ISSN 1870-4069

# A Word Embedding Analysis towards Ontology Enrichment

Mikael Poetsch[1], Ulisses Brisolara Correa[1,2], Larissa Astrogildo de Freitas[1]

[1] Federal University of Pelotas (UFPel), Pelotas, RS, Brazil
{mpoetsch, ub.correa, larissa}@inf.ufpel.edu.br
[2] Sul-rio-grandense Federal Institute of Education, Science, and Technology (IFSul), Charqueadas, RS, Brazil
ulissescorrea@charqueadas.ifsul.edu.br

**Abstract.** Word Embedding is a set of language modeling and feature learning techniques in Natural Language Processing where words or phrases are mapped to vectors of real numbers. This approach could be used in many tasks of Natural Language Processing, such as Text Classification, Part-Of-Speech Tagging, Named Entity Recognition, Sentiment Analysis, and others. In this paper we created different Word Embedding models, using TripAdvisor's hotel reviews. The corpus was pre-processed, in order to reduce noise, and then submitted to four Word Embedding algorithms: Word2Vec, FastText, Wang2Vec, and GloVe. Finally, HOntology concepts and relations are compared with the outputs of models created aiming to improve it, enriching this domain ontology.

**Keywords:** word embedding, domain ontology, natural language processing.

## 1 Introduction

Word Embedding (WE) is a set of language modeling and feature learning techniques in Natural Language Processing (NLP) where words or phrases are mapped to vectors of real numbers [16]. These models create vector spaces where words that are semantically similar are mapped to nearby points.

Nowadays, one trend in NLP is to use vectors of words whose syntactic similarities correlate with semantic similarities. These vectors are used to calculate similarities between terms [10,19,23].

Bengio et al. (2003) [2] were among the first to introduce the term Word Embedding, where words or phrases are mapped to vectors of real numbers. The first work that shows the usefulness of WE pre-trained models in NLP tasks was proposed by [6].

The dissemination of WE can be attributed to [16], when the author created Word2Vec, a toolkit that allows the training of WE models, as well as make it easy to use pre-trained models.

Likewise other NLP tasks, research about WE applied to Portuguese texts is still in its first steps [23]. [10] made syntactic, semantic, intrinsic and extrinsic

analysis of different trained approaches of WE. Authors publicly share trained models in the repository http://nilc.icmc.usp.br/embeddings.

Another research in Portuguese is proposed by [19]. In [19] the intrinsic analysis using different trained models of WE with some tweets dataset were made.

Vargas and Pardo [23] used WE to aspect (explicit and implicit) identification and clustering that are critical tasks for sentiment analysis.

In this work, we created different trained models of WE, available in https://github.com/mikaellllll/Word-Embedding-Setor-Hoteleiro, using hotels reviews posted on the TripAdvisor with the intention of expanding the area of WE applied in the Portuguese. Finally, the models created are compared with a hotel ontology, with the aim of enriching it.

The remaining of this text is organized as follow. Section 2 discusses relevant related works. Next, Section 3 presents an overview of concepts involved in this work. In Section 4 we present the methodology used to assess WE performance, as well as the experimental results obtained. Finally, Section 5 summarizes our contributions and future work.

## 2 Related Works

Technical literature presents few related work, we found the following works about WE applied to Portuguese texts: [10] , [19], and [23].

Hartmann et al. (2007) [10] trained 31 WE models using the Word2Vec, FastText, Wang2Vec and GloVe algorithms. Their experiments were made varying the dimension of the model (using 50, 100, 300, 600 and 1000 inner dimensions) and using different corpora (mixed, encyclopedic genres, informative, didactic). The corpora used in his work are large. To compare the results, an intrinsic evaluation was performed on syntactic and semantic analogies.

The experiments of [10] shows that GloVe produced the best results for syntactic and semantic analysis, with 46.70% accuracy using 300 dimensions. Wang2Vec Skip-Gram produced the best results for POS-tagging, with 95.94% accuracy using 1000 dimensions.

