Exploiting Portuguese Lexical Knowledge Bases for Answering Open Domain Cloze Questions Automatically

Hugo Gonçalo Oliveira, Inês Coelho, Paulo Gomes

CISUC, University of Coimbra
DEI, Pólo 2, Pinhal de Marrocos
3030-290 Coimbra, Portugal
hroliv@dei.uc.pt, micoelho@student.dei.uc.pt, pgomes@dei.uc.pt

Abstract

We present the task of answering cloze questions automatically and how it can be tackled by exploiting lexical knowledge bases (LKBs). This task was performed in what can be seen as an indirect evaluation of Portuguese LKB. We introduce the LKBs used and the algorithms applied, and then report on the obtained results and draw some conclusions: LKBs are definitely useful resources for this challenging task, and exploiting them, especially with PageRanking-based algorithms, clearly improves the baselines. Moreover, larger LKBs, created automatically and not sense-aware led to the best results, as opposed to handcrafted LKBs structured on synsets.

Keywords: lexical knowledge bases, cloze question-answering, semantic relations, evaluation

1. Introduction

More than challenging tasks, word tests are attractive scenarios for assessing lexical-semantic resources, such as lexical knowledge bases (LKBs), where words are organised according to their meaning. Those tests are suitable for application-based evaluations, where the LKB is exploited for performing an independent task. This kind of evaluation is referred by Brank et al. (2005) as one of the most common approaches for evaluating ontologies, in addition to manual evaluation, comparison with a gold standard, and comparison with a dataset.

During the last years, we have been working on the development of several public LKBs for Portuguese – PAPEL (Gonçalo Oliveira et al., 2008), CARTÃO (Gonçalo Oliveira et al., 2011) and Onto.PT (Gonçalo Oliveira and Gomes, 2013) – all created automatically, but structured differently. These are suitable targets for indirect evaluations consisting of the completion of independent tasks, such as those that have been performed, for English, with the help of Princeton WordNet (Fellbaum, 1998), which include word tests. Besides complementing the results of manual evaluation, application-based may be used to assess the utility of the target LKB and enable their objective comparison with other LKBs, in the specific task addressed. While searching for word tests in Portuguese, with enough examples and ready to be computationally processed, we became aware of a set of cloze questions, also known as fill-in-the-blank questions, on no specific domain, and created in the scope of an assisted Portuguese learning system.

Those questions were used in this work, whose main goals are twofold. We aim to: (i) study to what extent LKBs can be exploited for answering cloze questions automatically; (ii) perform an application-based comparison of several available Portuguese LKBs, organised and created differently. We came to acknowledge that answering cloze questions, especially when using only a LKB, is a challenging task and took additional conclusions on the best algorithms, among those we have selected to this task. They were originally designed to exploit LKBs organised differently, whether structured in plain words or in word senses and synsets. The best results were obtained with PageRank, an algorithm based on random-walks on the LKB graph, using a large LKB, created automatically, not sense-aware, and covering a broad range of relation types.

After this introduction, we describe work focused on two tasks that share some similarities to that we are approaching. Then, we enumerate the resource used and the algorithms applied. The obtained results for different LKB/algorithm configurations are presented in the following sections, first regarding the full set of questions, then only the subset with all answers covered by all the exploited LKBs. Before concluding, we illustrate the results of this work with two questions and the answer ranking for all tested configurations.

2. Related work

For English, in the late 1990s and early 2000s, attention has been given to answering synonymy questions from the Test of English as a Foreign Language (TOEFL)[1] automatically. Proposed methods included corpus-based approaches (e.g. Landauer and Dumais (1997)), lexicon-based approaches (e.g. Jarmasz and Szpakowicz (2003)) and the combination of both (e.g. Turney et al. (2003)).

After presenting optimal results for a combined approach (97.5% accuracy), Turney et al. (2003) claimed that this problem was solved for English and turned the attention to other interesting problems, such as analogies. In 2012, however, Bullinaria and Levy (2012) achieved 100% accuracy, using word co-occurrence statistics and a large corpus. Yet, the same authors comment that this result is misleading, because parameters had been specifically tuned and are unlikely to generalisee well to new tasks or corpora.

There are other recent approaches to the TOEFL synonymy test that exploit Princeton WordNet, including Siblini and

---

[1] The state of the art for this task is presented in http://aclweb.org/aclwiki/index.php?title=TOEFL_Synonym_Questions_(State_of_the_art)

Kosseim (2013) and Pilehvar et al. (2013), who achieved 91.25% and 96.25% accuracy, respectively. Siblini and Kosseim (2013) used WordNet to create a network with words connected to the synsets that include them, and synsets connected to other related synsets. Connections were weighted according to the relation type (e.g. hypernymy has a lower weight than domain-of) and the relatedness between two words was given by the lowest cost path. Pilehvar et al. (2013) compute word relatedness based on semantic signatures, obtained from applying PageRank over a network with the WordNet senses.

The broad utilisation of the TOEFL synonymy test on this scope confirms that a word test can guide the research on lexical semantics. But another task worth mentioning is automatic multiple-choice questions answering, based on a provided document. Since 2013, the Entrance Exams task (Peñas et al., 2013) has been part of the Question Answering for Machine Reading challenge in the Cross-Language Evaluation Forum (QA4MRE@CLEF). This task uses real English exams and assesses how artificial systems understand a given textual document, by answering a set of multiple-choice questions. The best approach to the Entrance Exams (Banerjee et al., 2013) generates answer hypothesis by combining the question and each possible answer; retrieves the most relevant document sentences for each hypothesis, based on the TF-IDF and n-gram overlap; and ranks the hypothesis according to their entailment with the retrieved sentences.

