# Lexical Semantics Annotation for Enriched Portuguese Corpora

Steven Neale, Rita Valadas Pereira, João Silva(✉), and António Branco

NLX - Natural Language and Speech Group, Department of Informatics,
Faculty of Sciences, University of Lisbon, Lisbon, Portugal
{steven.neale,ana.pereira,jsilva,antonio.branco}@di.fc.ul.pt

**Abstract.** The semantic annotation of corpora has an important role to play in ensuring that sentences occurring in natural language texts are correctly understood based on their intended context. Two examples of lexical semantic units that contribute to this knowledge are word senses – which allow words with multiple meanings to be understood based on the context in which they are used – and named entities – which can be disambiguated and linked back to the specific encyclopedic resources that describe them.

In this paper, we describe the construction of lexical semantically-annotated corpora for Portuguese, annotated with both word senses linked to senses in a Portuguese wordnet and named entities linked to Portuguese Wikipedia entries using DBpedia. The result is a gold-standard lexical semantically-annotated resource that is useful in supporting the training and evaluation of tools for the disambiguation of these lexical units in Portuguese.

**Keywords:** Annotated corpora · Lexical semantics · Word senses · Named entities · Portuguese

## 1 Introduction

The representation of complex semantic linguistic features and information in the annotation of corpora has resulted in the availability of increasingly sophisticated resources for natural language processing (NLP) tasks. Word sense disambiguation (WSD) – annotating words that have more than one meaning in the lexicon with a suitable label that refers to correct sense of that word in the context in which it was used – and named entity disambiguation (NED) – determining the identity of the entities mentioned in a text and then linking them to the corresponding information in existing knowledge bases – are two such semantic tasks whose representation in corpora can be vital for the training and evaluation of NLP tools.

An example of a simple case for WSD would be the English word 'bank' – this might refer either to the financial institution or to the slope of land by the side of a river, and the chosen sense would usually be represented by an entry from a

© Springer International Publishing Switzerland 2016
J. Silva et al. (Eds.): PROPOR 2016, LNAI 9727, pp. 296–305, 2016.
DOI: 10.1007/978-3-319-41552-9_30

stand-alone knowledge base or ontology such as WordNet [6], where nouns, verbs, adjectives and adverbs are stored as sets of synonyms or 'synsets' and linked by their semantic relations. Similarly, a simple case for NED would be a phrase such as 'The President lives in the White House', in which 'President' would be recognized as an entity and annotated with a tag denoting 'person', and 'White House' with a tag denoting 'location' and would usually be represented by a link to a descriptive entry in a database such as DBpedia [8], a large-scale multilingual knowledge base extracted from 111 language editions of Wikipedia.[1] However, most of the available annotated corpora for Portuguese contain other types of semantic information, such as the semantic role of phrases within sentences [2] or the semantic type of named entities [1] – specific information about word senses and named entities senses is not yet represented.

In this paper, we describe the creation of new Portuguese corpora annotated with lexical semantic units, CINTIL-WordSenses and CINTIL-NamedEntities. The new corpora are built upon the CINTIL International Corpus of Portuguese [1] and are annotated with synset identifiers selected from the Portuguese Multi-WordNet [9] (word senses) and with links to appropriate Portuguese Wikipedia entries extracted from DBpedia (named entities). Our contribution is a pair of gold-standard annotated datasets for Portuguese that can support the development of WSD, named entity recognition and classification (NERC) and NED tools, either as dedicated training materials or as a baseline against which Portuguese tools can be evaluated.

We first describe some related work (Sect. 2) before outlining the construction of our corpora (Sect. 3) – focusing on the CINTIL corpus we build upon, its enrichment with word senses, linking disambiguated named entities to it, and the resulting corpora statistics. Next, we describe some of the issues encountered and the future work necessary to develop the corpora (Sect. 4), before offering our concluding remarks (Sect. 5).

