Gradually Improving the Computation of Semantic Textual Similarity in Portuguese

Hugo Gonçalo Oliveira[1(✉)], Ana Oliveira Alves[1,2(✉)], and Ricardo Rodrigues[1,3]

[1] CISUC, DEI, University of Coimbra, Coimbra, Portugal
{hroliv,ana,rmanuel}@dei.uc.pt
[2] ISEC, Polytechnic Institute of Coimbra, Coimbra, Portugal
[3] ESEC, Polytechnic Institute of Coimbra, Coimbra, Portugal

Abstract. There is much research on Semantic Textual Similarity (STS) in English, specially since its inclusion in the SemEval evaluations. For other languages, it is not as common, mostly due to the unavailability of benchmarks. Recently, the ASSIN shared task targeted STS in Portuguese and released training and test collections. This paper describes an incremental approach to ASSIN, where the computed similarity is gradually improved by exploiting different features (e.g., token overlap, semantic relations, chunks, and negation) and approaches. The best reported results, obtained with a supervised approach, would get second place overall in ASSIN.

Keywords: Natural language processing · Semantic Textual Similarity · Semantic relations · Supervised machine learning

1 Introduction

Computing the similarity of words or sentences in terms of their meaning is an active area of research in natural language processing (NLP) and understanding. This is confirmed by the shared tasks organised on STS, such as SemEval STS [1,2], which required the manual compilation of annotated data for benchmarking this specific task. Briefly, STS aims at computing a score for the semantic similarity between two sentences. Most successful approaches for English learn a similarity function that combines different metrics, such as word or chunk overlap, semantic relations, or distributional similarity.

Since their beginning, STS tasks have targeted English and had an increasing number of participants. For instance, the 2016 edition had more than 100 runs submitted by a total of 43 teams. We can therefore say that, for English, STS is becoming a mature task. On the other hand, excluding Spanish, included in recent editions of SemEval, this task is in its earliest days for other languages, including Portuguese. In fact, until recently, there was not a public dataset for computing semantic similarity between Portuguese sentences. Last year, this started to change, after the ASSIN shared evaluation [3] and the release of a dataset of STS in Portuguese.

E. Oliveira et al. (Eds.): EPIA 2017, LNAI 10423, pp. 841–854, 2017.
DOI: 10.1007/978-3-319-65340-2_68

This paper presents a post-evaluation approach to ASSIN and the gradual improvement of the results as features and techniques are added. Most features are inspired by related work for English, but adapted for Portuguese. More than describing a winning approach, which it is not, we detail every single step towards improving the baselines that rely exclusively on the surface text.

The remainder of this paper starts with a brief overview of related work, with the best results of English STS together with commonly used features. We address this task for Portuguese, focusing on ASSIN, its collections and best approaches. The pre-processing tools used here are then described, and results based on set similarity measures presented. After this, information in Portuguese LKBs is considered to compute similarity without supervision. Before concluding, the best supervised measures are combined with lexical and syntactic features to learn a similarity function, this time with supervision, with different regression algorithms and training datasets. Though some features were left unexplored, the final results are very close to the best position overall in ASSIN.

2 Related Work

The SemEval shared evaluations include STS tasks since 2012 [1,2]. Results are typically assessed by the correlation (e.g., Pearson, hereafter ρ, between -1 and 1—the higher, the better) and the Mean Squared Error (MSE—the lower, the better) between the values computed by the system and those given by several human judges, for the same collection of pairs.

Most successful approaches are supervised. To learn a similarity function, they combine different features, some of which as basic as token or n-gram overlap, but also similarity measures computed in WordNet [4] or other semantic networks, topic models, or distributional similarity models. In recent editions, deep semantic models have been successfully used, including by the best systems in SemEval 2016 [5,6]. Best ρ has ranged from 0.618, in SemEval 2013, to 0.824, in SemEval 2012. For the adopted baseline—the cosine of the vectors that represent the words in each sentence of the pair—this number has ranged from 0.311, in 2012, to 0.587, in 2015. For Spanish STS, the best system [7] in SemEval 2015 achieved $\rho = 0.69$, also with an approach that combined string similarity, semantic similarity and alignment features.

Another related task, in SemEval 2014, was the *Evaluation of Compositional Distributional Semantic Models on Full Sentences through Semantic Relatedness* [8]. The main difference is that it evaluates compositionality. It is thus more focused on issues like lexical variation, syntactic alternations, negation, and does not require dealing with multiword expressions or accessing world knowledge, including named entities. Yet, the system with the best results [9] tackled it as STS and achieved $\rho = 0.827$ and $MSE = 0.325$. Another system that participated in both the previous tasks, though with less success, was ASAP [10,11].