Our cloze question task deals with the selection of a suitable word for a blank in a given sentence. Therefore, both of the aforementioned tasks have some similarities with the task we have at our hands. Similarly to ours, both of them provide four alternative answers, including one correct and three distractors. Moreover, as in the synonymy tests, the possible answers for our task are single words. In the Entrance Exams they are natural language statements. On the other hand, in the Entrance Exams, the questions rely on a provided context, in the form a document, that should be, to some degree, understood. In the synonymy tests, the target word is the only available context. Although smaller, our cloze questions can also be seen as some kind of context where the correct alternative must fit, so that it is coherent. In terms of complexity, we can say that our task is between the other two.

3. Resources used

This section describes the external resources used in this work, including the cloze questions set, illustrated with a sample question, and the LKBs exploited.

3.1. Cloze Questions

REAP.PT (Silva et al., 2012) is a computer assisted language learning tutoring system for European Portuguese. After becoming aware of REAP.PT, we asked its developers for cloze questions, created in the scope of their project, which we would use to assess the LKBs we were developing. They answered positively and kindly provided us a set with 3,890 cloze questions, generated from sentences of the CETEMPúblico newspaper corpus (Santos and Rocha, 2001), with candidate stems selected from the Portuguese

Academic Word List (Baptista et al., 2010). Both the selection of stems (Correia et al., 2010) and distractors (Correia et al., 2012) were automatically refined to be in accordance. For example, available lexical resources were used to find (and replace) distractors that could be syononyms of the correct answer.

Cloze questions, illustrated below, consist of: (i) a sentence where one word is missing (stem); (ii) a shuffled list of alternatives, including the missing word and a set of three distractors. The goal of the task is to select the missing word (in bold) from the list of alternatives.

Houve influência da _____ oriental sobre a grega, porém não se pode superestimar a importância dessa influência.

(There was some influence of the oriental _____ on the greek, but we cannot overestimate the importance of this influence.)

    (a) cultura *(culture)*   (c) praticante *(practicioner)*
    (b) exibição *(exhibition)*   (d) inteligência *(intelligence)*

3.2. Lexical Knowledge Bases

In this work, we have exploited several public domain LKB for Portuguese, including not only those we developed (PAPEL, CARTÃO, Onto.PT), but also LKBs created by other researchers (TeP and OpenWordNet.PT). In the following paragraphs, we briefly describe each one of them.

TeP (Maziero et al., 2008) is an electronic thesaurus for Brazilian Portuguese, created manually. Its current version, TeP 2.0, contains more than 44,000 lexical items, organised in 19,888 synsets, and also 4,276 antonymy relations between synsets. Similarly to a wordnet, TeP is structured in synsets, but they are only connected by antonymy relations. Therefore, we decided to transform TeP into a lexical network, where plain words are connected either by synonymy, if they belong to the same synset, or by antonymy, if they belong to antonymous synsets.

PAPEL (Gonçalo Oliveira et al., 2008) is a lexical-semantic network extracted automatically from a proprietary Portuguese dictionary. Its last version, 3.5, contains about 102,000 lexical items, connected by about 191,000 semantic relation instances covering a rich set of types, including synonymy, hypernymy, several subtypes of meronymy, causation, purpose, state, quality, manner and property.

CARTÃO (Gonçalo Oliveira et al., 2011) is a resource with the same structure of PAPEL, that integrates PAPEL and relation instances extracted from two open dictionaries. Its current version contains about 146,000 lexical items and 286,000 relation instances. For this work, we have also augmented CARTÃO with additional synonymy instances obtained from TeP, OpenThesaurus.PT[2] and OpenWordNet.PT (de Paiva et al., 2012), all public lexical resources for Portuguese, created manually and organised in synsets. This resulted in CARTÃO+.

Onto.PT (Gonçalo Oliveira and Gomes, 2013) is a large wordnet for Portuguese, created automatically after integrating the relation instances of CARTÃO in the

---

synsets of TeP and OpenWordNet.PT. Its current version, 0.6 (Gonçalo Oliveira and Gomes, 2014), contains about 168,000 lexical items and 238,000 word senses, organised in about 117,000 synsets, connected by 341,000 relation instances, that cover the same types as PAPEL. OpenWordNet.PT (de Paiva et al., 2012) is a Portuguese wordnet that resulted from the manual translation of a set of base synsets in Princeton WordNet 3.0. Semantic relations were inherited from the latter, given the synset matches. Currently, it contains about 48,000 lexical items and 54,000 word senses, organised in about 39,000 synsets, connected by 84,000 relation instances, that cover the same types as WordNet 3.0.

4. Algorithms

The goal of this task is to select a word from a set of shuffled alternatives, according to its suitability to fill a blank, in context. Given its similarities to word sense disambiguation (WSD), we tackled this task either with state-of-the-art unsupervised WSD algorithms or others inspired on those, which see a LKB as an undirected graph with semantic relations connecting words or word structures. By no means we intended to cover extensively all the algorithms reported in the literature, that would suit this task. Since our main goal was to compare LKBs, this is not a problem, as long as the same algorithms are used for LKBs of the same kind. All the algorithms share a preamble, where the question sentence is POS-tagged and lemmatised. For coherence, the correct answer is put in the stem, and the OpenNLP toolkit[3] is used to POS-tag the complete sentence. Lemmatisation is achieved by applying a set of rules based on the POS-tags. After tagging, the correct answer is removed from the sentence.