## 2 Related Work

While there are various semantically-annotated corpora in existence in other languages, there are relatively few examples in Portuguese. In the case of NED, recent work by Santos et al. [13] describes their efforts to resolve the linking of named entities in Spanish and Portuguese texts to Wikipedia pages, and outlines that their approach is based on extracted dumps of the Spanish and Portuguese versions of Wikipedia and XLEL-21, a dataset developed to support the training and evaluation of cross-language named entity linking systems in twenty-one languages other than English. However, our understanding is that these datasets are based not on spontaneously occurring natural language in context, but rather on lists of singular entities mapped to the links of corresponding Wikipedia pages.

In their work on WSD for Portuguese, Nóbrega and Pardo [11] describe evaluating their work against a manually-annotated subsection of the CSTNews corpus [12], the original version of which contains 140 news texts grouped by

[1] Wikipedia, the free encyclopedia: http://en.wikipedia.org.

topic – 2,088 sentences amounting to 47,240 words. Due to the difficulty and time constraints of their annotation task, they annotated only a small portion of the original corpus – the most frequent 10 % of nouns in each of the 50 clusters that they divided the texts into – resulting in 4,366 annotated words in total. An additional caveat of their work is that their approach relies on translating ambiguous terms to and from English and using the English WordNet for the disambiguation and annotation tasks.

Both of these examples highlight the importance – for both the training and the evaluation of tools – of having a large, dedicated corpora of Portuguese, accurately annotated with word senses and with named entities. For WSD, annotating ambiguous terms with senses from a Portuguese-specific lexicon is important, while for NED annotating named entities as they occur in natural language – as opposed to relying solely on datasets with stand-alone entities mapped to Wikipedia or DBpedia – will be beneficial for the training of NED and also NERC tools. Finally, the example provided by previous work on WSD for Portuguese highlights how useful it would be to have corpora of a much larger size, to better account for the variety found in natural language.

## 3 Constructing the Corpora

This section describes the CINTIL International Corpus of Portuguese, the lexical semantically-annotated corpora that have been built upon it - CINTIL-WordSenses and CINTIL-NamedEntities - and the tools and processes used to annotate these new resources.

### 3.1 The CINTIL International Corpus of Portuguese

The CINTIL International Corpus of Portuguese [1] is a linguistic resource of 1 million tokens containing data from both written sources and transcriptions of spoken Portuguese – the written part, sourced mainly from newspaper articles and short novels, comprises approximately 700,000 tokens. After first being manually annotated with (a) accurate sentence, paragraph and token boundaries; part-of-speech and morphosyntactic categories; inflectional features; and named entities boundaries and semantic types, the corpus was then used to train (b) the sentence chunker; tokenizer; POS tagger; lemmatizer; conjugator; nominal inflector; and named entity recognizer that collectively form the LX-Suite [5]. For the existing version of CINTIL on top of which our lexical semantically-annotated corpora are built, this progressive process of manual annotation, verification and subsequent re-training of the auxiliary annotation tools has ensured that all of the tokens in the existing corpus have been hand annotated and verified, and that the whole of the corpus has been used to train the LX-Suite.

The CINTIL corpus continues to be extended and developed following its original construction. CINTIL-DeepGramBank [4] includes deep grammatical representations, with the output of LX-Gram [3] – a dedicated deep linguistic grammar for Portuguese – having been manually verified by Portuguese

linguististic experts to extend CINTIL with representations of deep linguistic treebanks. This was followed by CINTIL-PropBank [2], whereby syntactic constituency trees from CINTIL-DeepGramBank have been leveraged and enriched with semantic role tags to construct a complete PropBank with both syntactic and semantic levels of annotation.

### 3.2 Enriching CINTIL with Word Senses

CINTIL-WordSenses is the result of the manual assignment of appropriate sense or meaning labels to words that are lexically ambiguous, taking into account the context in which they appear in a given sentence. For example, given a phrase such as 'John deposited Mary's money in the bank', words such as 'deposit' and 'money' give us enough context to determine that the word 'bank' refers to the financial institution, and not to the slope of land at the side of a river.