An earlier approach to STS in Portuguese [12] exploited a knowledge base to identify related words in different sentences. The proposed measure was tested in natural language descriptions of bugs in software engineering projects, which had their similarity annotated by two human judges. But it was not until 2015 that Portuguese NLP researchers could tackle STS and compare it with other approaches, using common benchmarks. The ASSIN shared task [3] targeted precisely Semantic Similarity and Textual Entailment in Portuguese. Announced in late 2015, the evaluation took 6 days, starting on 27 February 2016. Training data comprised 3,000 sentence pairs for Brazilian Portuguese (PT-BR) and another 3,000 for European Portuguese (PT-PT). Test data comprised 2,000 PT-BR pairs and 2,000 PT-PT pairs. Both collections are available in XML, together with the evaluation tools[1]. While recent editions of English STS have used text from varied sources, including news, glosses, forum posts, forum answers or image descriptions, sentences in the ASSIN collections were obtained exclusively from Google News. Table 1 shows three pairs in the ASSIN training collections, including their ids, their two sentences (t, for text, and h, for hypothesis), and the average similarity given by four human judges that followed the same guidelines. Similarity values range from 1 (completely different sentences, on different subjects) to 5 (sentences mean essentially the same).

Table 1. Examples from the training collections.

Collection	Id	Pair	Sim
PT-PT	2675	**t:** *O Chelsea só conseguiu reagir no final da primeira parte.*	1.25
		h: *Não podemos aceitar outra primeira parte como essa.*	
PT-PT	315	**t:** *Todos que ficaram feridos e os mortos foram levados ao hospital.*	3.0
		h: *Além disso, mais de 180 pessoas ficaram feridas.*	
PT-BR	1282	**t:** *As multas previstas nos contratos podem atingir, juntas, 23 milhões de reais.*	5.0
		h: *Somadas, as multas previstas nos contratos podem chegar a R$ 23 milhões.*	

ASSIN had 6 participating teams, which submitted 14 runs for the similarity task in PT-BR and 17 in PT-PT. The best official runs were obtained by different systems for PT-PT and PT-BR. For PT-BR, the best run [13] achieved $\rho = 0.70$ with $MSE = 0.38$, obtained by computing the cosine similarity of a vector representation of each sentence, based on the sum of the TF-IDF scores and word2vec [14] vectors of each word. For PT-PT, the best run [15] achieved $\rho = 0.73$ with $MSE = 0.61$, obtained after learning a similarity function with a Kernel Ridge Regression that used several similarity metrics as input, computed between the two sentences of each pair, including overlap and set similarity measures on multiple text representations (lowercase, character trigrams,...). An adaptation to Portuguese of ASAP, dubbed ASAPP [16], also participated

[1] http://nilc.icmc.usp.br/assin/.

in ASSIN, with best runs achieving $\rho = 0.65$ and $MSE = 0.44$, for PT-BRE, and $\rho = 0.68$ and $MSE = 0.70$ for PT-PT. This work revisits ASAPP.

3 Sentences Pre-processing

Our approach for computing the similarity of the pairs in the ASSIN collections starts with a pre-processing step, where sentences are split by tokens, part-of-speech (POS) tagged and lemmatized. Tokenization and POS tagging were performed by a set of tools based on the Apache OpenNLP Toolkit[2], targeting Portuguese. Besides basic tokenization operations, the tokenizer separates contractions, handles clitics and normalizes some abbreviations. Identified tokens are used by the basic OpenNLP Portuguese POS tagger, trained with a Maximum Entropy model. The lemmatizer converts words to their dictionary form, considering the output of the POS tagger, a morphology lexicon and a set of handcrafted rules [17]. Chunks and named entities were also identified with OpenNLP-based tools. These tools are freely available[3].

Alternatively to lemmatization, tokens can be stemmed with the PTStemmer tool, also freely available[4]. The main difference is that a lemma is still a word, though in the masculine singular form, if a noun or an adjective (e.g., *jogo* for *jogos*), or in the infinitive, if a verb (e.g., *jogar* for *joga* or *jogaram*), while stems are just the roots of the words (e.g., *jog* for *jogos*, *joga* and *jogaram*).

