The Portuguese verbal inflection system is a most complex part of the Portuguese morphology, and of the Portuguese language, given the high number of conjugated forms for each verb (ca. 70 forms in non pronominal conjugation), the number of productive inflection rules involved and the number of non regular forms and exceptions to such rules.

This complexity is further increased when the so-called pronominal conjugation is taken into account. The Portuguese language has verbal clitics, which according to some authors are to be analyzed as integrating the inflectional suffix system: the forms of the clitics may depend on the Number (Singular vs. Plural), the Person (First, Second, Third or Second courtesy), the Gender (Masculine vs. Feminine), the grammatical function which they are in correspondence with (Subject, Direct object or Indirect object), and the anaphoric properties (Pronominal vs. Reflexive); up to three clitics (e.g. *deu-se-lho* / gave-One-ToHim-It) may be associated with a verb form; clitics may occur in so called enclisis, i.e. as a final part of the verb form (e.g. *deu-o* / gave-It), or in mesoclisis, i.e. as a medial part of the verb form (e.g. *dá-lo-ia* / give-it-Condicional) — when the verb form occurs in certain syntactic or semantic contexts (e.g in the scope of negation), the clitics appear in proclisis, i.e. before the verb form (ex.: *não o deu* / NOT it gave); clitics follow specific rules for their concatenation.

With LX-Conjugator, pronominal conjugation can be fully parameterizable and is thus exhaustively handled. Additionally, LX-Conjugator exhaustively handles a set of inflection cases which tend not to be supported together in verbal conjugators: Compound tenses; Double forms for past participles (regular and irregular); Past participle forms inflected for number and gender (with transitive and unaccusative verbs); Negative imperative forms; Courtesy forms for second person.

This service handles also the very few cases where there may be different forms in different variants: when a given verb has different orthographic representations for some of its inflected forms (e.g. *arguir* in European vs. *argüir* in American Portuguese), all such representations will be displayed.

2.3 LX-Lemmatizer

The LX-Lemmatizer is an online service for fully-fledged lemmatization and morphological analysis of Portuguese verbs. It takes a verb form and delivers all the possible corresponding lemmata (infinitive forms) together with inflectional feature values.

This service is supported by a tool based on general string replacement rules for word endings whose outcome is validated by the reverse procedure of conjugation of the output and matching with the original input. These rules are supplemented by a list of overriding exceptions. It thus handles an open set of verb forms provided these input forms bear an admissible verbal inflection ending. Hence, this service processes both lexically known and unknown verbs, thus coping with neologisms.

LX-Lemmatizer handles the same range of forms handled and generated by the LX-Conjugator. As for pronominal conjugation forms, the outcome displays the clitic detached from the lemma. The LX-Lemmatizer and the LX-Conjugator can be used in "roll-over" mode. Once the outcome of say the LX-Conjugator on a given input lemma is displayed, the user can click over any one of the verbal forms in that conjugation table. This activates the LX-Lemmatizer on that input verb form, and then its possible lemmas, together with corresponding inflection feature values, are displayed. Now, any of these lemmas can also be clicked on, which will activate back the LX-Conjugator and will make the corresponding conjugation table to be displayed.

2.4 LX-Inflector

The LX-Inflector is an online service for the lemmatization and inflection of nouns and adjectives of Portuguese. This service is also based on a tool that relies on general rules for ending string replacement, supplemented by a list of overriding exceptions. Hence, it handles both lexically known and unknown forms, thus handling possible neologisms (with orthographic suffixes for nominal inflection).

As input, this service takes a Portuguese nominal form — a form of a noun or an adjective, including adjectival forms of past participles –, together with a bundle of inflectional feature values — values of inflectional features of Gender and Number intended for the output.

As output, it returns: inflectional features — the input form is echoed with the corresponding values for its inflectional features of Gender and Number, that resulted from its morphological analysis; lemmata — the lemmata (singular and masculine forms when available) possibly corresponding to the input form; inflected forms — the inflected forms (when available) of each lemma in accordance with the values for inflectional features entered. LX-Inflector processes both simple, prefixed or non prefixed, and compound forms.

2.5 LX-Suite

The LX-Suite is an online service for the shallow processing of Portuguese. It accepts raw text and returns it sentence splitted, tokenized, POS tagged, lemmatized and morphologically analyzed.

This service is based on a pipeline of a number of tools, including those supporting the services described above. Those tools, for lemmatization and morphological analysis, are inserted at the end of the pipeline and are preceded by three other tools: a sentence splitter, a tokenizer and a POS tagger.

The sentence splitter marks sentence and paragraph boundaries and unwraps sentences split over different lines. An f-score of 99.94% was obtained when testing it on a 12,000 sentence corpus.

The tokenizer segments the text into lexically relevant tokens, using whitespace as the separator; expands contractions; marks spacing around punctuation or symbols; detaches clitic pronouns from the verb; and handles ambiguous strings (contracted vs. non contracted). This tool achieves an f-score of 99.72%.

The POS tagger assigns a single morpho-syntactic tag to every token. This tagger is based on Hidden Markov Models, and was developed with the TnT software (Brants, 2000). It scores an accuracy of 96.87%.

