Small in Size, Big in Precision: A Case for Using Language-Specific Lexical Resources for Word Sense Disambiguation

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

Abstract

Linked open data (LOD) presents an ideal platform for connecting the multilingual lexical resources used in natural language processing (NLP) tasks, but the use of machine translation to fill in gaps in lexical coverage for resource-poor languages means that large amounts of data are potentially unverified. For graph-based word sense disambiguation (WSD), one approach has been to first translate terms into English in order to disambiguate using richer, fuller lexical knowledge bases (LKBs) such as WordNet.

In this paper, we show that this approach actually creates more ambiguity and is far less accurate than using language-specific resources, which, regardless of their smaller size, can provide results comparable in accuracy to the state-of-the-art reported for graph-based WSD *in* English. For LOD, this demonstrates the importance of continuing to grow and extend language-specific resources in order to continually verify and reintegrate them as accurate resources.

1 Introduction

In the context of natural language processing (NLP), word sense disambiguation (WSD) refers to the computational problem of determining the 'sense' or meaning of a word when used in a particular context (Agirre and Edmonds, 2006). To use a classic example, the word 'bank' could be interpreted in the sense of the financial institution or as the slope of land at the side of a river, depending on the context in which it is used. Target words are disambiguated based on their context (determined based on the words surrounding them), and the potential senses that they could relate to (Nóbrega and Pardo, 2014).

Linked Open Data (LOD) – the implementation of best practices ensuring that not just documents but the data within them are structured and interconnected on the web – is particularly useful in tying together the resources used for knowledge-based WSD, which leverages existing collections and indexes of potential senses to choose the most appropriate for a given target word (Agirre and Edmonds, 2006). WSD research has tended to derive knowledge bases from stand-alone dictionaries and ontologies such as WordNet, where nouns, verbs, adjectives and adverbs are stored as 'synsets' and linked by their semantic relations (Fellbaum, 1998). Recent projects such as BabelNet (Navigli and Ponzetto, 2012) are now focusing on integrating these resources with encyclopedic information and making the connected data available as LOD.

Our work focuses on Portuguese, for which specific work on WSD – particularly involving Portuguese knowledge resources – is still limited, and usually either focused on particular domain areas and applications or achieved by translating terms to English in order to disambiguate using English knowledge sources (Nóbrega and Pardo, 2014). While there are similarities between Portuguese and other languages for which more substantial lexical resources and WSD research are already available – French and Spanish, for example – there are still enough differences to motivate specific research in Portuguese. The sheer number of 'false friends' – similar words with very different meanings – between Portuguese and Spanish (Director General of Translation, 2006) demonstrates the necessity of having Portuguese-specific resources available for lexically-motivated tasks such as WSD.

This paper describes a comparison between two approaches to performing graph-based WSD

Proceedings of the Second Workshop on Natural Language Processing and Linked Open Data, pages 6–15,
Hissar, Bulgaria, 11 September 2015.

in Portuguese; 1) using the smaller, language-specific Portuguese MultiWordNet (MultiWordNet, nd) as the underlying lexical knowledge base (LKB) for the WSD, and 2) translating open-class words in the input text from Portuguese to English in order to run WSD using the much larger English WordNet as the underlying LKB. The contributions from our results are twofold:

- Performing graph-based WSD using a smaller, language-specific LKB (Portuguese MultiWordNet) provides better results than translating terms to English in order to run WSD using the much larger English WordNet.
- The results obtained when performing graph-based WSD using a small, language-specific LKB (such as the Portuguese MultiWordNet) are comparably accurate with state-of-the-art results previously reported for graph-based WSD in English using WordNet.

These contributions suggest that for LOD, relying on machine translation to fill in the lexical gaps between resource-rich and research-poor languages (as with BabelNet) must only be a stopgap measure, and that work to grow and extend local, language-specific lexical resources such as WordNets should continue so that verified, accurate data can be properly linked and reintegrated with existing LOD later for use in NLP tasks such as WSD.

We first explore some related work (Section 2), before describing an implementation of graph-based WSD for Portuguese (Section 3). Next, we present our evaluation of the two approaches to WSD in Portuguese, using a gold-standard, human-annotated corpus for comparison (Section 4). Finally, we discuss the possible ramifications of our findings in the context of LOD (Section 5), before presenting our conclusions (Section 6).

