A Flexible Interface Tool for Manual Word Sense Annotation

Steven Neale, João Silva and António Branco
University of Lisbon, Faculty of Sciences
Department of Informatics
{steven.neale, jsilva, antonio.branco}@di.fc.ul.pt

Abstract

This paper introduces LX-SenseAnnotator, a user-friendly interface tool for manual word sense annotation. The demonstration will show how input texts are loaded by the tool, the options available to the annotator for displaying and browsing texts, and how word senses are displayed and manually assigned. The flexibility of LX-SenseAnnotator, including the support of a variety of languages and the handling of pre-processed texts with different tagsets, will also be addressed.

1 Introduction

Annotated corpora are a cornerstone of Natural Language Processing (NLP), supporting the analysis of large quantities of text across a wide variety of contexts (Leech, 2004) and the development and evaluation of processing tools. There has been an increased interest in "high quality linguistic annotations of corpora" at the semantic level, with word senses in particular being "both elusive and central to many areas of NLP" (Passonneau et al., 2012). Sense annotated corpora are useful, for example, as training data for Word Sense Disambiguation (WSD) tools (Agirre and Soroa, 2009), many of which are based on the Princeton WordNet approach to the lexical semantics of nouns, verbs, adjectives and adverbs (Fellbaum, 1998).

This format is widely used to build sense-annotated corpora in a variety of languages—examples include parallel corpora such as the English/Italian MultiSemCor (Bentivogli and Pianta, 2002) and corpora in languages such as Japanese, Bulgarian, German, Polish and many more (Global WordNet Association, 2013). Despite the need for these corpora to train and test new and developing WSD approaches (Wu et al., 2007), tools for manual sense-annotation are not easy to come by.

Finding any information at all about such tools is difficult, and those that are described are often done so in the context of the specific purposes for which they were developed. For example, the tools used to manually annotate the English MASC Corpus (Passonneau et al., 2012) and Chinese Word Sense Annotated Corpus (Wu et al., 2007) both seem intrinsically tied to those particular corpora. Such examples demonstrate the need for a more open, flexible solution for manual word-sense annotation that is more "readily adaptable to different annotation problems"(O'Donnell, 2008).

As part of our research on WSD in Portuguese, we have encountered the need for a more user-friendly way to manually annotate corpora with information about word senses. Based on these requirements, we present LX-SenseAnnotator, a flexible user-interface tool for browsing texts and annotating them with senses pulled from a Princeton-style WordNet. We are using this tool to produce a gold-standard corpus annotated with senses from our Portuguese WordNet for use in our own WSD tasks, and in this paper describe how its usability and flexibility make it well-suited to similar manual annotation tasks using source texts and WordNet-based lexicons in a variety of different languages.

2 Importing Text

The current implementation of LX-SenseAnnotator is designed for the import of text files that have already been tagged and morphologically analyzed (in particular, POS-tagged and lemmatized) in an

existing pipeline of NLP tools (Branco and Silva, 2006). POS-tagging in particular makes the separation of the input text according to which words are and are not sense-taggable (as described in the next section) very straightforward. It is of course assumed that the preprocessed tags in the input text have been verified and are correct.

A goal for LX-SenseAnnotator is to support the import of source text in a variety of different formats. The code that currently reads and interprets input text is stored in a stand-alone C++ function, making it easy for the tool to be tuned to allow texts pre-processed using different types of tagsets to be imported depending on the goals of particular users. Coupled with the possibility of reading data from different WordNets (any lexical semantic network in any language that adheres to the Princeton WordNet format can be handled), a wide range of languages and different tagsets for each of those languages can be served by LX-SenseAnnotator.

3 Displaying and Browsing Texts

Before being loaded into the text edit panel, each word from the input text is analyzed according to its need for sense-tagging. In accordance with the Princeton format, nouns, verbs, adjectives and adverbs are separated from the rest of the text as potential candidates for sense-tagging and marked in red, so that they can be easily seen against the rest of the text, which is coloured in a dark blue except for those words that have already been annotated, which are marked in green (Figure 1). An additional search is performed on the words identified as being sense-taggable to ensure that they actually exist in the uploaded WordNet, in this case our Portuguese version—those that do not are also excluded from the red, sense-taggable portions of the text.