Saleiro et al. (2017) [19] created WE models from tweets in Portuguese. First, authors begin with a relatively small sample and focus on three challenges, (i) the volume of training dataset, (ii) the size of the vocabulary, and (iii) the intrinsic metrics. The intrinsic metrics aims to establish a good combination between the number of dimensions and the size of the vocabulary. Through this work, the authors realized that producing WE from tweets is challenging due to the specificity of the vocabulary in the social media. Results show that using less than 50% of the available training examples for each vocabulary size might result in overfitting.

Automatic aspect identification and clustering are critical tasks for sentiment analysis. [23] presented a new approach to group explicit and implicit aspects from online reviews (about Books, Cameras, and Smartphones). The authors

achieved the 95.7% accuracy using a Skip-Gram based Word2Vec model, with 300 dimensions.

## 3 Theoretical Reference

This section describes a few key concepts to understand our proposed approach.

### 3.1 Corpus

Corpus is a set of collected text to be subject of linguistic research. This resource should be composed of texts, described in natural language by native speakers [20]. Corpora have been used as a key validation resource of solutions of NLP Tasks [8].

Developing a corpus demand to deal with several key factors, such as: definition of texts domain, definition of texts origin, definition of the amount of data to be collected, and how the corpus will be used (practical applications) [8]. Besides that, [8] highlights the importance of planning in corpus construction. Lack of planning could turn invalid experimental data produced using a malformed corpus.

As to size, according to [21], corpus can be classified as small, small-medium, medium, medium-large and large size. [21] created these classes based on research reported in Linguistics conferences and meetings. In this paper we use a medium-size corpus, with 656 thousand words.

### 3.2 Word Embedding

WE is a set of language modeling and feature learning techniques in Natural Language Processing (NLP) where words or phrases are mapped to vectors of real numbers [16]. These models create vector spaces where words that are semantically similar are mapped to nearby points and where some operations can extract logical results.

Mikolov et al. (2013) [16] present Word2Vec, that allow us to apply vector arithmetics to work with concepts. For instance, if we subtract the vector representation of 'man' from the vector representation of 'king', then we add the vector representation of 'women' the result will be near of the vector representation of 'queen' ($king - man + woman = queen$).

In this work we trained WE models using several different techniques (Word2Vec, FastText, Wang2Vec, and Glove), in order to evaluate which one can present better results in the task of Domain Ontology Enrichment.

**Word2Vec.** Word2Vec is a widely used method and have two model architectures for computing continuous word vectors using simple model. In the first model, called Continuous Bag-of-Words (CBOW), the non-linear hidden layer is removed and the projection layer is shared for all words; thus, all words get projected into the same position.

Whereas, in the Skip-Gram model the prediction of the current word is based on the context. The author use each current word as an input to a log-linear classifier with continuous projection layer and predict words within a certain range before and after the current word.

**FastText.** FastText is an approach based on the Skip-Gram model, this algorithm include representing sentences with bag-of-words and bag-of-n-grams, using sub-word information. The method attempts to capture morphological information to insert in WE model.

**Wang2Vec.** Wang2Vec is a modification of Word2Vec that intends to improve the WE obtained for syntactic tasks. The modification of CBOW is named Continuous Window (CWindow). The modification of Skip-Gram is named Structured Skip-Gram. CWindow and Structured Skip-Gram were made in order to make the network aware of the relative positioning of context words.

**GloVe.** The Global Vectors (GloVe) approach was proposed by [18]. It combines the advantages of the global matrix factorization and the local context window methods. This method model efficiently leverages statistical information by training only on the non-zero elements in a word to word co-occurrence matrix.

### 3.3 Ontology

According to [9], an ontology is a formal representation of knowledge based on conceptualization. This kind of resource encompasses a representation of the concepts, relations between concepts and instances.