4 Sentence Similarity as Set Similarity

After pre-processing, simple computations were performed for each pair. Each sentence was considered as a set of words—T and H—and their overlap was computed with measures typically used for set similarity, including Jaccard, Overlap and Dice, respectively in Eqs. 1, 2, and 3. The cosine of the word vectors built from the words in the sentences was also computed as a set similarity measure (see Eq. 4).

$$Jacc(T,H) = \frac{|T \cap H|}{|T \cup H|} \qquad (1)$$

$$Dice(T,H) = \frac{|T \cap H|}{|T| + |H|} \qquad (3)$$

$$Ovl(T,H) = \frac{|T \cap H|}{|min(T,H)|} \qquad (2)$$

$$Cos(T,H) = \frac{|T \cap H|}{\sqrt{|T|}\sqrt{|H|}} \qquad (4)$$

In order to select basic parameters and set our baselines, the previous measures were computed for each pair of sentences and combinations of the following:

- Normalization: tokens could be matched exactly as they occur, after lemmatization, or after stemming;

[2] http://opennlp.apache.org/.

[3] NLP tools available from https://github.com/rikarudo/.

[4] PTStemmer is available from https://code.google.com/archive/p/ptstemmer/.

Table 2. Set similarity results for PT-PT.

Closed	Lem	POS	Sim	ρ	MSE
×	×	×	Cos	0.622	0.850
✓	×	×	Ovl	0.640	0.529
✓	×	×	Cos	0.664	0.552
×	✓	×	Ovl	0.618	0.550
×	✓	×	Dice	0.653	0.621
✓	✓	×	Cos	**0.698**	**0.446**
×	×	✓	Ovl	0.577	0.775
×	×	✓	Cos	0.605	0.918
✓	×	✓	Ovl	0.624	0.564
✓	×	✓	Cos	0.649	0.618
×	✓	✓	Ovl	0.599	0.601
×	✓	✓	Cos	0.632	0.675
✓	✓	✓	Ovl	0.642	0.484
✓	✓	✓	Cos	0.675	0.544
✓	*stems*		Jacc	0.700	1.140
✓	*stems*		Dice	0.703	0.458
✓	*stems*		Cos	**0.706**	**0.443**

Table 3. Set similarity results for PT-BR.

Closed	Lem	POS	Sim	ρ	MSE
×	×	×	Cos	0.539	0.718
✓	×	×	Ovl	0.574	0.583
✓	×	×	Cos	0.587	0.591
×	✓	×	Cos	0.560	0.598
✓	✓	×	Jacc	**0.610**	0.921
✓	✓	×	Cos	**0.610**	**0.484**
×	×	✓	Jacc	0.521	1.357
×	×	✓	Cos	0.519	0.784
✓	×	✓	Jacc	0.571	1.049
✓	×	✓	Ovl	0.557	0.545
×	✓	✓	Dice	0.538	0.648
✓	✓	✓	Cos	0.585	0.523
✓	*stems*		Jacc	0.625	0.853
✓	*stems*		Dice	0.609	0.489
✓	*stems*		Cos	**0.626**	**0.467**

- All tokens (except punctuation signs) could be considered, or only open-class words (nouns, verbs, adjectives and adverbs);
- A match could require just the same token/lemma or also the same POS.

As all the measures used output values between 0 (sets have no common elements) to 1 (sets have exactly the same elements), they were normalized to the 1–5 interval, as in the ASSIN collection. Tables 2 and 3 show a selection of results of set similarity and different configurations in the training collections. For each combination of parameters, we show the best correlation ρ and MSE.

These experiments lead to relevant conclusions. First, they are higher than all the baselines considered in editions of the SemEval English STS, which suggests that this exercise might be easier. Better results are obtained when using all tokens and not only open-class words, even better when tokens are normalized. However, matching also the POS leads to worse results. This can be caused by noise in the POS tagger or because lemmatization differentiates the base form of words depending on their category (e.g., noun, verb, and adjective). Best results were obtained with stems instead of lemmas, *grouping* words of the same family, regardless of their POS. The best configuration for both training collections used the Cosine of the stem vectors of all words and reached $\rho = 0.706$ and $MSE = 0.443$ for PT-PT, and 0.626 and 0.467 for PT-BR. With lemmas, Cosine was also the best measure, but with lower ρ and MSE, respectively 0.698 and 0.446 for PT-PT and 0.610 and 0.484 for PT-BR.

5 Exploiting Known Semantic Relations

Language is flexible in such a way that the same idea can be transmitted through different words, generally related by well-known semantic relations. For instance, synonyms may be used to denote the same meaning (e.g., *big* and *large*) and hypernyms are generalisations of their hyponyms (e.g., *animal* and *dog*). Those relations are implicitly mentioned in dictionaries, and explicitly encoded in LKBs, such as WordNet [4].