2 Related Work

2.1 Knowledge and graph-based WSD

While WSD has traditionally delivered its best results using supervised and unsupervised machine learning methods, domain-specific knowledge-based WSD can now perform as well or better than a more generic, supervised machine learning-based WSD approach (Agirre et al., 2009). For example, in the medical domain good results have been obtained in WSD tasks by creating an LKB from the Unified Medical Language System (UMLS) Metathesarurus, a collection of more than one million biomedical concepts and five million concept names (Stevenson et al., 2011; Preiss and Stevenson, 2013).

Progress in knowledge-based WSD has largely been driven by the development of graph-based disambiguation methods, as pioneered by a number of researchers (Navigli and Velardi, 2005; Mihalcea, 2005; Sinha and Mihalcea, 2007; Navigli and Lapata, 2007; Agirre and Soroa, 2008). Graph-based methods allow LKBs such as WordNets to be represented as weighted graphs, where word senses correspond to nodes and the relationships or dependencies between pairs of senses correspond to the edges between nodes. The strength of the edge between two nodes, corresponding to the relationship or dependency between two synsets, can then be calculated using semantic similarity measures such as the Lesk algorithm (Lesk, 1986).

For WSD tasks, graph-based representations of LKBs can then be used to choose the most likely sense of a word in a given context, based on the dependencies between nodes in the graph (Agirre and Soroa, 2009). Algorithms such as PageRank (Brin and Page, 1998) allow for the weights and probabilities of directed links between target words and words in their local context to be spread over the entirety of the graph (Agirre and Soroa, 2009). Nodes (senses) 'recommend' each other based on their own importance – with the importance of any given node being higher or lower depending on the importance of other nodes which recommend it – and then follow a 'random walk' over the rest of the graph based on the importance of the nodes to whose edges they are attached (Mihalcea, 2005; Agirre and Soroa, 2009).

At the end of this random walk, the probability of a random walk from the target word's node ending on any other node in the graph has been calculated, thus allowing the most appropriate sense of the target word to be detemined. By utilizing the full extent of the graph-based representation of the LKB in this way, the performance of WSD in general (non-specific) domains has been shown to improve, becoming almost as efficient as supervised learning-based methods in some tasks (Agirre et al., 2014).

2.2 Linked Open Data and aligned LKBs

In parallel to the growing use and adaptation of different types of LKBs in knowledge and graph-based WSD, the lexical resources on which these LKBs and WSD methods depend are becoming increasingly linked, interconnected and accessible. Projects like MultiWordNet (MultiWordNet, nd) and EuroWordNet (Vossen, 2004) are built around the idea of aligning and mapping the identifier codes of WordNet-style synsets to each other, and in many languages. For knowledge-based WSD, this connectivity makes multilingual and language-specific WSD tasks and workflows much simpler to construct.

Recent LOD projects such as DBpedia (Lehmann et al., 2012) and BabelNet (Navigli and Ponzetto, 2012) are now collecting data from encyclopedic sources such as Wikipedia to create large-scale, structured multilingual knowledge bases. BabelNet, in particular, integrates both lexical and encyclopedic resources – chiefly WordNet and Wikipedia – to create a 'wide-coverage, multilingual semantic network' of not only information and concepts but also the semantic relationships between them (Navigli and Ponzetto, 2012). Like DBpedia – which connects the extracted knowledge from 111 different language editions of Wikipedia (Lehmann et al., 2012) – BabelNet is also multilingual, using machine translation techniques to fill in the lexical gaps in resource-poor languages (Navigli and Ponzetto, 2012).

2.3 Current state of WSD in Portuguese

Portuguese-specific WSD has also followed the knowledge-based trend. Early work focused on the automatic generation of disambiguation rules based on representations of meaning in pre-annotated corpora (Specia et al., 2005), before exploring hybrid approaches that leverage the relationships between different knowledge sources to support such rules (Specia, 2006; Specia et al., 2007). More recent work has focused on graph-based methods, leveraging WordNets as LKBs (Nóbrega and Pardo, 2014). However, this work assumes that translating Portuguese terms into English and then querying the English WordNet is sufficient for representing most of the senses found in Portuguese texts.