There are many methodologies that helped in creating an ontology, such as: Enterprise [7] and OnToKnowledge [22]. In general, ontologies previously constructed could be adapted and expanded based in its first version. Ontologies are created and not discovered, so one domain can have different ontologies created by different methodologies.

In this work we evaluate WE models to enrich a domain ontology created with some of the methodologies aforementioned.

## 4 Experiments

In order to evaluate the applicability of WE models to domain ontologies enrichment we used a corpus to train a set of WE models and then manually analyzed models searching for relations between ontology's concepts absent in ontology.

The corpus used in our experiments is a set of hotel reviews from TripAdvisor [12]. Before the WE training all hotel reviews are pre-processed as follow.

First step was mapping emails to a token "EMAIL", mapping numbers to a token "0", mapping URLs to a token "URL", remove texts within square

brackets, and remove sentences with less than 5 tokens. This pre-processing approach was first proposed by [10].

Second step was removing punctuation of corpus and third step was correcting the spelling of the corpus through the LibreOffice's VERO library[3].

Fourth step was identifying n-grams using N-Gram Statistics Package [1]. N-gram is a sequence of words in a sentence [24]. For instance, the bi-grams of the sentence "O hotel é ótimo" ["The hotel is excellent"] are "# O" ["# The"], "O hotel" ["The hotel"], "hotel é" ["hotel is"], "é ótimo" ["is excellent"], "ótimo #" ["excellent #"], were # is empty. [26] states that use 4 or more grams can cause noise. For this reason, in this paper, we use n-grams with $n < 4$. Besides that, n-grams with less than 10 occurrences were discarded.

Fifth step was removing corpus stopwords. This was made using the Python library Natural Language Toolkit (NLTK)[4].

Sixth step was lemmatizing and identifying the grammatical category of words. We used SpaCy[5] to do so. For example, the lemma of "tem" ["has"] and "tinha" ["had"] is "ter" ["have"].

Sixth step was creating WE models. We used just the grammatical categories of verbs and nouns and its lemmas in the four methods (Word2Vec, FastText, Wang2Vec, and GloVe).

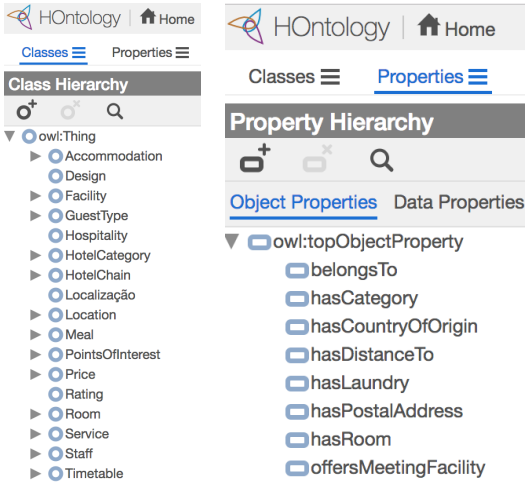
We used 300 dimensions, because according to [10] the increase in performance not reward the increase in memory. Besides that, as configuration we used 10 windows, because according to [17] a bigger number can lead to loss of performance.

To create CBOW and Skip-Gram of Word2Vec and FastText we used Gensim [25]. To create CWindow and Structured Skip-Gram of Wang2Vec we used the codes created by [11] and Glove we used the codes created by [18].

To visualize the WE models we used Bokeh[6] (Figures 2, 3, 4, and 5). More specifically the technique t-distributed Stochastic Neighbor Embedding (t-SNE). t-SNE was created by [15] to reduce the dimensions of the WE models to two dimensions facilitating the visualization of high-dimensional data. It is important to note that there are other techniques to decrease the dimensions, as "Principal Component Analysis"or "TruncatedSVD". We chose to use t-SNE because many works in literature has used this technique, such as: "Learning a Parametric Embedding by Preserving Local Structure [13]" and "Visualizing non-metric similarities in multiple maps [14]".