5.1 Portuguese Lexical Knowledge Bases

As it seemed natural to explore semantic relations when computing semantic similarity, we decided to use a LKB for Portuguese. After analysing the landscape of alternatives, we extracted semantic relation instances of different types, represented as "*a* related-to *b*" from 9 LKBs, namely: WordNet.Br [18] (just verbs), OpenWordNet-PT (OWN.PT) [19] and PULO [20]—three wordnets; TeP [21] and OpenThesaurus.PT[5]—two synonymy-based thesauri; PAPEL [22] and relations from Dicionário Aberto [23] and Wiktionary.PT[6]—three lexical networks extracted from Portuguese dictionaries; and the semantic relations of Port4Nooj [24]—a set of linguistic resources.

The aforementioned LKBs have substantially different sizes and the creation of most involved some degree of automation, which means that they contain noise, including rarely used words and meanings, not so useful relations, and also actual errors. Therefore, we decided to rely on redundancy to build more reliable and useful semantic networks, namely *Redun2* and *Redun3*, which include all the the relation instances respectively in at least two or three of the exploited LKBs. Table 4 shows the number of distinct relation instances in all the LKBs ($\geq$1) and the types with more instances, side-by-side with the same number for *Redun2* ($\geq$2) and *Redun3* ($\geq$3), and two examples for each relation type. A comparison of the LKBs and the creation of the redundancy-based LKBs is described in detail elsewhere [25].

5.2 Combining Semantic Relations and Sentence Overlap

In order to consider semantic relations, set similarity was first computed and then adjusted, considering the semantic network distances of the lemmas of the non-overlapping tokens (T' and H'). The later was computed according to Eq. 5—for each lemma in $T'_i \in T'$ we used the maximum similarity with a lemma $H'_j \in H'$—and summed in a factor γ (Eq. 6). The set similarity measure was adjusted by

[5] http://paginas.fe.up.pt/~arocha/AED1/0607/trabalhos/thesaurus.txt.
[6] http://pt.wiktionary.org.

Table 4. Relations of the semantic networks and their redundancy.

Relation	≥1	≥2	≥3	Examples
Synonymy	327,316	94,525	29,750	*(realçar,sublinhar), (afronta,ofensa)*
Hypernymy	277,764	29,879	4,639	*(mover,tremer), (campo,prado)*
Causation	15,373	4,705	1,590	*(frio,crestar), (distinção,preferência)*
Property-of	52,365	6,934	877	*(oral,boca), (defeituoso,ter_defeito)*
Antonymy	50,171	1,727	470	*(tristeza,alegria), (próximo,distante)*
Part-of	24,656	2,036	153	*(núcleo,átomo), (mês,ano)*
Purpose-of	17,348	1,410	134	*(polir,lixa), (traçar,compasso)*
Other	29,590	3,130	298	–
Total	794,583	144,346	37,983	

adding γ to the numerator, which, for all measures used, was the intersection of the tokens/lemmas/stems. Equation 7 shows how Jaccard becomes Jaccard$^+$. The other measures were analogously adapted. Considering that the maximum similarity (1, before normalization) is achieved when the tokens/lemmas/stems are the same, for each lemma in T', similarity had to be lower than 1, therefore $0 \leq \gamma \leq |T'|$. After several experiments in the training collections, we empirically set the values of the similarity equation to $\alpha = 0.75$ and $\beta = 0.05$.

$$Sim(T'_i, H'_j) = \begin{cases} \alpha, & \text{if } distance(T'_i, H'_j) = 1 \\ \beta, & \text{if } distance(T'_i, H'_j) = 2 \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

$$\gamma = \sum_{i=1}^{|T'|} \sum_{j=1}^{|H'|} Max(Sim(T'_i, H'_j)) \tag{6}$$

$$Jacc^+(T, H) = \frac{|T \cap H| + \gamma}{|T \cup H|} \tag{7}$$

The adapted measures were computed with each semantic network, individually, and with *Redun2* and *Redun3*. Although some differences were insignificant, *Redun3* lead more consistently to the best results. Tables 5 and 6 present, respectively for the PT-PT and for the PT-BR training collections, a selection of the best results obtained with different similarity measures and a different kind of normalization using all the relations of *Redun3*. The lower part of the tables shows a selection of the best results when using only synonymy and hypernymy relations, *Redun2*, and when using each LKB individually.

Cosine was again the best measure. Jaccard was very close in terms of correlation, but with considerably higher MSE. Moreover, although we would think that synonymy and hypernymy relations would fit this task better, using all relation types shown to be a better choice. Although lower than expected, the baseline results were improved. For PT-PT, correlation went from 0.706 to 0.721 and for PT-BR it went from 0.626 to 0.632.