All open-category words (nouns, verbs, adjectives, adverbs) of the sentence are used as its context. After the preamble, algorithms proceed differently considering the different organisations of the LKBs, which imply different kinds of nodes in the graph. For lexical networks, where the nodes are plain lexical items, we tested three algorithms, described as follows:

- **Adjacencies similarity (*AdjSim*):**

    1. Augment the question context with words related directly to the context words;

    2. Set the context of each alternative answer to its word and words related directly;

    3. Compute similarities between the contexts of the question and each of its alternatives;

    4. Select the alternative with the highest similarity.

- **Minimum combined distance (*MinDist*):**

    1. Compute the distance between each alternative word and the context words.

    2. Select the alternative that minimises the sum of the distances.

- **Personalized PageRank on a word graph (*PR-words*),** an adaptation of the WSD method (Agirre et al., 2013):

    1. Distribute starting weights uniformly over the nodes corresponding to context words;

    2. Run PageRank for 30 iterations;

    3. Select the alternative with the highest rank.

For wordnet-like LKBs, where the nodes are groups of words (synsets), we tested another three algorithms:

- **Adapted Lesk (*AdLesk*)** where, instead of gloss words[4], the synset context includes the words in the synset and in synsets related directly:

    1. Assign the most suitable synset in the LKB to each context word, using the Lesk WSD algorithm adapted for wordnets (Banerjee and Pedersen, 2002);

    2. Augment the question context with all the contexts of all the assigned synsets;

    3. Compute the intersection between the context of synsets including an alternative word and the question context;

    4. Select the alternative in the synset maximising the intersection.

- **Personalized PageRank on a synset graph (*PR-syn*),** the WSD method by Agirre et al. (2013):

    1. Distribute starting weights uniformly over the synsets that contain context words;

    2. Run PageRank for 30 iterations;

    3. From the synsets that include alternative words, select the alternative in the highest ranked.

- **Personalized PageRank on a sense graph (*PR-sen*),** another adaptation of the WSD method, using the graph of word senses extracted from the wordnet[5]:

    1. Distribute starting weights uniformly over all the senses of context words;

    2. Run PageRank for 30 iterations;

    3. From the senses of the alternative words, select the alternative in the highest ranked.

5. Answering the full set of cloze questions

The algorithms described in the previous section were used for exploiting the LKBs presented earlier towards the goal of answering the full set of 3,890 cloze questions. This section reports on the obtained results. As different similarity measures could be used for *Adj-Sim*, we only show the values for simple intersection (*AdjSim-int*) and cosine similarity (*AdjSim-cos*). The presented results can be compared to two baselines: (i) the lower-bound random-choice baseline (25%, since there are always four alternative answers), not presented in the result tables; (ii) another baseline that selects the most frequent alternative in the AC/DC corpora (Santos and Bick, 2000), the first entry of the result tables (*MostFrequent*).

---

[3]OpenNLP is available from `https://opennlp.apache.org/`

[4]OpenWordNet.PT does not contain Portuguese glosses.

[5]A word sense is given by the inclusion of a word in a synset.

5.1. Results presentation

For each configuration, consisting of a LKB and an applied algorithm, table 1 has four columns with results. The first column ($Correct_1$) shows the number and proportion of questions to which the correct answer, and only this one, was given. However, especially in the non-PageRank algorithms, there is not enough information to give a single answer. Therefore, table 1 also presents the number of questions where there was a tie in the best answer (Ties). To select only one answer from the tied ones, we followed two different strategies: (i) random tied alternative ($Correct_r$); (i) most frequent tied alternative ($Correct_f$), also according to the AC/DC frequencies. For the majority of situations, the latter leads to better results.

5.2. Best results

The best results are obtained by PageRanking CARTÃO. This configuration led to about 41% of the questions answered correctly, which is far from being an excellent performance, and shows that we have a challenging task at our hands. It should be noted that some questions have a short context – 355 have less than 8 content words – and others use named entities, which are not expected to be found in a LKB – 1,752 contain at least one capitalised word. On the other hand, this result is clearly higher than the random selection baseline and about 10% higher than the most frequent baseline. This confirms that the structure of CARTÃO makes sense and it is advantageous to use it for answering this kind of open domain questions.

5.3. Main conclusions

The results show that the PageRank based algorithms perform better than the others, which is in line with the best approaches for WSD (Agirre et al., 2013). In opposition to the algorithms that only consider the adjacencies, PageRank takes advantage of random walks and considers the structure of the whole graph. Moreover, nodes are ranked according to their relevance to the context with a higher precision than in the ranks of the other algorithms, and this results in much fewer ties. On the other hand, the adjacency-based algorithms rely only on the intersection between the question and the alternative context in the LKB, which is an important limitation, especially when there are no intersections between them. Combined with the low rank precision, this leads to many ties. Still, after selecting one of the tied alternatives, with few exceptions, these algorithms also outperform the baselines. Regarding ties, the distance-based algorithm is more or less between the other two kinds of algorithm. As for the number of correct answers, for all but PAPEL, its results are also between the other algorithms.

Another conclusion is that, although the algorithms for each kind of LKB (plain words and synset-oriented) are slightly different, more questions are answered correctly when using plain words LKBs. Not only CARTÃO, but also CARTÃO+ and PAPEL outperform Onto.PT and OpenWordNet.PT. This suggests that, for this specific task, the notion of sense is not very important. The results of TeP are closer to those of the synset-oriented LKBs. But we should recall that TeP, originally a synset-oriented LKB, only covers synonymy and antonymy, while the other

plain-word LKBs cover a broad range of relations (see section 3.2.), which are clearly important here.

Though better results are obtained with CARTÃO than with CARTÃO+, the resource size also seems to matter. Apart from this exception, better results are obtained with larger resources of the same kind. CARTÃO led to better results than PAPEL, and PAPEL to better results than TeP. On the synset-oriented LKBs, Onto.PT led to better results than OpenWordNet.PT. We may conclude that resources covering a larger part of the lexicon, both in terms of words/concepts and relations, are better suited for this task, because they have an increased probability of both covering more alternative answers and having more information (relations) on them. We can even go further and say that, in this situation, the coverage of the resource seems to be more critical than its full reliability. Although PAPEL, CARTÃO[6] and Onto.PT are created automatically[7] and should thus have more reliability issues than manually created LKBs, they are larger and led to better results than those of TeP and OpenWordNet.PT.