Next, we searched and visualized domain ontologies. We obtained the hotel domain ontology, HOntology, from the repository OntoLP[7]. To visualize the domain ontology we used Web Protégé[8]. As can be seen in Figure 1. Figure 1(a) presents HOntolgy classes and Figure 1(b) presents HOntolgy properties.

[3] https://pt-br.libreoffice.org/projetos/vero/
[4] https://www.nltk.org
[5] https://spacy.io/
[6] https://bokeh.pydata.org/en/latest
[7] http://ontolp.inf.pucrs.br/Recursos/downloads-Hontology.php
[8] https://webprotege.stanford.edu




(a) Classes (b) Properties

**Fig. 1.** HOntology.

HOntology is a multilingual ontology of the accommodation sector described in English, Portuguese, Spanish and French [5]. This ontology reuses concepts of another vocabulary as Dbpedia.org[9] and Schema.org[10]. At first, this resource was created to support managerial decision making and end-user applications. Some experts updated HOntology based on existing concepts in others ontologies, as can be seen in the work of [5].

HOntology was developed in seven phases: (1) identify existing ontologies on related domains, (2) select the main concepts and properties, (3) organize concepts and properties hierarchically, (4) manually translate concepts and properties, (5) expand concepts and properties based on online reviews manually evaluated, (6) manually translate the new concepts and properties,(7) export the ontology [5]. Our work could be applied to complement phase 5.

Currently, HOntology contains 282 concepts categorized into 16 top-level concepts, with maximum depth of concept hierarchy equals to 5 [5].

After visualize the WE models we compare them. Figure 2 shows the two variations of Word2Vec, CBOW and Skip-Gram. Glove model is depicted in Figure 3. Figure 4 presents the FastText models based in CBOW and Skip-Gram. In Figure 5, them Wang2Vec models (CWindow and Structured-Skip-Gram) can be seen. Each figure presents the whole structure of the corpus, following different WE models.

We perform two types of analysis. First, we analyze classes and subclasses with each WE model. The classes and subclasses were plotted to facilitate the identification in the clusters.

[9] https://wiki.dbpedia.org
[10] https://schema.org

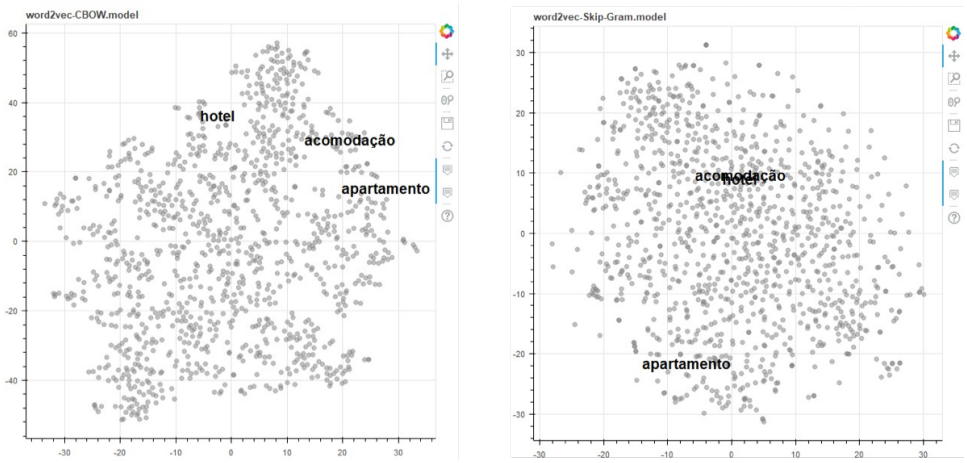



**Fig. 2.** Word2Vec models, "Acomodação"["Accommodation"] class and its subclasses "Apartamento" ["Apartment"] and "Hotel" ["Hotel"].