Spanish, which shares a degree of similarity with Portuguese, has been more widely explored in the context of WSD. Agirre and Soroa (2009) evaluated their graph-based WSD algorithm using the Spanish WordNet of approximately 67,000 senses (Atserias et al., 2004) as their LKB. They obtained promising results that approach those reported using the supervised 'most frequent sense' (MFS) baseline system for the SemEval-2007 Task 09 dataset (Màrquez et al., 2007). More recently, graph-based WSD performed over Spanish Babelnet senses as the LKB was shown to improve over the MFS baseline in the Multilingual Word Sense Disambiguation task at SemEval-2013 (Navigli et al., 2013).

These results are encouraging for the case of Portuguese, demonstrating that knowledge-based WSD produces good results using LKBs specific to similar languages. For Portuguese, it would thus seem more appropriate to grow Portuguese-specific lexical resources and to link them with existing resources in other languages as LOD, than to rely either on translating the input words to be disambiguated, as in (Nóbrega and Pardo, 2014), or on filling the gaps in one language by translating from the fuller lexical resources of other languages, as in BabelNet (Navigli and Ponzetto, 2012).

3 Implementing Graph-Based WSD for Portuguese

For the evaluations described in this paper, we use UKB, a collection of tools and algorithms (Agirre and Soroa, 2009; Agirre et al., 2014) for performing graph-based WSD over a pre-existing knowledge base. We use UKB for two reasons:

- UKB includes tools for automatically creating graph-based representations of LKBs in WordNet-style formats.
- The algorithm used by UKB for performing WSD over the graph itself has been consistently shown to produce results in line with or above the state-of-the-art (Agirre and Soroa, 2009; Agirre et al., 2014).

For the purpose of our work, we are thus able to perform highly-efficient disambiguation over an accurate graph-based representation of our chosen LKBs, meaning that any differences in results can be confidently attributed to the quality of either the input texts that are being disambiguated or to the LKBs themselves.

UKB first accepts input texts in a 'context' format, where each sentence in a text is treated as an individual context containing the target word and all other open-class words (nouns, verbs, adjectives and adverbs) from the original sentence. This context file can be easily extracted and arranged from input texts pre-tagged with lemmas and part-of-speech (PoS) tags, which we produce using the LX-Suite (Branco and Silva, 2006), a collection of shallow processing tools for Portuguese.

UKB then performs WSD for each sentence in the context file, using a PageRank-based (Brin and Page, 1998) random walk to return the probability of each node (synset) in a given graph being semantically related to a target word, and returning the appropriate synset identifier for the most likely node. It is this use of the words surrounding a target word in the context file – which are also included as nodes in the graph and whose relevance thus affects the final decision on which sense to assign – that separates UKB from similar algorithms and consistently delivers state-of-the-art results (Agirre and Soroa, 2009; Agirre et al., 2014).

The graphs used for the evaluation in this paper were created, using the tools supplied with UKB, from two different source LKBs – the Portuguese MultiWordNet (MultiWordNet, nd) and version 3.0 of the Princeton English WordNet (Fellbaum, 1998). These LKBs are described in more detail in the following section.

4 Evaluation

This section describes our comparison of the assignment of word senses by a human annotator with the output of two options for performing graph-based WSD in Portuguese:

- UKB-based WSD over the Portuguese MultiWordNet.
- UKB-based WSD over the English WordNet (using terms automatically translated from Portuguese to English)

For UKB-based WSD over the Portuguese MultiWordNet, we create the required dictionary files and corresponding graph from approximately 19,700 verified synsets. Because the synset identifiers are mapped to the corresponding synsets in the English WordNet, we are able to make use of the semantic relations in the English WordNet when building the graph – although the dictionary used is small at 19,700, the fuller representation of semantic relations for English ensures that the computed similarity between Portuguese dictionary items is more reliable. Semantic relations between glosses in the English WordNet are also used when building the graph, which our own experimentation and previous reporting of results using UKB (Agirre and Soroa, 2009; Agirre et al., 2014) have both shown to result in more accurate WSD.

For UKB-based WSD over the English WordNet, we follow the model used by Nóbrega and Pardo (2014) of translating ambiguous terms into English and then disambiguating them using the English WordNet. In practice, this involves translating the context file from Portuguese to English after the input text is preprocessed and tagged using the shallow processing tools, so as to have translated not just the target words but also the surrounding open class words in each sentence. The translated context file is then disambiguated by UKB using a dictionary file and corresponding graph created from the English WordNet, comprising approximately 117,000 synsets.