5.4. Results by POS

Since different LKBs cover and organise words of different parts-of-speech (POS) differently, in order to study its impact, table 2 presents the results according to the POS of the answers, which is always the same for the four alternatives. Regardless the category, PageRank still leads to the best results. However, for nouns and adjectives, the best LKB is CARTÃO+, while for the verbs and adverbs it is CARTÃO. If the overall results already suggested that the synonymy information in the handcrafted thesauri added some noise to CARTÃO+, now it is more clear that the problem is on the verbs and adverbs. After analysing the structure of TeP, we noticed that, on average, its verb items have 2.6 senses and its verb synsets have 5.7 words (the largest has 53). For nouns and adjectives, these numbers are 1.7 and 1.4 senses, respectively and 3.5 item for both.

For most configurations, results are better for questions with noun and adjective answers. Besides the higher ambiguity of verbs, this might also result from a better coverage of those POS by the LKBs, in opposition to verbs and adverbs. Also interesting are the results of *PR-syn* with Onto.PT for adjectives, which are the second best, outperforming PAPEL and CARTÃO.

6. Results for the subset of covered questions

As mentioned earlier, the size/coverage of the LKB has some impact on the number of questions answered correctly. To have a better insight on the quality of the LKBs'

---

[6]The evaluation of previous versions of PAPEL and CARTÃO (Gonçalo Oliveira and Gomes, 2013) estimated that the correction of their relations depended on their type. In PAPEL, it ranged from 99% (synonymy) and 91-94% (hypernymy), to 72-78% (property-of) and 69% (purpose-of). In CARTÃO, it ranged from 99% (synonymy) and 88-90% (hypernymy), to 71-77% (property-of) and 73-74% (purpose-of)

[7]The evaluation of a previous version of Onto.PT (Gonçalo Oliveira and Gomes, 2013) estimated that its hypernymy instances were about 65% accurate, while the remaining relation instances were, on average, about 80% accurate. There have been recent improvements though (Gonçalo Oliveira and Gomes, 2014).

Resource	Algorithm	Correct$_1$		Ties		Correct$_r$		Correct$_f$	
Baseline	*MostFrequent*	1,264	(32.49%)	53	(1.36%)	1,279	(32.88%)	-	(-%)
TeP 2.0	*AdjSim-int*	589	(15.14%)	1,960	(50.39%)	1,254	(32.24%)	1,346	(34.60%)
	AdjSim-cos	634	(16.30%)	1,890	(48.59%)	1,253	(32.21%)	1,356	(34.86%)
	MinDist	546	(14.04%)	2,116	(54.40%)	1,307	(33.60%)	1,372	(35.27%)
	PR-words	1,390	(35.73%)	179	(4.60%)	1,443	(37.10%)	1,449	(37.25%)
PAPEL 3.5	*AdjSim-int*	679	(17.46%)	1,934	(49.72%)	1,272	(32.70%)	1,370	(35.22%)
	AdjSim-cos	739	(19.00%)	1,732	(44.52%)	1,286	(33.06%)	1,333	(34.27%)
	MinDist	1,063	(27.33%)	334	(8.59%)	1,199	(30.82%)	1,238	(31.83%)
	PR-words	1,524	(39.18%)	49	(1.26%)	1,537	(39.51%)	1,540	(39.59%)
CARTÃO	*AdjSim-int*	883	(22.70%)	1,578	(40.57%)	1,421	(36.53%)	1,457	(37.46%)
	AdjSim-cos	956	(24.58%)	1,255	(32.26%)	1,353	(34.78%)	1,382	(35.53%)
	MinDist	1,225	(31.49%)	319	(8.20%)	1,359	(34.94%)	1,402	(36.04%)
	PR-words	**1,598**	**(41.08%)**	14	(0.36%)	**1,601**	**(41.16%)**	**1,603**	**(41.21%)**
CARTÃO+	*AdjSim-int*	1,049	(26.97%)	1,106	(28.43%)	1,430	(36.76%)	1,458	(37.48%)
	AdjSim-cos	1,117	(28.71%)	828	(21.29%)	1,377	(35.40%)	1,398	(35.94%)
	MinDist	1,173	(30.15%)	289	(7.43%)	1,312	(33.73%)	1,339	(34.42%)
	PR-words	1,585	(40.75%)	6	(0.15%)	1,588	(40.82%)	1,589	(40.85%)
Onto.PT 0.6	*AdLesk*	66	(1.70%)	3,428	(88.12%)	1,061	(27.28%)	1,172	(30.13%)
	PR-syn	1,463	(37.61%)	15	(0.39%)	1,470	(37.79%)	1,470	(37.79%)
	PR-sen	1,436	(36.92%)	14	(0.36%)	1,440	(37.02%)	1,442	(37.07%)
OpenWN.PT	*AdLesk*	123	(3.16%)	3,098	(79.64%)	1,090	(28.02%)	1,146	(29.46%)
	PR-syn	1,230	(31.62%)	171	(4.40%)	1,285	(33.03%)	1,288	(33.11%)
	PR-sen	1,222	(31.41%)	222	(5.71%)	1,244	(31.98%)	1,267	(32.57%)

Table 1: Results when answering the full set of cloze questions, for different algorithms and LKBs.