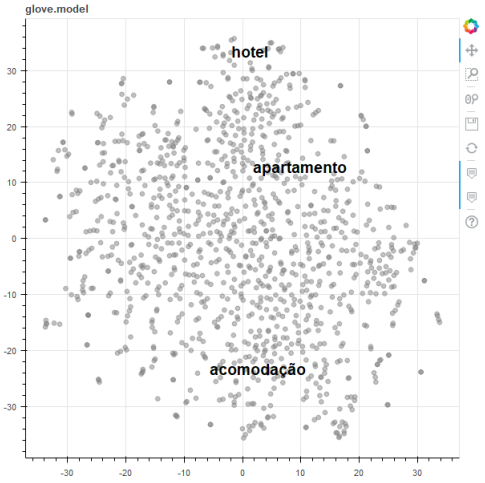



**Fig. 3.** Glove model, "Acomodação" ["Accommodation"] class and its subclasses "Apartamento" ["Apartment"] and "Hotel" ["Hotel"].

For instance, to the "Acomodação" ["Accommodation"] class some of its subclasses were not found in the models created. This may occur because the ontology is a formal representation of information while the reviews collected are informal texts, where formal terms are not common. Still, we observe that many subclasses were not found. For example: "Instalação do Banheiro" ["Bathroom Facility"], "Preço do Café da Manhã" ["Breakfast Price"] and "Serviço de Câmbio" ["Exchange Service"]. In general, these subclasses are related to

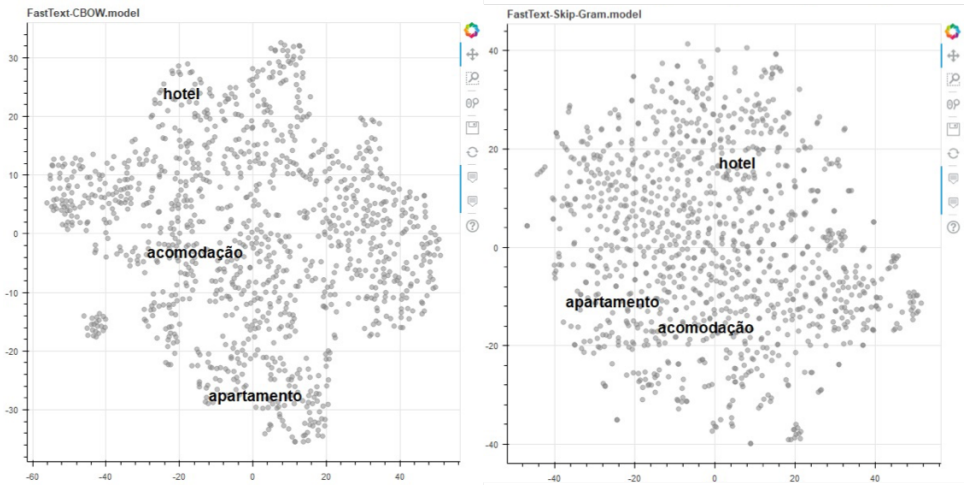



**Fig. 4.** FastText models, "Acomodação" ["Accommodation"] class and its subclasses "Apartamento" ["Apartment"] and "Hotel" ["Hotel"].