We have not been able to use the WordReference API (WordReference.com, nd) that Nóbrega and Pardo (2014) used for translating from Portuguese to English, for which user access is no longer being granted. Instead, we have created our own tool for translating terms from the context file word-by-word using BabelNet. Each individual Portuguese word to be translated is given together with its part of speech to BabelNet, which returns the most appropriate 'BabelSynset' for that word.

BabelSynsets are constructed from linked information from a variety of sources in different languages (including Wikipedia (Wikipedia, nd), WordNet (Fellbaum, 1998), Wiktionary (Wiktionary, nd), Wikidata (Wikidata, nd), OmegaWiki (OmegaWiki, nd) and various others) with gaps in resource-poor languages filled using machine translation. Every BabelSynset contains a list of translations of its main sense in different languages, and each of the possible translations for the word in each language has a weighting or probability attached to it. From this, we choose the best weighted translation from the English options and use this as the translation for the original Portuguese word in the context file.

CINTIL	UKB + PT	UKB + EN Translations
Manually disambiguated	45,502	45,502
Automatically disambiguated	59,190	112,678
Manually *and* automatically disamb.	45,386	41,441
Same sense assigned	29,540	12,563
Precision	65.09	30.32
Recall	64.92	27.61
F1	65.00	28.90

Table 1: Comparison of the performance of UKB-based WSD over the Portuguese MultiWordNet and by translating terms to English to be run over the English WordNet.

4.1 Gold-Standard Test Corpus

The CINTIL International Corpus of Portuguese (Barreto et al., 2006) was chosen as the gold-standard for our evaluation. It comprises approximately 1 million tokens manually annotated with lemmas, part-of-speech, inflection, and named entities, which are compatible with the input and output formats of the tools in the LX-Suite. The corpus contains data from both written sources and transcriptions of spoken Portuguese – we have used the data from the written part, sourced mainly from newspaper articles and short novels and comprising approximately 700,000 tokens, of which 193,443 are open class words.

Word senses were manually chosen and assigned to open-class words by a team of human annotators using the LX-SenseAnnotator tool (Neale et al., 2015), a graphical user interface for assigning senses from WordNet-style lexicons to pre-tagged input texts. The lexicon from which annotators were able to choose senses was the same Portuguese MultiWordNet (approximately 19,700 verified synsets) used in the evaluation. Because annotators were only able to select from the words and synets present in the Portuguese MultiWordNet, not all of the open-class words in the corpus were able to be annotated.

4.2 Performance for Portuguese

Running the UKB algorithm over the manually disambiuated CINTIL corpus, we can see how well the two approaches – disambiguation using the smaller Portuguese MultiWordNet or translating words to English and then disambiguating using the much larger English WordNet – perform when compared with disambiguation by a human annotator. As described earlier in section 4, the mapping of synset identifiers between the Portuguese and English WordNets allows the same graph to be used in both approaches (built based on the semantic relations between English synsets coupled with the semantic relations between English glosses) - it is the sizes of the dictionary files that link words to synsets in the graph that greatly differ.

Table 1 shows that 45,502 of the 193,443 open class words have been manually disambiguated. When running UKB over the dictionary files and graph built from the Portuguese MultiWordNet, 45,386 of the manually disambiguated words are also automatically disambiguated, from a total of 59,190 tagged by the algorithm. Note that although annotators may have chosen *not* to disambiguate certain words if they felt that the senses presented to them by the Portuguese MultiWordNet did not convey the required meaning, the UKB algorithm will always assign something from the options available to it, choosing the most probable sense from those provided.

This explains the greater number of senses automatically disambiguated than manually disambiguated, but without manual disambiguation we have no measure of whether the additional automatic disambiguation was correct or not. Thus, we here define recall as the number of words with the same sense assigned by UKB *and* the human annotator, divided by the number of words manually disambiguated (45,502). The UKB-based WSD was able to assign the same sense to the word as was chosen by the annotator for 29,540 of the 45,386 words for which the same sense was assigned manually and automatically, giving a precision of 65.09% and recall of 64.92%.