Resource	Algorithm	Nouns (1,769 questions)		Verbs (1,077)		Adjectives (809)		Adverbs (235)	
		Correct$_1$	Correct$_f$	Correct$_1$	Correct$_f$	Correct$_1$	Correct$_f$	Correct$_1$	Correct$_f$
Baseline	*MostFrequent*	35.44%	-	32.59%	-	28.43%	-	30.21%	-
TeP 2.0	*AdjSim-int*	15.15%	40.64%	18.76%	27.48%	12.48%	32.14%	7.66%	30.21%
	AdjSim-cos	15.66%	40.81%	22.10%	27.86%	12.48%	32.01%	7.66%	31.91%
	MinDist	15.26%	40.53%	14.58%	29.81%	12.48%	32.63%	7.66%	29.79%
	PR-words	41.78%	42.62%	29.06%	29.25%	35.97%	38.32%	20.00%	29.79%
PAPEL 3.5	*AdjSim-int*	24.14%	38.67%	13.56%	32.22%	11.25%	33.13%	6.38%	30.21%
	AdjSim-cos	26.46%	36.69%	14.95%	32.03%	11.62%	32.76%	6.81%	31.49%
	MinDist	28.15%	33.47%	26.28%	30.45%	29.05%	31.52%	20.00%	26.81%
	PR-words	44.88%	45.00%	33.80%	33.89%	38.81%	38.94%	22.13%	27.23%
CARTÃO	*AdjSim-int*	30.86%	41.89%	17.08%	32.96%	16.69%	34.98%	7.66%	33.19%
	AdjSim-cos	32.56%	37.48%	20.33%	33.24%	17.55%	34.86%	8.09%	33.62%
	MinDist	33.13%	38.78%	28.60%	32.59%	34.36%	37.45%	22.55%	26.38%
	PR-words	45.56%	45.62%	**35.75%**	**35.75%**	41.04%	41.04%	**31.91%**	**33.62%**
CARTÃO+	*AdjSim-int*	33.75%	43.02%	24.51%	30.27%	20.64%	36.34%	8.94%	32.77%
	AdjSim-cos	35.05%	39.01%	27.76%	31.75%	21.63%	35.72%	9.79%	32.77%
	MinDist	31.94%	36.69%	27.30%	32.13%	32.39%	35.23%	22.13%	25.11%
	PR-words	**47.03%**	**47.09%**	31.66%	31.66%	**41.90%**	**42.03%**	31.06%	31.91%
Onto.PT 0.6	*AdLesk*	0.90%	31.32%	0.28%	27.02%	4.94%	31.15%	2.98%	31.91%
	PR-syn	39.40%	39.46%	32.68%	32.87%	41.78%	41.78%	32.34%	34.04%
	PR-sen	39.46%	39.51%	32.03%	32.03%	40.42%	40.42%	28.09%	30.21%
OpenWN.PT	*AdLesk*	2.20%	30.86%	0.74%	27.11%	7.42%	28.68%	6.81%	32.34%
	PR-syn	33.97%	34.43%	32.50%	32.78%	30.28%	31.15%	14.47%	31.49%
	PR-sen	34.26%	34.71%	32.40%	32.59%	29.67%	31.77%	11.49%	19.15%

Table 2: Results when answering the full set of cloze questions, for different algorithms and LKBs, according to POS.

contents, for the 1,162 questions where all alternative answers exist in all the exploited LKBs, table 3 shows the number of questions answered correctly before (Correct$_1$) and after (Correct$_f$) untying by selecting the most frequent tied answer.

There is no great change in the top LKBs and algorithms, which leads to no additional conclusions. Moreover, as expected, the number of ties significantly drops. A surprising outcome is that the PageRank-based results are worse than those on the full set of questions. For instance, for CARTÃO, 2.5% less questions are answered correctly in this case. The performance of the baseline also drops and, in the adjacency-based algorithms, some of the results are now better and others are worst than in the full set. This apparently means that the probability of the LKBs not covering an incorrect alternative is higher than the probability of selecting the correct alternative out of four possibilities.

Using this subset of questions, we also investigated on the impact of the size of the question context on the number of correct answers. For this purpose, for each configuration, we measured the correlation between the context size and the proportion of questions with that context size, answered correctly. We only considered context with 5 to 18 content words, the range to which there were at least 20 questions for each size. The obtained values were also unexpected, as the correlation changes drastically depending on the configuration, and there appears to be no correlation between context size and correct answers. The highest positive correlation occurs for *AdjSim-cos* in CARTÃO (0.34), and the highest negative correlation occurs for *PR-sen* in Onto.PT (-0.91). In fact, all PageRank-based algorithms are negatively correlated with the question context – -0.002 (TeP), -0.25 (PAPEL), -0.40 (CARTÃO), -0.70 (CARTÃO+), -0.62 and -0.91 (Onto.PT), -0.37 and -0.21 (OpenWN.PT). This suggests that larger contexts add noise to the Personalized PageRank algorithm, especially in the LKBs created automatically, which already tend to be noisier.

Resource	Algorithm	Correct$_1$		Ties		Correct$_r$		Correct$_f$	
Baseline	*MostFrequent*	345	(29.69%)	1	(0.09%)	345	(29.69%)	0	(0.00%)
TeP 2.0	*AdjSim-int*	182	(15.66%)	519	(44.66%)	313	(26.94%)	343	(29.52%)
	AdjSim-cos	209	(17.99%)	474	(40.79%)	320	(27.54%)	349	(30.03%)
	MinDist	152	(13.08%)	635	(54.65%)	315	(27.11%)	361	(31.07%)
	PR-words	380	(32.70%)	0	(0.00%)	380	(32.70%)	380	(32.70%)
PAPEL 3.5	*AdjSim-int*	229	(19.71%)	515	(44.32%)	387	(33.30%)	393	(33.82%)
	AdjSim-cos	265	(22.81%)	436	(37.52%)	383	(32.96%)	396	(34.08%)
	MinDist	308	(26.51%)	118	(10.15%)	362	(31.15%)	367	(31.58%)
	PR-words	432	(37.18%)	0	(0.00%)	432	(37.18%)	432	(37.18%)
CARTÃO	*AdjSim-int*	267	(22.98%)	381	(32.79%)	398	(34.25%)	397	(34.17%)
	AdjSim-cos	331	(28.49%)	238	(20.48%)	391	(33.65%)	401	(34.51%)
	MinDist	337	(29.00%)	121	(10.41%)	396	(34.08%)	401	(34.51%)
	PR-words	**448**	**(38.55%)**	0	(0.00%)	**448**	**(38.55%)**	**448**	**(38.55%)**
CARTÃO+	*AdjSim-int*	328	(28.23%)	194	(16.70%)	394	(33.91%)	399	(34.34%)
	AdjSim-cos	376	(32.36%)	96	(8.26%)	396	(34.08%)	407	(35.03%)
	MinDist	321	(27.62%)	103	(8.86%)	368	(31.67%)	369	(31.76%)
	PR-words	436	(37.52%)	0	(0.00%)	436	(37.52%)	436	(37.52%)
Onto.PT 0.6	*AdLesk*	0	(0.00%)	1,097	(94.41%)	265	(22.81%)	300	(25.82%)
	PR-syn	403	(34.68%)	4	(0.34%)	405	(34.85%)	405	(34.85%)
	PR-sen	393	(33.82%)	4	(0.34%)	393	(33.82%)	393	(33.82%)
OpenWN.PT	*AdLesk*	0	(0.00%)	1,129	(97.16%)	284	(24.44%)	327	(28.14%)
	PR-syn	355	(30.55%)	2	(0.17%)	355	(30.55%)	355	(30.55%)
	PR-sen	356	(30.64%)	3	(0.26%)	356	(30.64%)	356	(30.64%)