Table 5. Results for PT-PT training pairs when semantic networks are exploited.

Norm	LKB	Rels	Sim	ρ	MSE
Lem	*Redun3*	All	Jacc^{+}	0.709	1.116
Lem	*Redun3*	All	Dice^{+}	0.708	0.448
Lem	*Redun3*	All	Cos^{+}	0.712	0.431
Stem	*Redun3*	All	Jacc^{+}	0.717	1.049
Stem	*Redun3*	All	Dice^{+}	0.717	0.419
Stem	*Redun3*	All	Cos^{+}	**0.721**	0.388
Stem	*Redun3*	Syn+Hyp	Jacc^{+}	0.713	1.007
Stem	*Redun3*	Syn+Hyp	Cosine^{+}	0.717	0.394
Stem	*Redun2*	All	Jacc^{+}	0.715	1.034
Stem	*Redun2*	All	Cos^{+}	0.712	0.404
Stem	Wikt.PT	All	Cos^{+}	0.720	0.388
Stem	PAPEL	All	Cos^{+}	0.719	**0.383**

Table 6. Results for PT-BR training pairs when semantic networks are exploited.

Norm	LKB	Rels	Sim	ρ	MSE
Lem	*Redun3*	All	Jacc^{+}	0.621	0.843
Lem	*Redun3*	All	Dice^{+}	0.617	0.466
Lem	*Redun3*	All	Cos^{+}	0.62	0.464
Stem	*Redun3*	All	Jacc^{+}	**0.632**	0.778
Stem	*Redun3*	All	Dice^{+}	0.628	0.456
Stem	*Redun3*	All	Cos^{+}	0.631	**0.453**
Stem	*Redun3*	Syn+Hyp	Jacc^{+}	0.631	0.787
Stem	*Redun3*	Syn	Cos^{+}	0.631	**0.453**
Stem	*Redun2*	All	Jacc^{+}	**0.632**	0.727
Stem	*Redun2*	All	Cos^{+}	0.624	0.466
Stem	TeP	Syn	Jacc^{+}	0.629	0.771
Stem	OT	Syn	Cos^{+}	0.628	0.458

5.3 Unsupervised Test Results

From the previous experiments on the training collections, we selected three configurations, based on the best combined results and used them to compute the similarity of the pairs in the ASSIN test collections. Table 7 shows the obtained results for PT-PT and PT-BR and the results of the best training baseline in the test collection.

Table 7. Test results when semantic networks are exploited, plus the Cosine baseline.

Normalization	Network	Relations	Measure	PT-PT		PT-PBR	
				ρ	MSE	ρ	MSE
Stem	*Redun2*	All	Jacc^{+}	0.669	0.746	0.669	0.764
Stem	*Redun3*	All	Jacc^{+}	0.669	0.723	0.666	0.825
Stem	*Redun3*	All	Cos^{+}	**0.677**	**0.686**	**0.667**	**0.454**
(baseline) Stem	–	–	Cos	0.656	0.658	0.653	**0.445**

In PT-PT, performance was lower than for the training, but in PT-BR it was higher. For both of them, the highest ρ and lowest MSE are obtained with the Cos^{+} and *Redun3*. In the ASSIN evaluation, this would be the fifth best run and fourth best system for PT-PT, in terms of ρ, and the third best run and system in terms of MSE. For PT-BR, it would get the third best ρ, second best system, and seventh best MSE, fourth best system. For an unsupervised approach, we see these results as very interesting, and would be ranked first considering only unsupervised approaches to ASSIN. Towards better results, some of these measures were combined with additional features in different learning methods.

6 Learning a Similarity Model

To improve the previous results, some of the unsupervised measures were used together with a set of additional features to learn a similarity function from each training collection and, later, from both. Here, we enumerate the features used, describe the learning algorithms explored, and report on the training and test results achieved.

6.1 Features

From the previous experiments, we selected the best baselines—Jaccard and Cosine—both with lemmas and stems, as well as two of the best configurations of the unsupervised approach—Jaccard$^+$ and Cosine$^+$—again with lemmas and stems. The previous were considered together with the following additional features, some of which also exploited in related work:

- **Lexical features:** number of negation words (*não*, *nada*, *nenhum*, *de modo algum*,...) in each sentence of the pair and their absolute difference; number of common tokens; number of common lemmas; and number of common stems.
- **Syntactic features:** number of noun, verb and prepositional chunks in each sentence of the pair and their absolute difference.
- **Semantic features:** number of named entities of each type (abstraction, product, event, number, organization, person, place, thing and time) in each sentence of the pair and their absolute difference; number of semantic relations of four types (synonymy, hypernymy, antonymy and other) between lemmas in one sentence of the pair and lemmas in the other, for each LKB.