When running UKB by automatically translating ambiguous Portuguese terms into English and then running them over the dictionary files and

graph built from the English WordNet, performance is greatly affected. Despite vastly more words being tagged with an assigned sense by the algorithm – 112,678 – a lower number of the words that were manually disambiguated end up being tagged as well – 41,441. The UKB-based WSD was able to assign the same sense to the word as was chosen by the annotator for just 12,563 of these words, giving a precision of 30.32% and recall of 27.61%

Corpus	LKB	F1
Senseval-2	WN3.0	70.3
Senseval-3	WN3.0	65.3
Semeval-07 (FG)	WN3.0	56.0
Semeval-07 (CG)	WN3.0	83.6
CINTIL	PT MWN	65.0

Table 2: Comparison of UKB-based WSD over the Portuguese MultiWordNet with previously reported state-of-the-art results (for nouns).

Table 2 compares the performance of UKB over the Portuguese MultiWordNet with the results obtained by Agirre et al. (2014), who most recently reported on the performance of UKB as F1 over four different datasets – the Senseval-2 (Palmer et al., 2001), Senseval-3 (Snyder and Palmer, 2004), Semeval-2007 fine-grained (Palmer et al., 2001; Snyder and Palmer, 2004; Pradhan et al., 2007) and Semeval-2007 coarse-graned (Navigli et al., 2007) English all-words tasks. Although the results they present cover various disambiguation options within UKB, we focus here on the results they obtained using the *ppr_w2w* UKB method (as we have). We also assume that they continue using version 3.0 of the English WordNet (complete with information on the semantic relationships between glosses) as their underlying LKB, as they have reported in previous evaluations (Agirre and Soroa, 2009). This combination of UKB option and underlying LKB is comparable with our own evaluation of UKB over the Portuguese MultiWordNet.

The 19,700 verified synsets from the Portuguese MultiWordNet version used in our evaluation are constructed from 16,728 words, of which only 45 are not nouns. While Agirre et al. separate their results by nouns, verbs, adjectives and adverbs and also offer an overall score (2014), to compare our results with their overall score would cast our own in a very favourable (and very unfair) light. Therefore, Table 2 only compares our results against those previously reported for nouns by Agirre et al. (2014).

5 Discussion

The results presented in the previous section highlight two important points:

- That performing WSD over a smaller, language specific LKB (such as the Portuguese MultiWordNet) is *more accurate* (tagged with the sames senses as were manually assigned by a human annotator) than translating ambiguous terms into English to perform WSD over larger LKBs (such as WordNet).
- That performing WSD over a smaller, language specific LKB (such as the Portuguese MultiWordNet) produces results with *comparable accuracy* to state-of-the-art results reported for (UKB-based) WSD over the much larger English WordNet.

Table 1 shows that the results obtained by running UKB over the dictionary and graph files created from the Portuguese MultiWordNet are far higher than those obtained by first translating the target and surrounding words in the context file into English, and then running UKB over the English WordNet. This is despite the fact that the Portuguese MultiWordNet is considerably smaller, at around 19,700 verified synsets, than the English WordNet, at a reported 117,000 synsets.

Nóbrega and Pardo themselves (2014), whose approach of translating ambiguous words to English in order to perform WSD using the English WordNet we have compared with our own language-specific results, describe some of the problems that translating terms to and from English can introduce. They observe that some very specific terms or concepts in Portuguese may not have a direct translation in English at all, while conversely there may be generic terms or concepts in Portuguese that have much more specific categories in English (Nóbrega and Pardo, 2014). While their coverage may be less due to their smaller size, language-specific LKBs limit such problems, with the terminology that *is* accounted for being specific to the language in question.

A glance at the original and translated context files used in our comparison shows that in many

cases incorrect translations before the WSD has even been performed have led to the difference in results using the two approaches. For example, a line from a news article in the CINTIL corpus reads:

> "O secretário de Imprensa da Casa Branca, Mike McCurry, disse que qualquer agressão iraquiana seria 'uma questão de grave preocupação"'

An accurate translation of which would be:

> "The White House press secretary, Mike McCurry, said that any Iraqi offensive would be 'a question of serious concern"'

From this sentence, extracting the open-class words in Portuguese produces the following line for the context file (formatted as lemma#pos#wordid):

secretário#n#w1 imprensa#n#w2
dizer#v#w3 agressão#n#w4
iraquiano#a#w5 ser#v#w6
questão#n#w7 grave#a#w8
preocupação#n#w9

Upon translating each of these words to English, we are left with the following line in our translated context file, to be passed to UKB and each term disambiguated using the dictionary and graph files from the English WordNet.