Table 3: Results when answering the subset of covered cloze questions, for different algorithms and LKBs.

7. Examples

To illustrate our work and its results, we show the answers ranking by all configurations, for two questions: (i) one answered correctly by the majority of the configurations (table 4); (ii) another that all configurations failed to answer correctly (table 5). In the first example, the correct answer was *admissão* (admission), which is strongly related to the context words *ingresso* (entering) and *concurso* (competition). On the other hand, in the example of table 5, most configurations selected the incorrect alternative *conduzir* (drive). This question was more difficult to the LKBs, because the word *conduzir* happens to be indirectly related with the context word *partida* which, in Portuguese, might have the meaning of a game/match but also of start, as in a race (where drivers participate). As for the correct alternative (*antecipar*), all the LKBs lack an explicit connection between this concept and the context words *jogador* (player) and *xadrez* (chess).

This question is one of many that include a proper noun (Daniel). While here, it is irrelevant, there are 1,752 questions with capitalised words, most of them named entities, including famous people (e.g. Guterres, Bruce Springsteen), organisations (e.g. Benfica, TVI, ONU) or brands (e.g. Audi, Apple), to mention a few. The LKBs are not expected to have any kind of information on them, which, of course, increases the difficulty of selecting the correct answer without any additional source of knowledge.

8. Concluding remarks

We have described the task of answering cloze questions automatically, which we have tackled by exploiting the structure of Portuguese LKBs. We acknowledged that this is a challenging task, as our best configuration only answered about 41% of the questions correctly. This is, nevertheless, substantially higher than the random choice baseline. Our results are much lower than those obtained for answering TOEFL synonymy questions, using WordNet (96%, by Pilehvar et al. (2013)), and they are in line with the best results obtained for answering multiple-choice

questions about a given text (Banerjee et al., 2013)[8]. But our task cannot be blindly compared to those. On the one hand, it is more complex than the synonymy questions. On the other hand, it is still different from the multiple choice tests, as those imply some deeper understanding both of the provided text and answers, which are statements and not single words. Moreover, all the used LKBs differ from WordNet, either on structure, creation approach, and size. Given our main goal, the performed experiments were important for taking the following conclusions:

- Outperforming the baselines indicates that the contents and organisation of the LKB makes sense and are useful for this task.

- Seeing the LKBs as a graph and applying PageRank-based algorithms consistently leads to better results.

- Exploiting LKBs based on plain words leads to better results, in opposition to exploiting those organised in senses/synsets.

- The size and the coverage of the LKB plays an important role. For this task, it is more critical to have a larger resource created automatically than a smaller one that is virtually 100% reliable.

These conclusions, however, suffer from the limitations of task-based evaluations and cannot be generalised. In the future, we will look for other datasets with word tests, which will help in the assessment of Portuguese LKBs. As far as we know, we are the only researchers using this cloze question set for evaluating lexical resources. It would definitely be interesting to have more people using not only this, but perhaps other common datasets, for evaluating their lexical resources and approaches, in a similar manner

---

[8]If no final answer is given for ties, we can convert our results to the c@1 measure, used in the Entrance Exams task, that favours unanswered questions to incorrect answers. As in our best results there are almost no ties, their result (41.08% for CARTÃO with PageRank) can be approximated to c@1 = 0.41.

Resource	Algorithm	(a)	(b)	(c)	(d)
Baseline	*MostFrequent*	2,784	6,808	**13,831**	2,044
TeP 2.0	*AdjSim-int*	**9.0**	0.0	0.0	0.0
	AdjSim-cos	**0.091147**	0.0	0.0	0.0
	MinDist	**5.64**	6.0	6.0	6.0
	PR-words	**0.00900**	3.151×10^{-5}	6.836×10^{-5}	1.157×10^{-4}
PAPEL 3.5	*AdjSim-int*	**4.0**	0.0	0.0	0.0
	AdjSim-cos	**0.07597**	0.0	0.0	0.0
	MinDist	**3.5**	4.5	4.93	4.64
	PR-words	**0.01370**	2.588×10^{-5}	2.316×10^{-4}	6.488×10^{-5}
CARTÃO	*AdjSim-int*	**4.0**	0.0	0.0	0.0
	AdjSim-cos	**0.05601**	0.0	0.0	0.0
	MinDist	**3.43**	3.79	4.79	4.43
	PR-words	**0.00893**	4.448×10^{-5}	2.006×10^{-4}	4.182×10^{-5}
CARTÃO+	*AdjSim-int*	**10.0**	0.0	0.0	1.0
	AdjSim-cos	**0.06389**	0.0	0.0	0.00892
	MinDist	**3.57**	4.0	4.93	4.64
	PR-words	**0.00533**	3.410×10^{-5}	8.899×10^{-5}	1.724×10^{-4}
Onto.PT 0.6	*AdLesk*	5.0	5.0	0.0	0.0
	PR-syn	**0.00493**	2.953×10^{-4}	0.00167	0.00104
	PR-sen	**0.00198**	3.710×10^{-5}	5.219×10^{-4}	1.690×10^{-4}
OpenWN.PT	*AdLesk*	**1.0**	0.0	0.0	0.0
	PR-syn	**0.01727**	2.567×10^{-5}	3.621×10^{-5}	4.500×10^{-5}
	PR-sen	**0.00904**	1.728×10^{-5}	2.005×10^{-5}	2.119×10^{-5}

Para o ingresso na carreira de técnico superior é preciso ter uma licenciatura e a _____ efectua-se mediante a realização de um concurso e de um estágio com a duração de um ano.