**The Word Sense Annotation Tool.** CINTIL-WordSenses was annotated using our own graphical user interface tool for assigning synset identifers from WordNet-style lexicons to pre-tagged input texts, LX-SenseAnnotator [10]. The tool was developed to provide a more user-friendly way to annotate texts with word sense information, as part of our research into WSD for Portuguese. The initial version of the tool was developed specifically to provide us with a flexible way to complete the word sense annotation task, after deciding that a gold-standard corpus for use in our WSD tasks was needed.

LX-SenseAnnotator requires that input texts have already been processed using the LX-Suite [5], resulting in tokens already being lemmatized, POS-tagged and morphologically analyzed – the POS-tagging in particular makes it very straightforward to separate the input text according to tokens that can (i.e. open-class words) and cannot be annotated. Loaded text is displayed in a text panel, with potential candidates for annotation marked in red and those words that have already been annotated marked in green. To the right of the interface, a second text panel displays the available senses of a given word to the annotator.

Annotators are able to choose from senses (synsets) extracted directly from the Portuguese MultiWordNet. On highlighting an ambiguous word, the main lemma, POS and the eight-digit synset identifier for each possible sense that could be assigned are displayed by LX-SenseAnnotator, as well as the available synonyms of the synset and a selection of other semantically-related words (hypernyms, hyponyms, holonyms etc.) that offer additional context. As will be described in the next sections, annotators were only able to select from the words and synsets present in the Portuguese MultiWordNet, and as expected not all of the open-class words in the corpus were annotated.

**The Word Sense Annotation Task.** Sections of the CINTIL corpus – divided into short segments of around 50 sentences each – were given to a team of linguistic annotators whose task was to select the correct sense for words in a given sentence, taking into account the discursive context in which the word

is used. The available senses (synsets) from which annotators can choose come directly from the Portuguese MultiWordNet, which currently stands at around 19,700 verified synsets. For example, given a simple phrase such as:

"Produção nacional e qualidade são os objectivos."

A rough translation of which would be:

"National production and quality are the goals."

We might wish to add the following synset identifiers to the open-class words, linking them with their correct sense (synset) in the wordnet:

"Produção (00600686) nacional e qualidade (00765551) são os objectivos (00884793)."

Within the Portuguese MultiWordNet – as with the English WordNet [6] – the open-class words that account for most of the occurrences of semantic ambiguity in natural language are represented by groups of terms or 'synsets', each of which represents a specific meaning or concept for which multiple words may be appropriate. Each synset is linked to others by semantic relations – such as synonymy, hyponymy, hypernymy etc. – and is labeled using an eight-digit number that acts as a unique identifier and can be used to represent the meaning of a word that a human annotator might assign it to (see the above example).

### 3.3 Linking Disambiguated Named Entities in CINTIL

CINTIL-NamedEntities has been created by manually linking pre-recognized named entities to appropriate Portuguese Wikipedia URIs via their entries in DBpedia. For example, the word 'Portugal' on being recognized as a named entity of semantic type 'LOC' (location) would be annotated with a link to the DBpedia entry for the country of Portugal in its geographical sense. Prior to the disambiguation and annotation task, the corpus was automatically tagged using LX-NER, a hybrid rule-based and statistical NERC tool for Portuguese [7]. For the task this paper describes we have focused particularly on the annotation of entities in the names category, which includes types for locations (LOC), organizations (ORG), persons (PER), events (EVT), works (WRK) and an additional miscellaneous (MSC) category.

**Brat.** The disambiguation and annotation task was completed using version 1.3 of the brat annotation tool [14],[2] a web-based annotation system with several features that make it a good choice for our task. b Brat runs directly in the browser – meaning that there is no installation of specific software required for the human annotator – and has an intuitive user interface that allows annotators to define the span of a named entity simply by dragging a selection over text and then selecting the tag to be assigned from a pre-defined drop-down list. Recent

[2] Available from: http://brat.nlplab.org.

versions of brat also have built-in support for normalization [15], which allow annotations to be associated with external resources. In the context of our task, this allows annotators to associate each disambiguated named entity with its corresponding Portuguese Wikipedia page entry in DBpedia.