6.2 Learning Algorithms

Three different regression algorithms, provided by the Weka [26] machine learning toolkit, were used to learn the similarity function. Table 8 presents the setup of the three best-peforming algorithms, after an exhaustive set of runs, namely:

- *M5Rules* [27] generates a decision list for regression problems using a separate-and-conquer strategy. In each iteration, it builds a model tree using the M5 algorithm and turns the "best" leaf into a rule.
- *Random Subspace* [28] is an ensemble learning algorithm that builds a decision tree classifier. It consists of random subspacing regression ensembles composed of multiple trees constructed systematically by pseudo-randomly selected subsets of components of the feature vector.
- Regression algorithm to infer the similarity function based on *Gaussian Processes* [29], in this case with a Radial Basis Function (RBF) Kernel as the Gaussian function. This implementation is simplified in Weka: it does not apply hyper-parameter-tuning and uses normalization to the target class (similarity value), so the features simplify the choice of a noise level.

Table 8. Weka setup for the three learning algorithms used.

Algorithm	Weka setup
M5Rules	`weka.classifiers.rules.M5Rules -M 4.0`
RandomSubspace w/M5	`weka.classifiers.meta.RandomSubSpace -P 0.5 -S 1 -num-slots 1 -I 10 -W weka.classifiers.trees.M5P -- -M 4.0`
Gaussian Process w/RBF Kernel	`weka.classifiers.functions.GaussianProcesses -L 1.0 -N 0 -K"weka.classifiers.functions.supportVector.RBFKernel -G 0.01 -C 250007"`

6.3 Training and Testing

Table 9 shows the average training performance for the three learning models in a 10-fold cross validation, for PT-PT and PT-BR individually. They are clearly higher than the unsupervised results in the training collections.

Table 10 shows the results of the learned models in the ASSIN test collections. Though using the same algorithm, different models were used for PT-PT and PT-BR, respectively trained in the PT-PT and on the PT-BR training collection. When compared to our unsupervised approach, there is a clear improvement—ρ is 0.032 points higher for PT-PT and 0.019 for PT-BR—but results are still below the best official ASSIN results. They would get second position in both collections, considering both ρ and MSE.

Table 9. Performance when training in the PT-PT and PT-BR collections.

Method	PT-PT		PT-BR	
	ρ	MSE	ρ	MSE
M5Rules	0.742	0.472	0.657	0.518
RandomSubspace	**0.756**	**0.457**	**0.662**	**0.515**
GaussianProcess	0.739	0.479	0.658	0.520

Table 10. Test results for models trained in the respective training collection.

Method	PT-PT		PT-BR	
	ρ	MSE	ρ	MSE
M5Rules	0.703	0.714	0.678	0.411
RandomSubspace	**0.709**	**0.698**	**0.686**	**0.403**
GaussianProcess	0.694	0.725	0.683	0.406

6.4 Training on both Collections

As a complementary experiment, we adopted a different training strategy. Since they are just variants of the same language, instead of training independent models for PT-PT and PT-BR, we concatenated the training collections and learned new (variant-ignoring) models from the resulting larger collection, which comprised 6,000 pairs. Tables 11 and 12 show, respectively, the training performance of the same learning algorithms on a 10-fold cross-validation in the larger collection, and the results of the new models in each test collection.

Despite the lower training performance, when using the Random Subspace ensemble, there are minor improvements in ρ, both for PT-PT and PT-BR, which is enough to match the correlation of the best official run for PT-BR — although the official results reported 0.70 correlation, they were rounded to two

Table 11. Training performance in a collection with both PT-PT and PT-BR training pairs.

Method	ρ	MSE
M5Rules	0.705	0.493
RandomSubspace	0.713	0.486
GaussianProcess	0.701	0.493

Table 12. Test results for models trained with both PT-PT and PT-BR training pairs.

Method	PT-PT		PT-BR	
	ρ	MSE	ρ	MSE
M5Rules	0.702	0.648	0.690	0.505
RandomSubspace	**0.711**	**0.657**	**0.697**	**0.499**
GaussianProcess	0.691	0.678	0.684	0.509

decimal places. Regarding MSE, it is slightly better for PT-PT, but worst for PT-BR. Also, the lower differences between training and test suggest that the variant specific features were more clear in the training than in the test, but a deeper analysis would be needed for this claim.