(In order to enter a higher technical career, you must have a degree and the _____ is carried out upon the conduction of a competition and a year-long internship.)

(a) admissão *(admission)*
(b) parceria *(partnership)*
(c) prioridade *(priority)*
(d) ambiguidade *(ambiguity)*

Table 4: Ranking of alternatives for a question correctly by all configurations.

Resource	Algorithm	(a)	(b)	(c)	(d)
Baseline	*MostFrequent*	9,849	25,547	23,850	**54,691**
TeP 2.0	*AdjSim-int*	1.0	1.0	**7.0**	3.0
	AdjSim-cos	0.00890	0.00866	**0.05765**	0.03082
	MinDist	5.778	5.778	5.556	**5.334**
	PR-words	2.940×10^{-4}	1.672×10^{-4}	$\mathbf{7.176 \times 10^{-4}}$	2.646×10^{-4}
PAPEL 3.5	*AdjSim-int*	0.0	0.0	**1.0**	0.0
	AdjSim-cos	0.0	0.0	**0.01077**	0.0
	MinDist	5.1	4.0	**3.2**	4.0
	PR-words	4.706×10^{-5}	1.820×10^{-4}	$\mathbf{8.702 \times 10^{-4}}$	2.962×10^{-4}
CARTÃO	*AdjSim-int*	0.0	0.0	**1.0**	0.0
	AdjSim-cos	0.0	0.0	**0.00727**	0.0
	MinDist	5.455	4.091	**3.455**	3.909
	PR-words	4.902×10^{-5}	1.993×10^{-4}	$\mathbf{5.607 \times 10^{-4}}$	2.311×10^{-4}
CARTÃO+	*AdjSim-int*	2.0	1.0	**10.0**	4.0
	AdjSim-cos	0.01238	0.00506	**0.043272**	0.02341
	MinDist	5.18	4.18	**3.64**	4.0
	PR-words	2.371×10^{-4}	2.064×10^{-4}	$\mathbf{6.493 \times 10^{-4}}$	2.438×10^{-4}
Onto.PT 0.6	*AdLesk*	0.0	0.0	0.0	0.0
	PR-syn	3.275×10^{-5}	1.689×10^{-4}	$\mathbf{2.618 \times 10^{-4}}$	1.689×10^{-4}
	PR-sen	1.328×10^{-5}	1.902×10^{-5}	$\mathbf{5.921 \times 10^{-5}}$	2.385×10^{-5}
OpenWN.PT	*AdLesk*	1.0	0.0	**2.0**	1.0
	PR-syn	1.298×10^{-4}	8.376×10^{-4}	5.665×10^{-4}	$\mathbf{9.047 \times 10^{-4}}$
	PR-sen	5.639×10^{-5}	1.950×10^{-4}	$\mathbf{2.363 \times 10^{-4}}$	2.159×10^{-4}

O Daniel é um jogador de xadrez brilhante, pois consegue _____ até dez jogadas para examinar as ramificações possíveis de uma partida.

(Daniel is a brilliant chess player, because he can _____ up to ten moves in order to explore the possible branches of a game.)

(a) antecipar *(anticipate)*
(b) limitar *(limit)*
(c) conduzir *(drive)*
(d) incluir *(include)*

Table 5: Ranking of alternatives for a question answered incorrectly by all configurations.

to what happens in evaluation campaigns, such as those organised for Portuguese, targeting different NLP topics (e.g. HAREM (Freitas et al., 2010), Págico (Mota et al., 2012)). Possible directions for this work include the improvement of the current algorithms, and the adaptation of others that suit this task, such as the WordNet-based algorithms for the answering the TOEFL questions (Siblini and Kosseim, 2013; Pilehvar et al., 2013). Focusing on the task and not in the LKBs, using n-gram frequency information, either from a large corpus or from a web search engine, is an alternative approach that could lead to potentially interesting results. Moreover, given that LKBs are limited to lexical information, the performance of this task would probably benefit from using additional sources of world knowledge.

Acknowledgements

9. References

Agirre, E., de Lacalle, O. L., and Soroa, A. (2013). Random walks for knowledge-based word sense disambiguation. *Computational Linguistics*, 40(1).

Banerjee, S. and Pedersen, T. (2002). An adapted Lesk algorithm for word sense disambiguation using WordNet. In *Proceedings of the 3rd International Conference on Computational Linguistics and Intelligent Text Processing (CICLing 2002)*, LNCS, pages 136–145, London, UK. Springer.

Banerjee, S., Bhaskar, P., Pakray, P., Bandyopadhyay, S., and Gelbukh, A. (2013). Multiple choice question (mcq) answering system for entrance examination. In *Proceedings of the Question Answering for Machine Reading Evaluation (QA4MRE) at CLEF 2013 Conference and Labs of the Evaluation Forum*, Valencia, Spain.

Baptista, J., Costa, N., Guerra, J., Zampieri, M., Cabral, M., and Mamede, N. J. (2010). P-AWL: Academic Word List for Portuguese. In *Proceedings of Computa-*

tional Processing of the Portuguese Language, 9th International Conference, PROPOR 2010, volume 6001 of *LNCS*, pages 120–123. Springer.