**Disambiguating and Annotating Named Entities.** The task of disambiguating and annotating named entities was performed on a version of the corpus that had already been pre-processed using our in-house NERC tool, LX-NER. As well as helping the linguistic annotators to identify the ambiguous named entities within the text, having the entities recognized and classified prior to the start of the task allowed for entities to be fed to a pre-annotation normalization script that could associate them with the Portuguese Wikipedia pages in DBpedia's database [8]. Therefore, the linguistic annotators' task was not concerned with the recognition or coarse classification of the entities themselves, but instead on first verifying that the normalization was correct, and then on choosing the appropriate DBpedia entry for the named entity in question.

To complete the disambiguation task, annotators had to take into account both the tag assigned to the entity by LX-NER and the discursive context in which the entity occurs. Often, the same word could be associated with different Wikipedia entries, depending on the context in which it is used – for example, the word 'Portugal' with the LOC tag should link to the Wikipedia page for the country of Portugal in the geographical sense, while the same word with the ORG tag would be better linked to the Wikipedia page for the Portuguese government.

### 3.4 Semantically-Annotated Corpora Statistics

The final result of the annotation of the corpora is a language resource for Portuguese of 23,825 word sense-annotated and 30,493 named-entity annotated sentences[3] (see Table 1). For word senses, from a total of 508,717 tokens there are 193,443 open-class (potentially ambiguous) words, of which 45,502 (23.52 %) were manually disambiguated and annotated with synset identifiers from the Portuguese MultiWordNet. For named entities, from a total of 684,467 tokens there were 26,371 entities recognized by LX-NER during pre-processing, of which 16,120 (61.13 %) were manually disambiguated and annotated with links to their appropriate DBpedia entries. Both corpora – CINTIL-WordSenses and CINTIL-NamedEntities – are available via META-SHARE.[4]

To create these first versions of both CINTIL-WordSenses and CINTIL-NamedEntities, the sections into which the corpus was divided for each task were each assigned to a single annotator. In the future, we plan to annotate each section of the corpus again with a different annotator (for inter-annotator agreement)

[3] In this first version of the word sense annotation task, fewer sentences were distributed to annotators than in the named entity disambiguation task. These gaps will be addressed in future versions of the word sense annotation task.

[4] Accessible from: http://www.meta-share.eu/.

**Table 1.** Composition of the lexical semantically-annotated corpora for Portuguese, CINTIL-WordSenses and CINTIL-NamedEntities.

| | WordSenses | NamedEntities |
|---|---|---|
| Sentences | 23,825 | 30,492 |
| Tokens | 508,717 | 684,467 |
| *Senses:* | | |
| Potentially Ambiguous | 193,443 | — |
| Manually Annotated | 45,502 (23.52 %) | — |
| *Entities:* | | |
| Recognized (LX-NER) | — | 26,371 |
| Manually Annotated | — | 16,120 (61.13 %) |

and to introduce a process of adjudication, such that the reliability of both resources can be properly quantified.

## 4 Future Work and Development of the Corpora

Following the creation of CINTIL-WordSenses and CINTIL-NamedEntities, there are a number of possible improvements – affecting both the word sense and named entity annotation tasks – that can be considered going forward in order to continue the development of the corpora.

### 4.1 Gaps in the Resources Essential to Disambiguation

For both annotation tasks, some gaps are apparent in the lexical resources used to disambiguate word senses and to normalize named entities. In the case of CINTIL-WordSenses, we encountered some examples where the task would be improved by extending the Portuguese MultiWordNet used by the human annotators. The usual case is of words not having an entry present in the wordnet, but there were other cases in which while synsets may currently exist that contain the word in question, these synsets represent alternative meanings to the one most appropriate for the given context. For example, there is a well-formed synset for the word 'corredor' ('athlete' or 'runner' in English) in the wordnet, but currently no synset representing the alternative meaning of the word 'corredor' ('hallway' in English). The homonymous Portuguese words 'corredor' ('hallway') and 'corredor' ('runner') – with one of these senses being well-represented and another not available at present – highlight the positive impact that growing the Portuguese wordnet will have on future versions of CINTIL-WordSenses.

Similarly, in the case of CINTIL-NamedEntities, it was not possible to provide normalized links for all of the named entities discovered by LX-NER tool, as many of them had no suitable entries in DBpedia – in these cases, annotators marked the entities in question as 'Not found' during the annotation.