secretary#n#w1 printing_press#n#w2
tell#v#w3 aggression#n#w4 iraqi#a#w5
being#v#w6 question#n#w7
grave#a#w8 concern#n#w9

As well as a number of words which could have been translated slightly better – 'say' would have been better than 'tell' for word three, 'offensive' better than 'aggression' for word four and 'serious' better than 'grave' for word eight – there is a more obvious problem with the translation of word two. The Portuguese word 'imprensa' has been (in this context) incorrectly translated as 'printing press', the actual mechanical device used to create printed materials. With it being highly unlikely that the White House employs a 'printing press' secretary, we can see how incorrect translations from Portuguese to English would lead to UKB being provided with problematic and potentially confusing contexts from which to disambiguate target words.

Of course, we must take into account that our translations from Portuguese to English are not likely to be as accurate as those obtained by Nóbrega and Pardo (2014). They describe using the WordReference API to extract dictionary definitions of Portuguese terms in English, but because that is no longer available we instead translate terms using the linked datasets in BabelNet, as described in section 4. Because lexical gaps in BabelNet are filled using machine translation for resource-poor languages, the resources on which our translations depend are unlikely to be as accurate from the outset as those from a verified dictionary API, and it would be interesting to explore whether alternative methods of producing our translations might give different results in our future work. However, we feel that the point demonstrated by the previous example still holds true – in trying to translate ambiguous terms from Portuguese to English in order to perform WSD over a larger underlying LKB in English, we are actually introducing more noise to the problem.

Table 2 shows that the accuracy of running UKB over the dictionary and graph files created from the Portuguese MultiWordNet is comparable with previously-reported state-of-the-art results – namely running UKB over the much larger English Wordnet to disambiguate words already *in* English. As well as the results shown in Table 1 and discussed in the preceding paragraphs, showing that translating Portuguese terms into English to make use of a much larger English LKB for disambiguation decreases accuracy, the results in Table 2 show that the smaller size of the Portuguese MultiWordNet does not have any considerable detrimental effect on the accuracy of the WSD process itself.

Besides the limited lexical coverage, there is no reason that using a smaller, language-specific LKB would produce any less accurate results for WSD. In fact, while language-specific dictionaries might be much smaller in certain languages, because the semantic relationships between concepts generally hold true across different languages, graphs representing these relationships as nodes and edges can actually be created from much fuller LKBs (as we have done using semantic relations from the English WordNet). This ensures that although not all words are covered locally, our

capacity to determine the relationships between them is still strong, providing consistently accurate results. Problems arise not necessarily from difficulty in determining the semantic relationships between concepts, but because the kinds of ambiguities and translation errors described above will occur when gaps in the lexical coverage of linked data are filled using machine translation.

For LOD, the implications are that while missing data for resource-poor languages can be filled in using machine translation (Navigli and Ponzetto, 2012), verified language-specific lexical resources still provide highly accurate results for tasks like WSD regardless of their comparative size – there is nothing to be gained by translating terms into other languages (such as English) to make use of fuller, larger LKBs. The increased connectivity and integration of lexical (and encyclopedic) resources in projects like DBpedia and BabelNet open up a world of possibilities for multilingual NLP, but filling the gaps using machine translation should only be a stopgap measure. Rather than abandon them in favour of the linked data already available, local efforts to grow, extend and expand language-specific lexical resources must continue, such that they can be continually re-integrated as LOD later as fuller, accurate and verified resources – thus increasing the overall quality of linked lexical data.

6 Conclusions

We have evaluated two approaches to performing graph-based WSD in Portuguese; 1) by using the smaller, language-specific Portuguese MultiWordNet as the underlying LKB, and 2) by first translating open-class words from Portuguese to English in order to use the much larger English WordNet as the underlying LKB. Comparing the results of both approaches with the human-assigned senses in a gold-standard annotated corpus, we have demonstrated that performing graph-based WSD using a smaller, language-specific LKB provides more accurate results than the approach of using the larger LKB by way of translating terms first. Furthermore, the accuracy of the language-specific approach is comparable with state-of-the-art results reported for graph-based WSD *in* English using WordNet.

For LOD, the implications of our results are that as well as in the short term making use of linked data where the gaps between resource-rich and resource-poor languages have been filled by machine translation, local efforts to grow and extend language-specific lexical resources such as WordNets should continue, so that these can be linked back to existing data as LOD later. This way, LOD will eventually consist not only of the connected semantic relationships across languages, but also fuller and verified lexical coverage, making rich multilingual NLP applications possible based on accurate linked data.