2.6 LX-NER

The LX-NER is an online service for the recognition of expressions for named entities in Portuguese. It takes a segment of Portuguese text and identifies, circumscribes and classifies the expressions for named entities it contains. Each named entity receives a standard representation.

This service handles two types of expressions, and their subtypes. (i) Number-based expressions: Numbers — arabic, decimal, non-compliant, roman, cardinal, fraction, magnitude classes; Measures — currency, time, scientific units; Time — date, time periods, time of the day; Addresses — global section, local section, zip code; (ii) Name-base expressions: Persons; Organizations; Locations; Events; Works; Miscellaneous.

The number-based component is built upon handcrafted regular expressions. It was developed and evaluated against a manually constructed test-suite including over 300 examples. It scored 85.19% precision and 85.91% recall. The name-based component is built upon HMMs with the help of TnT (Brants, 2000). It was trained over a manually annotated corpus of approximately 208,000 words, and evaluated against an unseen portion with approximately 52,000 words. It scored 86.53% precision and 84.94% recall.

2.7 CINTIL Concordancer

The CINTIL-Concordancer is an online concordancing service supporting the research usage of the CINTIL Corpus.

The CINTIL Corpus is a linguistically interpreted corpus of Portuguese. It is composed of 1 Million annotated tokens, each one of which verified by human expert annotators. The annotation comprises information on part-of-speech, lemma and inflection of open classes, multi-word expressions pertaining to the class of adverbs and to the closed POS classes, and multi-word proper names (for named entity recognition).

This concordancer permits to search for occurrences of strings in the corpus and returns them together with their window of left and right context. It is possible to search for orthographic forms

or through linguistic information encoded in their tags. This service offers several possibilities with respect to the format for displaying the outcome of a given search (e.g. number of occurrences per page, size of the context window, sorting the results in a given page, hiding the tags, etc.)

This service is supported by Poliqarp, a free suite of utilities for large corpora processing (Janus and Przepiórkowski, 2006).

2.8 MWN.PT Browser

The MWN.PT Browser is an online service to browse the MultiWordnet of Portuguese.

The MWN.PT is a lexical semantic network for the Portuguese language, shaped under the ontological model of wordnets, developed by our group. It spans 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.

It includes the subontologies under the concepts of Person, Organization, Event, Location, and Art works, which are covered by the top ontology made of the Portuguese equivalents to all concepts in the 4 top layers of the Princeton wordnet and to the 98 Base Concepts suggested by the Global Wordnet Association, and the 164 Core Base Concepts indicated by the EuroWordNet project.

This browsing service offers an access point to the MultiWordnet, browser[1] tailored to the Portuguese wordnet. It offers also the possibility to navigate the Portuguese wordnet diagrammatically by resorting to Visuwords.[2]

3 Outline

This is an outline of the script to be followed.

Step 1 : Presentation of the LX-Center.
Narrative: The text in Section 2.1 above.
Action: Displaying the page at
`http://lxcenter.di.fc.ul.pt`.

Step 2 : Presentation of LX-Conjugator.
Narrative: The text in Section 2.2 above.
Action: Running an example by selecting "see an example" option at the page
`http://lxconjugator.di.fc.ul.pt`.

Step 3 : Presentation of LX-Lemmatizer.
Narrative: The text in Section 2.3 above.
Action: Running an example by selecting "see an example" option at the page
`http://lxlemmatizer.di.fc.ul.pt`;
clicking on one of the inflected forms in the conjugation table generated; clicking on one of the lemmas returned.

Step 4 : Presentation of LX-Inflector.
Narrative: The text in Section 2.4 above.
Action: Running an example by selecting "see an example" option at the page
`http://lxinflector.di.fc.ul.pt`.

Step 5 : Presentation of LX-Suite.
Narrative: The text in Section 2.5 above.
Action: Running an example by selecting "see an example" option at the page
`http://lxsuite.di.fc.ul.pt`.

Step 6 : Presentation of LX-NER.
Narrative: The text in Section 2.6 above.
Action: Running an example by copying one of the examples in the page
`http://lxner.di.fc..ul.pt`
and hitting the "Recognize" button.

Step 7 : Presentation of CINTIL Concordancer.
Narrative: The text in Section 2.7 above.
Action: Running an example by selecting "see an example" option at the page
`http://cintil.ul.pt`.

Step 8 : Presentation of MWN.PT Browser.
Narrative: The text in Section 2.8 above.
Action: Running an example by selecting "see an example" option at the page
`http://mwnpt.di.fc.ul.pt/`.

4 Requirements

This demonstration requires a computer (a laptop we will bring along) and an Internet connection.

[1] `http://multiwordnet.itc.it/`
[2] `http://www.visuwords.com/`

References

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

D. Janus and A. Przepiórkowski. 2006. "POLIQARP 1.0: Some technical aspects of a linguistic search engine for large corpora". *Proceedings PALC 2005*.

T. Brants. 2000. "TnT-A Statistical Part-of-speech Tagger". *Proceedings ANLP2000*.