There were also cases in which the entity in question is more specific than the available DBpedia entry (or vice-versa), cases that annotators currently account for by adopting a 'generalized' annotation approach. For example, given a more abstract or ambiguous entity (to use an English example, 'President of the United States'), the annotator will link the entity to the most specific appropriate entry in DBpedia (i.e. 'Barack Obama') if it can be identified by context, or otherwise to the DBpedia entry of the more general concept (i.e. 'President of the United States'). This generalized annotation approach is also used when a DBpedia entry for a very specific named entity cannot be found. For example, many of the journalistic texts in the original corpus come from one of the numerous supplementary sections of the Portuguese newspaper 'Diário de Notícias', such as 'DN - Internacional'. These specific sections of the newspaper often appear as named entities in the corpus, but because DBpedia does not contain entries for the individual sections the entities must instead be linked to the less-specific page of the newspaper, 'DN - Diário de Notícias'.

### 4.2 Problems Inherited from Prior Tagging and Annotation

The prior annotation of the original corpus has also introduced some considerations for the development of both the word sense and named entity annotation tasks, usually as a result of discrepancies in the existing POS tagging or incorrectly recognized named entities. For example, previous annotation schemes allowed for nominal multi-word expressions (MWEs) to be compositionally annotated – for example, 'sala de estar' ('living room' in English), adverbial MWEs such as 'em princípio' ('in principle' in English), or prepositional MWEs such as 'por volta de' ('around' in English). The initial version of the word sense annotation tool used to annotate CINTIL-WordSenses reads input texts as individual tokens, and so currently only allows for nominal tokens (i.e. 'sala', 'princípio', 'volta') to be annotated with synset identifiers. Given that the disambiguation of these expressions relies on the role that each word plays in the compositional sense formed with the other words in the expression, future versions of the task – and the resulting corpus – could be greatly improved by updating the word sense annotation tool to allow for compositionally annotated expressions to be understood, and to link to compositional terms representing these expressions (as opposed to nominal tokens only) in the Portuguese MultiWordNet.

Considering also CINTIL-NamedEntities, a number of entities have been incorrectly tagged prior to the task by LX-NER – consider the sentence:

> "Estas declarações não escondem as divergências entre Paris (LOC) e Washington (LOC)"

Which could be roughly translated as:

> "These statements make clear the differences of opinion between Paris (LOC) and Washington (LOC)"

In this example, 'Paris' and 'Washington' would both have been better tagged as ORG than LOC – however, we have chosen not to disambiguate such occurrences

at this stage. Instead, we intend to improve CINTIL-NamedEntities in future versions by first correcting the erroneous output from the LX-NER tool, and then by retraining our tools on the corrected output.

## 5 Conclusions

We have described the creation of two new lexical semantically-annotated corpora for Portuguese – CINTIL-WordSenses and CINTIL-NamedEntities – manually disambiguated and annotated with word senses and with named entities linked to appropriate Portuguese Wikipedia entries using DBpedia, respectively. Our work contributes gold-standard lexical semantically-annotated resources that can be used in the development of WSD, NERC and NED tools for Portuguese, either as dedicated data for the training of such tools or as a baseline against which their output can be evaluated.

We are now continuing with the development of the corpora, for which our immediate attention is on implementing inter-annotator agreement and adjudication steps in the next version to properly quantify the accuracy and reliability of the resource. Following this, we will then focus on addressing some of the issues encountered in Sect. 4 – manually correcting some of the existing errors inherited from pre-processing (concerning MWEs and recognized entities in particular) will allow us to retrain our existing tools on the corrected output, and to continue developing the corpora in the knowledge that it becomes a more reliable resource with each version.

**Acknowledgements.** The results reported in this paper were partially supported by the Portuguese Government's P2020 program under the grant 08/SI/2015/3279: ASSET-Intelligent Assistance for Everyone Everywhere, by FCT-Fundao para a Cincioa e Tecnologia under the grant PTDC/EEI-SII/1940/2012: DP4LT-Deep Language Processing for Language Technology, and by the ECs FP7 program under the grant number 610516: QTLeap-Quality Translation by Deep Language Engineering Approaches.