7 Concluding Remarks

An incremental approach to the ASSIN shared task on Portuguese STS was described. Simple similarity measures based on the surface text were first used and gradually improved after exploiting Portuguese LKBs and combining different features in a supervised approach. The reported experiments should be seen as the development of ASAPP [16], an approach to ASSIN.

The best results achieved would get second place overall in ASSIN. To beat the best official results, more experiments would be needed and additional features explored. Our approach would probably benefit from distributional features based on a vector representation for words, such as word embeddings [14] for Portuguese (e.g., [30]), as well as topic distributions (e.g., using LDA [31]) and n-gram overlap. Adding those features will be the next steps of this work, which will also explore feature reduction methods.

Acknowledgements. This work was financed by the ERDF European Regional Development Fund through the COMPETE Programme (operational programme for competitiveness) and by National Funds through the FCT Fundação para a Ciência e a Tecnologia (Portuguese Foundation for Science and Technology) within project REMINDS – UTAP-ICDT/EEI-CTP/0022/2014.

References

1. Agirre, E., Banea, C., Cer, D., Diab, M., Gonzalez-Agirre, A., Mihalcea, R., Rigau, G., Wiebe, J.: Semeval-2016 task 1: semantic textual similarity, monolingual and cross-lingual evaluation. In: Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016), San Diego, California, pp. 497–511. ACL Press, June 2016

2. Agirre, E., Diab, M., Cer, D., Gonzalez-Agirre, A.: Semeval-2012 task 6: a pilot on semantic textual similarity. In: Proceedings of the 1st Joint Conference on Lexical and Computational Semantics, vol. 1: Proceedings of the Main Conference and the Shared Task, and Proceedings of the Sixth International Workshop on Semantic Evaluation, vol. 2, pp. 385–393. ACL Press (2012)
3. Fonseca, E., Santos, L., Criscuolo, M., Aluísio, S.: Visão geral da avaliação de similaridade semântica e inferência textual. Linguamática **8**(2), 3–13 (2016)
4. Fellbaum, C. (ed.): WordNet: An Electronic Lexical Database (Language, Speech, and Communication). The MIT Press, Cambridge (1998)
5. Rychalska, B., Pakulska, K., Chodorowska, K., Walczak, W., Andruszkiewicz, P.: Samsung Poland NLP team at SemEval-2016 task 1: necessity for diversity; combining recursive autoencoders, wordnet and ensemble methods to measure semantic similarity. In: Proceedings of 10th International Workshop on Semantic Evaluation (SemEval-2016), San Diego, California, pp. 602–608. ACL Press, June 2016
6. Brychcín, T., Svoboda, L.: UWB at semeval-2016 task 1: semantic textual similarity using lexical, syntactic, and semantic information. In: Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016), San Diego, California, pp. 588–594. ACL Press, June 2016
7. Hänig, C., Remus, R., de la Puente, X.: ExB themis: extensive feature extraction from word alignments for semantic textual similarity. In: Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), Denver, Colorado, pp. 264–268. ACL Press, June 2015
8. Marelli, M., Bentivogli, L., Baroni, M., Bernardi, R., Menini, S., Zamparelli, R.: Semeval-2014 task 1: evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment. In: Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), Dublin, Ireland, pp. 1–8. ACL Press, August 2014
9. Zhao, J., Zhu, T., Lan, M.: ECNU: one stone two birds: ensemble of heterogenous measures for semantic relatedness and textual entailment. In: Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), Dublin, Ireland, pp. 271–277. ACL Press, August 2014
10. Alves, A., Ferrugento, A., Lourenço, M., Rodrigues, F.: ASAP: automatic semantic alignment for phrases. In: Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), Dublin, Ireland, pp. 104–108. ACL Press, August 2014
11. Alves, A., Simões, D., Gonçalo Oliveira, H., Ferrugento, A.: ASAP-II: from the alignment of phrases to textual similarity. In: Proceedings of 9th International Workshop on Semantic Evaluation (SemEval 2015), Denver, Colorado, pp. 184–189. ACL Press, June 2015
12. Pinheiro, V., Furtado, V., Albuquerque, A.: Semantic textual similarity of portuguese-language texts: an approach based on the semantic inferentialism model. In: Proceedings of the 11th Conference on the Computational Processing of the Portuguese Language, PROPOR 2014, São Carlos/SP, Brazil, pp. 183–188, 6–8 October 2014 (2014)
13. Hartmann, N.: Solo queue at ASSIN: combinando abordagens tradicionais e emergentes. Linguamática **8**(2), 59–64 (2016)
14. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. In: Proceedings of the Workshop track of the International Conference on Learning Representations (ICLR), Scottsdale, Arizona (2013)