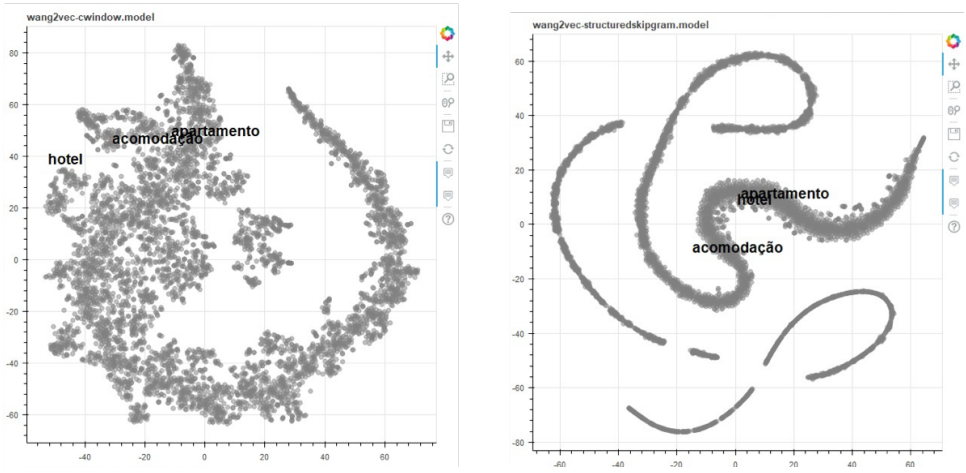



**Fig. 5.** Wang2Vec models, "Acomodação" ["Accommodation"] class and its subclasses "Apartamento" ["Apartment"] and "Hotel" ["Hotel"].

location or services offered by an accommodation. It is interesting to notice that the subclasses "Café da Manhã" ["Breakfest"] and "Jantar" ["Dinner"] of class "Refeição" ["Meal"] were found, but "Almoço" ["Lunch"] was not.

According to our results, "Acomodação" ["Accommodation"] class has no similarity to "Apartamento" ["Apartment"] subclass in Word2Vec, FastText and Glove due to low percentage of similarity and long distance between the words.

Wang2Vec and Word2Vec CBOW presents similarity bigger than 90% between "Acomodação" ["Accommodation"] class and "Hotel" ["Hotel"] subclass, as can be seen in TABLE 1,.

**Table 1.** Similarity between "Acomodação" ["Accommodation"] class and its subclasses.

| | **Subclasses** | |
|---|---|---|
| **Model** | **Apartment** | **Hotel** |
| Word2Vec CBOW | 36,7% | 92,3% |
| Word2Vec Skip Gram | 38,5% | 61,9% |
| FastText CBOW | 36,0% | 73,6% |
| FastText Skip Gram | 41,9% | 54,5% |
| Wang2Vec CWindow | 98,9% | 90,5% |
| Wang2Vec Structured Skip Gram | 99,9% | 99,9% |
| Glove | 30,8% | 24,8% |

Wang2Vec Structured-Skip-Gram presented the best similarity for both subclasses, "Apartamento" ["Apartment"] and "Hotel" ["Hotel"]. Observing the similarity between all classes and subclasses (depth 1) of HOntology, it was possible to notice that in the most of the models (Word2Vec, FastText, and Glove) there is no consensus on the degree of similarity between classes and subclasses, with except for the Wang2Vec. "Acomodação" ["Accommodation"] and "Refeição" ["Meal"] classes exist on all models. "Pontos de Interesse" ["Points of Interest"] exist only in FastText e Wang2Vec models. "Categoria de Hotel" ["Hotel Category"], "Endereço" ["Address"], "Serviço" ["Service"] and "Tipo de Hóspede" ["Guest Type"'] exist only on FastText model.

We also analyzed HOntology classes and words with each WE model, considering or not considering grammatical categories (verbs and nouns). In TABLE 2 we show words most similar to "Acomodação" ["Accommodation"] class.

In Word2Vec CBOW the word most similar to "Acomodação" ["Accommodation"] class was "Instalação" ["'Facility"], this word also is a class of HOntology. Thus, we could insert a new relation between these two classes.

In Wang2Vec CWindow one of the words most similar was the trigram "É limpo E" ["is clear and"]. The third column of the TABLE 2 should include verbs, which indicates an error in the POS-tagging, where "É limpo E" ["is clear and"] was classified as verb.