We plan to build on our work by making further comparisons to other graph-based WSD approaches, such as the disambiguation options available in BabelNet itself performed over its own linked data as an LKB, and by experimenting with alternative techniques and APIs for translating the open-class words from the context file into English in the first instance. It would also be interesting to combine approaches, augmenting results from accurate local lexical resources with results sourced via translated terms fed to larger resources. We also plan to explore whether LOD can play an effective role in the growth and extension of local lexical resources themselves, investigating whether there is an effective way that the expansion of local WordNets can be in some part automated based on manually checked and verified translations sourced from existing multilingual LOD.

Acknowledgments

This work has been undertaken and funded as part of the EU project QTLeap (EC/FP7/610516) and the Portuguese project DP4LT (PTDC/EEI-SII/1940/2012).

References

Eneko Agirre and Philip Edmonds. 2006. *Word Sense Disambiguation: Algorithms and Applications*. Springer.

Eneko Agirre and Aitor Soroa. 2008. Using the Multilingual Central Repository for Graph-Based Word Sense Disambiguation. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, may. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2008/.

Eneko Agirre and Aitor 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, pages 33–41, Athens, Greece. Association for Computational Linguistics.

Eneko Agirre, Oier Lopez De Lacalle, and Aitor Soroa. 2009. Knowledge-based WSD on Specific Domains: Performing Better Than Generic Supervised WSD. In *Proceedings of the 21st International Jont Conference on Artifical Intelligence*, IJCAI'09, pages 1501–1506, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Eneko Agirre, Oier López de Lacalle, and Aitor Soroa. 2014. Random Walks for Knowledge-based Word Sense Disambiguation. *Comput. Linguist.*, 40(1):57–84, March.

Jordi Atserias, Luís Villarejo, and German Rigau. 2004. Spanish WordNet 1.6: Porting the Spanish WordNet across Princeton Versions. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*, LREC '04.

Florbela Barreto, António Branco, Eduardo Ferreira, Amália Mendes, Maria Fernanda Bacelar Nascimento, Filipe Nunes, and João Silva. 2006. 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'06, pages 1438–1443.

António Branco and João 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, pages 179–182, Trento, Italy. Association for Computational Linguistics.

Sergey Brin and Lawrence Page. 1998. The Anatomy of a Large-Scale Hypertextual Web Search Engine. In *Seventh International World-Wide Web Conference (WWW 1998)*.

Director General of Translation. 2006. Nova Versão da Lista de Falsos Amigos Português-Espanhol / Español-Portugués. *A Folha: Boletim da Língua Portuguesa nas Instituições Europeias (Comissão Europeia)*, 23:19–27.

Christine Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick van Kleef, Sören Auer, and Christian Bizer. 2012. DBpedia - A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia. *Semantic Web Journal*, 6(2):167–195.

Michael Lesk. 1986. Automatic Sense Disambiguation Using Machine Readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone. In *Proceedings of the 5th Annual International Conference on Systems Documentation*, SIGDOC '86, pages 24–26, New York, NY, USA. ACM.

Lluís Màrquez, Luis Villarejo, M. A. Martí, and Mariona Taulé. 2007. SemEval-2007 Task 09: Multilevel Semantic Annotation of Catalan and Spanish. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, SemEval '07, pages 42–47, Stroudsburg, PA, USA. Association for Computational Linguistics.

Rada Mihalcea. 2005. Unsupervised Large-vocabulary Word Sense Disambiguation with Graph-based Algorithms for Sequence Data Labeling. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, pages 411–418, Stroudsburg, PA, USA. Association for Computational Linguistics.

MultiWordNet. n.d. The MultiWordNet project. `http://multiwordnet.fbk.eu/english/home.php`. Accessed: 2015-01-13.

Roberto Navigli and Mirella Lapata. 2007. Graph Connectivity Measures for Unsupervised Word Sense Disambiguation. In *Proceedings of the 20th International Joint Conference on Artifical Intelligence*, IJCAI'07, pages 1683–1688, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Roberto Navigli and Simone Paolo Ponzetto. 2012. BabelNet: The Automatic Construction, Evaluation and Application of a Wide-Coverage Multilingual Semantic Network. *Artificial Intelligence*, 193:217–250.