Brank, J., Grobelnik, M., and Mladenic', D. (2005). A survey of ontology evaluation techniques. In *Proceedings of Conference on Data Mining and Data Warehouses*, SiKDD 2005, pages 166–170.

Bullinaria, J. A. and Levy, J. P. (2012). Extracting semantic representations from word co-occurrence statistics: stop-lists, stemming, and SVD. *Behavior Research Methods*, 44(3):890–907.

Correia, R., Baptista, J., Mamede, N., Trancoso, I., and Eskenazi, M. (2010). Automatic generation of cloze question distractors. In *Second Language Studies: Acquisition, Learning, Education and Technology*, Tokyo, Japan, September.

Correia, R., Baptista, J., Eskenazi, M., and Mamede, N. (2012). Automatic generation of cloze question stems. In *Proceedings of the 10th International Conference on Computational Processing of the Portuguese Language (PROPOR 2012)*, volume 7243 of *LNCS*, pages 168–178, Coimbra, Portugal, April. Springer.

de Paiva, V., Rademaker, A., and de Melo, G. (2012). Openwordnet-pt: An open brazilian wordnet for reasoning. In *Proceedings of the 24th International Conference on Computational Linguistics*, COLING (Demo Paper).

Fellbaum, C., editor. (1998). *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press.

Freitas, C., Carvalho, P., Oliveira, H. G., Mota, C., and Santos, D. (2010). Second HAREM: advancing the state of the art of named entity recognition in Portuguese. In *Proceedings of the 7th International Conference on Language Resources and Evaluation*, LREC 2010, La Valleta, Malta, May. ELRA.

Gonçalo Oliveira, H. and Gomes, P. (2013). ECO and Onto.PT: A flexible approach for creating a Portuguese wordnet automatically. *Language Resources and Evaluation*, to be published (online September 2013).

Gonçalo Oliveira, H. and Gomes, P. (2014). Onto.PT: recent developments of a large public domain portuguese wordnet. In *Proceedings of the 7th Global WordNet Conference*, GWC'14, pages 16–22.

Gonçalo Oliveira, H., Santos, D., Gomes, P., and Seco, N. (2008). PAPEL: A dictionary-based lexical ontology for Portuguese. In *Proceedings of Computational Processing of the Portuguese Language - 8th International Conference (PROPOR 2008)*, volume 5190 of *LNCS/LNAI*, pages 31–40, Aveiro, Portugal, September. Springer.

Gonçalo Oliveira, H., Antón Pérez, L., Costa, H., and Gomes, P. (2011). Uma rede léxico-semântica de grandes dimensões para o português, extraída a partir de dicionários electrónicos. *Linguamática*, 3(2):23–38, December.

Jarmasz, M. and Szpakowicz, S. (2003). Roget's thesaurus and semantic similarity. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2003)*, CILT, pages 212–219, Borovets, Bulgaria. John Benjamins, Amster-

dam/Philadelphia.

Landauer, T. K. and Dumais, S. T. (1997). A solution to Plato's problem: The Latent Semantic Analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review*, 104(2):211–240.

Maziero, E. G., Pardo, T. A. S., Felippo, A. D., and Dias-da-Silva, B. C. (2008). A Base de Dados Lexical e a Interface Web do TeP 2.0 - Thesaurus Eletrônico para o Português do Brasil. In *VI Workshop em Tecnologia da Informação e da Linguagem Humana (TIL)*, pages 390–392.

Mota, C., Simões, A., Freitas, C., Costa, L., and Santos, D. (2012). Págico: Evaluating Wikipedia-based information retrieval in Portuguese. In Calzolari, N., Choukri, K., Declerck, T., Do?an, M. U., Maegaard, B., Mariani, J., Odijk, J., and Piperidis, S., editors, *Proceedings of 8th International Conference on Language Resources and Evaluation*, LREC 2012, pages 2015–2022. ELRA.

Peñas, A., Miyao, Y., Hovy, E., Forner, P., and Kando, N. (2013). Overview of QA4MRE 2013 Entrance Exams task. In *CLEF 2013 Evaluation Labs and Workshop. Online Working Notes*. Pamela Forner and Roberto Navigli and Dan Tufis.

Pilehvar, M. T., Jurgens, D., and Navigli, R. (2013). Align, disambiguate and walk: A unified approach for measuring semantic similarity. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1341–1351, Sofia, Bulgaria, August. ACL.

Santos, D. and Bick, E. (2000). Providing Internet access to Portuguese corpora: the AC/DC project. In *Proceedings of 2nd International Conference on Language Resources and Evaluation*, LREC 2000, pages 205–210. ELRA.

Santos, D. and Rocha, P. (2001). Evaluating CETEMPúblico, a free resource for Portuguese. In *Proceedings of 39th Annual Meeting of the Association for Computational Linguistics*, ACL 2001, pages 442–449. ACL Press, 9-11 July.

Siblini, R. and Kosseim, L. (2013). Using a weighted semantic network for lexical semantic relatedness. In *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013*, pages 610–618, Hissar, Bulgaria, September. INCOMA Ltd. Shoumen, BULGARIA.

Silva, A., Marques, C., Baptista, J., Ferreira, A., and Mamede, N. (2012). REAP.PT serious games for learning portuguese. In *Proceedings of the 10th International Conference on Computational Processing of the Portuguese Language (PROPOR 2012)*, volume 7243 of *LNCS*, pages 248–259, Coimbra, Portugal, April. Springer.

Turney, P. D., Littman, M. L., Bigham, J., and Shnayder, V. (2003). Combining independent modules to solve multiple-choice synonym and analogy problems. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2003)*, CILT, pages 482–489, Borovets, Bulgaria. John Benjamins, Amsterdam/Philadelphia.