Finally, we analyze HOntology relations with each WE model. Among the HOntology relations we can mention: "pertenceA" ["belongsTo"], "temCategoria" ["hasCategory"], "temPaísDeOrigem" ["hasCountryOfOrigin"], "temDistânciaA" ["hasDistanceTo"], "temLavanderia" ["hasLaundry"], "temEndereçoPostal" ["hasPostalAddress"], "temQuarto" ["hasRoom"], and "ofereceLocalDeReunião" ["offersMeetingFacility"] (TABLE 3). For example, the relation "pertenceA" ["be-

longsTo"] represent that "Hotel" ["Hotel"] belongs to "Rede de Hotéis" ["Hotel Chain"].

**Table 2.** Words Most Similar to "Acomodação" ["Accommodation"] Class.

| Model | Word Most Similar | Word Most Similar Considering Grammatical Category |
|---|---|---|
| Word2Vec CBOW | Instalação | Adequar, Gostar |
| Word2Vec Skip Gram | Boutique | Reduzir, Surpreender |
| FastText CBOW | Fomos Muito Bem | Fomos Muito Bem, Muito A Desejar |
| FastText Skip Gram | Acomodar | Acomodar, Surpreender |
| Wang2Vec CWindow | Hotel É Muito | Praticar, **É Limpo E** |
| Wang2Vec Structured Skip Gram | É Uma Cidade | Cado, Significar |
| Glove | Instalação | Adequar, Treinar |

**Table 3.** Words that correspond to the relations of HOntology.

| Relation | Domain | Extension | Word |
|---|---|---|---|
| pertenceA | Hotel | Rede de Hotéis | pertencer, hotel, rede, hotel de rede, hotel da rede, hotéis da rede, da rede acor, da rede ibis |
| 'belongsTo' | 'Hotel' | 'Hotel Chain' | |
| temCategoria | Hotel | Categoria de Hotel | ter, categoria, hotel, padrão, turista |
| 'hasCategory' | 'Hotel' | 'Hotel Category' | |
| temPaísDeOrigem | Tipo de Hóspede | - | ter, país, origem, casal, família |
| 'hasCountryOfOrigin' | 'Gest Type' | | |
| temDistânciaA | Hospitalidade | Pontos de Interesse | ter, distância, distante do centro, hospitaleiro, hospitalidade, pouco distante do |
| 'hasDistanceTo' | 'Hospitality' | 'Points Of Interest' | |
| temLavanderia | Hospitalidade | Lavanderia | ter, lavanderia, hospitalidade, hospitaleiro |
| 'hasLaundry' | 'Hospitality' | 'Laundry' | |
| temEndereçoPostal | Hospitalidade | Localização | ter, endereçar, hospitalidade, hospitaleiro |
| 'hasPostalAddress' | 'Hospitality' | 'Location' | |
| temQuarto | Hotel, Hospitalidade | Quarto de Hotel, Hotel | ter, quarto, quarto de hotel, hotel, hospitalidade, hospitaleiro |
| 'hasRoom' | 'Hotel', 'Hospitality' | "Hotel Room', 'Hotel' | |
| ofereceLocalDeReunião | Hotel | Reunião | oferecer, hotel, reunião, reunir |
| 'offersMeetingFacility' | 'Hotel' | 'Meeting' | |

After defining the words for each relation defined in HOntology, we find the similarity between all words searched for each model. Wang2Vec presented a big similarity between words searched and FastText did not obtain any conclusive results to "pertenceA" [" belongsTo"].

Negative similarity means that the relation between words has opposite meanings. If two words have similarity equals to zero, they are unrelated. And,

if two words have similarity equals to -100%, they have a perfectly opposite relation.

## 5 Final Remarks and Future Work

In this work we created different models of WE using hotel reviews published on TripAdvisor in order to assess its applicability to the task of domain ontology enrichment.