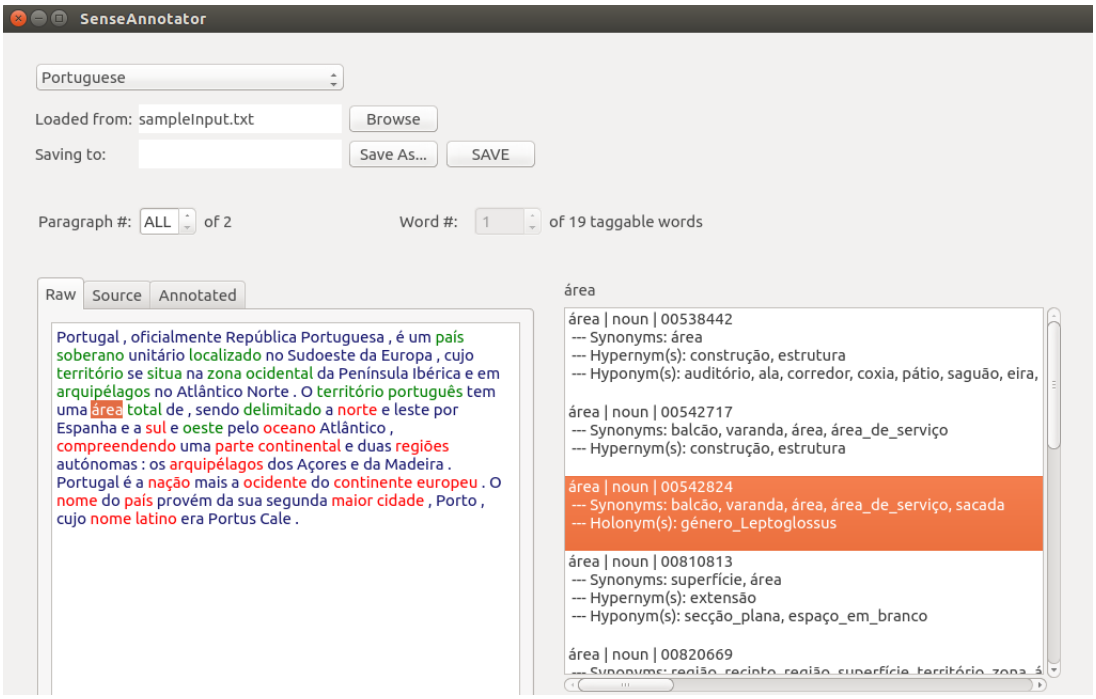

Figure 1: Displaying a list of senses for the word ‘área’ (English ‘area’) using LX-SenseAnnotator.

Once the pre-processed text has been uploaded, the human annotator has a choice of viewing it in the text edit panel in three different views—source text, sense-annotated text and raw text—which can be cycled between using a simple tab widget at the top of the panel. The source text tab displays exactly the original source text (complete with all of the tags present in the imported text). The sense-annotated text tab displays the text with all of the tags from the original source text, and appends the newly added sense tags to the text as the human annotator works—essentially, a continually-updated view of how the output file will be.

The raw text tab displays the text in the cleanest view for reading—all of the tags from the original source text, as well as newly added sense tags, are omitted, allowing for easier reading by the annotator (Figure 2). Our own shallow processing tools used to pre-process the text prior to input include a tokenizer which, among other functions, expands Portuguese contracted forms. For example, 'do' is expanded into two separate tokens, 'de+o' ('of+the' in English). To further aid readability, the current Portuguese LX-SenseAnnotator implementation reverses such tokenizations in the raw text tab.

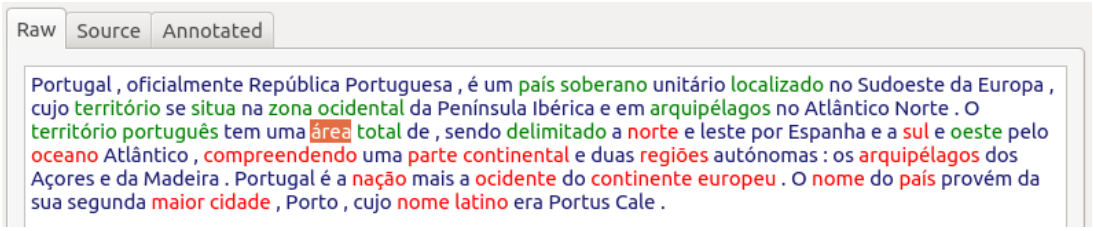

Figure 2: Browsing a text (in Portuguese), with annotated words displayed in green and words yet to be annotated in red.

4 Displaying and Assigning Senses

In any of the three viewing tabs, the annotator can either click on a red, sense-taggable word, or use a scroll-box to browse through the text with currently sense-taggable words. Selecting a word highlights it, and displays the available senses in a separate sense results panel to the right of the text edit panel (Figure 1). The available senses are sourced by querying the presence of the lemma of the selected word in any of the synsets in the index.sense file (limited to the appropriate POS—noun, verb, adjective or adverb). If the word is in a synset, the 8-digit offset of that synset is searched for in the corresponding data file (data.noun, data.verb, etc.) and the results displayed in the right-hand panel as a list of possible options for the selected word.

Information is provided with each sense result to give annotators everything they need to help them decide which sense to assign to a selected word. Using the information from the data (.noun, .verb, etc.) file where the synset was found, each sense result is populated with the main lemma, the POS and the 8-digit offset of the synset. To provide context, this is supplemented with the other words from the synset, which are presented as synonyms, and a selection of the pointers for that synset pulled from the data file (hyper and hyponyms, holonyms, antonyms, entailments, etc.).