Roberto Navigli and Paola Velardi. 2005. Structural Semantic Interconnections: A Knowledge-Based Approach to Word Sense Disambiguation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(7):1075–1086, July.

Roberto Navigli, Kenneth C. Litkowski, and Orin Hargraves. 2007. SemEval-2007 Task 07: Coarse-grained English All-words Task. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, SemEval-2007, pages 30–35, Stroudsburg, PA, USA. Association for Computational Linguistics.

Roberto Navigli, David Jurgens, and Daniele Vannella. 2013. SemEval-2013 Task 12: Multilingual Word Sense Disambiguation. In *International Workshop on Semantic Evaluation*, SemEval-2013.

Steven Neale, João Silva, and António Branco. 2015. A Flexible Interface Tool for Manual Word Sense Annotation. In *Proceedings of the 11th Joint ACL-ISO Workshop on Interoperable Semantic Annotation*, ISA-11, pages 67–71, London, UK. Association for Computational Linguistics.

Fernando Antônio Asvedo Nóbrega and Thiago Alexandre Salgueiro Pardo. 2014. General Purpose Word Sense Disambiguation Methods for Nouns in Portuguese. *Computational Processing of the Portuguese Language*, 8775:94–101.

OmegaWiki. n.d. OmegaWiki. `http://en.omegawiki.org`. Accessed: 2015-07-10.

Martha Palmer, Christiane Fellbaum, Scott Cotton, Lauren Delfs, and Hoa Trang Dang. 2001. English Tasks: All-Words and Verb Lexical Sample. In *Proceedings of SENSEVAL-2 Second International Workshop on Evaluating Word Sense Disambiguation Systems*, SensEval-2, pages 21–24, Toulouse, France, July. Association for Computational Linguistics.

Sameer S. Pradhan, Edward Loper, Dmitriy Dligach, and Martha Palmer. 2007. SemEval-2007 Task 17: English Lexical Sample, SRL and All Words. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, SemEval '07, pages 87–92, Stroudsburg, PA, USA. Association for Computational Linguistics.

Judita Preiss and Mark Stevenson. 2013. Dale: A word sense disambiguation system for biomedical documents trained using automatically labeled examples. In *Proceedings of the 2013 NAACL HLT Demonstration Session*, pages 1–4. Association for Computational Linguistics.

Ravi Sinha and Rada Mihalcea. 2007. Unsupervised Graph-basedWord Sense Disambiguation Using Measures of Word Semantic Similarity. In *Proceedings of the International Conference on Semantic Computing*, ICSC '07, pages 363–369, Washington, DC, USA. IEEE Computer Society.

Benjamin Snyder and Martha Palmer. 2004. The English All-Words Task. In Rada Mihalcea and Phil Edmonds, editors, *Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, SensEval-3, pages 41–43, Barcelona, Spain, July. Association for Computational Linguistics.

Lucia Specia, Maria das Graças V. Nunes, and Mark Stevenson. 2005. Exploiting Rules for Word Sense Disambiguation in Machine Translation. *Procesamiento del Lenguaje Natural*, 35:171–178.

Lucia Specia, Mark Stevenson, and Maria Graças V. Nunes. 2007. Learning expressive models for words sense disambiguation. In *In Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 41–48.

Lucia Specia. 2006. A Hybrid Relational Approach for WSD: First Results. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, COLING ACL '06, pages 55–60, Stroudsburg, PA, USA. Association for Computational Linguistics.

Mark Stevenson, Eneko Agirre, and Aitor Soroa. 2011. Exploiting Domain Information for Word Sense Disambiguation of Medical Documents. *Journal of the American Medical Informatics Association*, 19(2):235–40.

Piek Vossen. 2004. Eurowordnet: A multilingual database of autonomous and language-specific wordnets connected via an inter-lingualindex. *International Journal of Lexicography*, 17(2):161–173.

Wikidata. n.d. Wikidata. `http://en.wikidata.org`. Accessed: 2015-07-10.

Wikipedia. n.d. Wikipedia, the free encyclopedia. `https://en.wikipedia.org`. Accessed: 2015-07-10.

Wiktionary. n.d. Wiktionary, the free dictionary. `http://en.wiktionary.org`. Accessed: 2015-07-10.

WordReference.com. n.d. WordReference API Documentation. `http://www.wordreference.com/docs/api.aspx`. Accessed: 2015-06-04.