## References

1. Barreto, F., Branco, A., Ferreira, E., Mendes, A., Nascimento, M.F.B., Nunes, F., Silva, J.: Open resources and tools for the shallow processing of Portuguese: the TagShare Project. In: Proceedings of the 5th International Conference on Language Resources and Evaluation, LREC 2006, pp. 1438–1443 (2006)
2. Branco, A., Carvalheiro, C., Pereira, S., Silveira, S., Silva, J., Castro, S., Graça, J.: A PropBank for Portuguese: the CINTIL-PropBank. In: Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC 2012). European Language Resources Association (ELRA), Istanbul (2012)
3. Costa, F., Branco, A.: LXGram: a deep linguistic processing grammar for Portuguese. In: Pardo, T.A.S., Branco, A., Klautau, A., Vieira, R., de Lima, V.L.S. (eds.) PROPOR 2010. LNCS, vol. 6001, pp. 86–89. Springer, Heidelberg (2010)

4. Branco, A., Costa, F., Silva, J., Silveira, S., Castro, S., Avelãs, M., Pinto, C., Graça, J.: Developing a deep linguistic databank supporting a collection of treebanks: the CINTIL deepgrambank. In: Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010). European Language Resources Association (ELRA), Valletta (2010)
5. Branco, A., Silva, J.: 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 2006, pp. 179–182. Association for Computational Linguistics, Trento (2006)
6. Fellbaum, C.: WordNet: An Electronic Lexical Database. MIT Press, Cambridge (1998)
7. Ferreira, E., Balsa, J., Branco, A.: Combining rule-based and statistical methods for named entity recognition in Portuguese. In: V Workshop em Tecnologia da Informação e da Linguagem Humana, TIL 2007, pp. 1615–1624 (2007)
8. Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P., Hellmann, S., Morsey, M., van Kleef, P., Auer, S., Bizer, C.: DBpedia - a large-scale, multilingual knowledge base extracted from wikipedia. Semant. Web J. **6**(2), 167–195 (2012)
9. MultiWordNet: The MultiWordNet project. http://multiwordnet.fbk.eu/english/home.php (nd). Accessed 13 Jan 2015
10. Neale, S., Silva, J., Branco, A.: A flexible interface tool for manual word sense annotation. In: Proceedings of the 11th Joint ACL-ISO Workshop on Interoperable Semantic Annotation, ISA-11, pp. 67–71. Association for Computational Linguistics, London (2015)
11. Nóbrega, F.A.A., Pardo, T.A.S.: General purpose word sense disambiguation methods for nouns in Portuguese. In: Baptista, J., Mamede, N., Candeias, S., Paraboni, I., Pardo, T.A.S., Volpe Nunes, M.G. (eds.) PROPOR 2014. LNCS, vol. 8775, pp. 94–101. Springer, Heidelberg (2014)
12. Cardoso, P.C.F., Maziero, E.G., Jorge, M.L.R.C., Seno, E.M.R., di Felippo, A., Rino, L.H.M., das Nunes, M.G.V., Pardo, T.A.S.: CSTNews - a discourse-annotated corpus forsingle and multi-document summarization of news texts in Brazilian Portuguese. In: Proceedings of the Third Annual RST and Text Studies Workshop, pp. 88–105 (2011)
13. Santos, J., Anastacio, I., Martins, B.: Named entity disambiguation over texts written in the Portuguese or Spanish languages. Lat. Am. Trans. IEEE (Rev. IEEE Am. Lat.) **13**(3), 856–862 (2015)
14. Stenetorp, P., Pyysalo, S., Topić, G., Ananiadou, S., Aizawa, A.: Normalisation with the BRAT rapid annotation tool. In: Proceedings of the 5th International Symposium on Semantic Mining in Biomedicine, Zürich, Switzerland (2012)
15. Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., Tsujii, J.: Brat: a web-based tool for nlp-assisted text annotation. In: Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics, pp. 102–107. Association for Computational Linguistics, Avignon (2012)