15. Fialho, P., Marques, R., Martins, B., Coheur, L., Quaresma, P.: INESC-ID@ASSIN: medição de similaridade semântica e reconhecimento de inferência textual. Linguamática **8**(2), 33–42 (2016)
16. Alves, A., Gonçalo Oliveira, H., Rodrigues, R.: ASAPP: alinhamento semântico automático de palavras aplicado ao português. Linguamçtica **8**(2), 43–58 (2016)
17. Rodrigues, R., Gonçalo-Oliveira, H., Gomes, P.: LemPORT: a high-accuracy cross-platform lemmatizer for portuguese. In: Proceedings of the 3rd Symposium on Languages, Applications and Technologies (SLATE 2014), OASICS, Germany, Schloss Dagstuhl–Leibniz-Zentrum für Informatik, pp. 267–274. Dagstuhl Publishing, June 2014
18. Dias-da-Silva, B.C.: Wordnet.Br: an exercise of human language technology research. In: Proceedings of 3rd International WordNet Conference (GWC), GWC 2006, South Jeju Island, Korea, pp. 301–303, January 2006
19. Paiva, V., Rademaker, A., Melo, G.: OpenWordNet-PT: an open Brazilian wordnet for reasoning. In: Proceedings of 24th International Conference on Computational Linguistics, COLING (Demo Paper) (2012)
20. Simões, A., Guinovart, X.G.: Bootstrapping a Portuguese wordnet from Galician, Spanish and English wordnets. In: Navarro Mesa, J.L., Ortega, A., Teixeira, A., Hernández Pérez, E., Quintana Morales, P., Ravelo García, A., Guerra Moreno, I., Toledano, D.T. (eds.) IberSPEECH 2014. LNCS, vol. 8854, pp. 239–248. Springer, Cham (2014). doi:10.1007/978-3-319-13623-3_25
21. Maziero, E., Pardo, T., Felippo, A., Dias-da-Silva, B.: 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), pp. 390–392 (2008)
22. Gonçalo Oliveira, H., Santos, D., Gomes, P., Seco, N.: PAPEL: a dictionary-based lexical ontology for Portuguese. In: Teixeira, A., Lima, V.L.S., Oliveira, L.C., Quaresma, P. (eds.) PROPOR 2008. LNCS, vol. 5190, pp. 31–40. Springer, Heidelberg (2008). doi:10.1007/978-3-540-85980-2_4
23. Simões, A., Sanromán, Á.I., Almeida, J.J.: Dicionário-Aberto: a source of resources for the Portuguese language processing. In: Caseli, H., Villavicencio, A., Teixeira, A., Perdigão, F. (eds.) PROPOR 2012. LNCS, vol. 7243, pp. 121–127. Springer, Heidelberg (2012). doi:10.1007/978-3-642-28885-2_14
24. Barreiro, A.: Port4NooJ: an open source, ontology-driven portuguese linguistic system with applications in machine translation. In: Proceedings of the 2008 International NooJ Conference (NooJ 2008), Budapest, Hungary, Newcastle-upon-Tyne: Cambridge Scholars Publishing (2010)
25. Gonçalo Oliveira, H.: Comparing and combining Portuguese lexical-semantic knowledge bases. In: Proceedings of 6th Symposium on Languages, Applications and Technologies (SLATE 2017), OASICS, Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik. pp. 16: 1–16: 14 (2017)
26. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.: The WEKA data mining software: an update. SIGKDD Explor. **11**(1), 10–18 (2009)
27. Holmes, G., Hall, M., Prank, E.: Generating rule sets from model trees. In: Foo, N. (ed.) AI 1999. LNCS, vol. 1747, pp. 1–12. Springer, Heidelberg (1999). doi:10.1007/3-540-46695-9_1
28. Ho, T.K.: The random subspace method for constructing decision forests. IEEE Trans. Pattern Anal. Mach. Intell. **20**(8), 832–844 (1998)
29. Mackay, D.: Introduction to Gaussian processes. In: Bishop, C.M. (ed.) Neural Networks and Machine Learning. Springer, Berlin (1998)

30. Rodrigues, J., Branco, A., Neale, S., Silva, J.: LX-DSemVectors: distributional semantics models for Portuguese. In: Silva, J., Ribeiro, R., Quaresma, P., Adami, A., Branco, A. (eds.) PROPOR 2016. LNCS, vol. 9727, pp. 259–270. Springer, Cham (2016). doi:10.1007/978-3-319-41552-9_27
31. Blei, D., Ng, A., Jordan, M.: Latent dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003)