After deciding on the most appropriate sense for the selected word, double clicking it in the right-hand sense results panel automatically assigns that sense to the occurence of the word selected in the left-hand text edit panel. In all three viewing tabs, the newly sense-annotated word becomes green, and in the sense-annotated text tab the annotation itself can be seen appended to the selected word. The word is removed from the list of words yet to be annotated, although words which have already been sense-annotated, now displayed in green, can still be selected to allow annotators to assign a different option should they change their mind later.

5 Usability and Flexibility

As mentioned earlier in the paper, LX-SenseAnnotator can read lexical data for any language providing that it adheres to the Princeton WordNet format. The current implementation loads our Portuguese WordNet from a specific directory, from which any number of individual directories containing WordNet-style lexicons can be included and cycled between within the GUI to display senses in different languages. This means that different texts in different languages can be annotated just as easily as each other, simply by loading senses from a different WordNet directory.

Parallel to this is the interpretation of the tags already applied to the input text at the time of import. As the code for handling source text is assigned to a separate, stand-alone C++ function, it is possible to create new classes to interpret tagsets. We plan to further streamline this process by incorporating a simple GUI for annotators to create and edit their own tags, which the program can use to automatically create new versions of the standalone function for interpreting new tagsets in different languages. As the number of supported tagsets grows as a result, so does the flexibility of LX-SenseAnnotator, helping to make manual sense annotation "as flexible for use with common tools and frameworks as possible" (Passonneau et al., 2012).

6 Conclusions

This paper has demonstrated LX-SenseAnnotator, an easy-to-use interface tool for annotating corpora with word-sense data based on a WordNet-style lexicon. There are increasing calls for a "community-wide, collaborative effort to produce open, high quality annotated corpora" that are both "easily accessible and available for use by anyone" (Passonneau et al., 2012). LX-SenseAnnotator can contribute to this effort, offering a flexible, user-friendly platform to build sense-annotated corpora and being particularly suited to creating gold-standard corpora for use in NLP research.

As well as working on improving flexibility in the form of customisable support for tagsets in future updates of LX-SenseAnnotator, there are other elements that are worth taking into consideration. The assumption that the tags in preprocessed input texts are correct has already been mentioned, but specific handling for incorrect tags assigned during preprocessing is important, and providing annotators with the option to highlight and correct such errors using LX-SenseAnnotator would be beneficial. It would also be advantageous if LX-SenseAnnotator were able to handle not just different texts in different languages, but also cases where multiple languages are used within the same text.

We plan to start using the current version of LX-SenseAnnotator to produce a gold-standard sense-annotated corpus in Portuguese for use in our own WSD research, during which process we hope to evaluate the tool from a usability perspective with a team of annotators. We also aim to release LX-SenseAnnotator in the near future as part of the LX-Center (NLX, nd), our existing collection of NLP tools and resources.

Acknowledgements

This work has been undertaken and funded as part of the DP4LT and QTLeap projects.

References

Agirre, E. and A. Soroa (2009). Personalizing PageRank for Word Sense Disambiguation. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '09, Athens, Greece, pp. 33–41. Association for Computational Linguistics.

Bentivogli, L. and E. Pianta (2002). Opportunistic Semantic Tagging. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation*, Las Palmas, Canary Islands, pp. 1401–1406. Association for Computational Linguistics.

Branco, A. and J. R. Silva (2006). A Suite of Shallow Processing Tools for Portuguese: LX-suite. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics: Posters and Demonstrations*, EACL '06, Trento, Italy, pp. 179–182. Association for Computational Linguistics.

Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database*. MIT Press.

Global WordNet Association (2013). WordNet Annotated Corpora. http://globalwordnet.org/wordnet-annotated-corpora/. Accessed: 2015-01-19.

Leech, G. (2004). Adding Linguistic Annotations. In M. Wynne (Ed.), *Developing Linguistic Corpora: A Guide to Good Practice*. AHDS Literature, Languages and Linguistics.

NLX (n.d.). LX-Center: NLX - Natural Language and Speech Group. http://lxcenter.di.fc.ul.pt/home/en/index.html. Accessed: 2015-02-23.

O'Donnell, M. (2008). Demonstration of the UAM CorpusTool for Text and Image Annotation. In *Proceedings of the ACL-08: HLT Demo Session*, Columbus, OH, USA, pp. 13–16. Association for Computational Linguistics.

Passonneau, R. J., C. Baker, C. Fellbaum, and N. Ide (2012). The MASC Word Sense Sentence Corpus. In *Proceedings of the 8th International Conference on Language Resources and Evaluation*, Istanbul, Turkey. European Language Resources Association.

Wu, Y., P. Jin, T. Guo, and S. Yu (2007). Building Chinese Sense Annotated Corpus with the Help of Software Tools. In *Proceedings of the Linguistic Annotation Workshop*, Prague, Czech Republic, pp. 125–131. Association for Computational Linguistics.