After analyzing the HOntology and the different WE models created (Word2Vec CBOW and Skip-Gram, Glove, FastText CBOW and Skip-Gram, Wang2Vec CWindow and Structured-Skip-Gram), we realize that some concepts that are found in the reviews are not found in the domain ontology. Then, we conclude that it is possible to use WE models to enrich domain ontologies by incorporating new properties, such as a relation between "Accommodation " and "Installation".

Wang2Vec algorithm presented the best results for relations identification, accusing a great similarity between words related to an already existing relation in the HOntology. However, according to [4] the relevance of these relations must be evaluated before being inserted into the domain ontology to avoid an excessive increase of complexity in the representation.

As future works we intend to use other WE algorithms. Also, we intend to explore other levels of HOntology, such as: subclasses (depth 2, 3 and 4) and instances. And analyze how their results can be used to enrich domain ontologies. In addition, we could analyze the use of hyponymy, hyperonymy, and coreference.

## References

1. Banerjee, S., Pedersen, T.: The design, implementation, and use of the ngram statistics package. In: 4th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing), pp. 370–381 (2003)
2. Bengio, Y., Ducharme, R., Vincent, P., Jauvin, C.: A neural probabilistic language model. Journal of Machine Learning Research 3, 1137–1155 (2003)
3. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics 5(1), 135–146 (2016)
4. Carvalho, M.G.P., Braganholo, V., Machado Campos, M. L., de Almeida Campos, M.L.: Enriquecimento de ontologias: uma abordagem para extração de conhecimento do campo definição. In: 3th Seminário de Pesquisa em Ontologias no Brasil (Ontobras) (2010)
5. Chaves, M., Freitas, L., Vieira, R.: Hontology: a multilingual ontology for the accommodation sector in the tourism industry. In: 4th International Conference on Knowledge Engineering and Ontology Development, pp. 1–6 (2012)
6. Collobert, R., Weston, J.: A unified architecture for natural language processing: deep neural networks with multitask learning. In: 25th International Conference on Machine Learning (ICML), pp. 160–167 (2008)
7. Dietz, J.: Enterprise Ontology: Theory and Methodology. Springer-Verlag, Berlin, Heidelberg (2006)

8. Fromm, G.: O uso de corpora na análise linguística. Factus 1(1), 69–76 (2003)
9. Gruber, T. R.: Toward principles for the design of ontologies used for knowledge sharing? International Journal of Human-Computer Studies, 43(5-6), 907–928 (1995)
10. Hartmann, N., Fonseca, E. R., Shulby, C., Treviso, M. V., Silva, J., Aluísio, S.: Portuguese word embeddings: Evaluating on word analogies and natural language tasks. In: 11th Brazilian Symposium in Information and Human Language Technology (STIL), pp. 122–131 (2017)
11. Ling, W., Dyer, C., Black, A. W., Trancoso, I.: Two/too simple adaptations of word2vec for syntax problems. In: Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 1299–1304 (2015)
12. López Barbosa, R. R.: Aplicación del análisis de sentimientos a la evaluación de datos generados en medios sociales. Ph.D. thesis, Universidad de Alcalá (2015)
13. Maaten, L.: Learning a parametric embedding by preserving local structure. In: Artificial Intelligence and Statistics, pp. 384–391 (2009)
14. Van der Maaten, L., Hinton, G.: Visualizing non-metric similarities in multiple maps. Machine learning 87(1), 33–55 (2012)
15. van der Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research 9(Nov), 2579–2605 (2008)
16. Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in Neural Information Processing Systems, pp. 3111–3119 (2013)
17. Minarro-Giménez, J. A., Marín-Alonso, O., Samwald, M.: Applying deep learning techniques on medical corpora from the world wide web: a prototypical system and evaluation. CoRR, pp. 1–14 (2015)
18. Pennington, J., Socher, R., Manning, C.: Glove: Global vectors for word representation. In: Conference on empirical methods in natural language processing (EMNLP), pp. 1532–1543 (2014)
19. Saleiro, P., Sarmento, L., Mendes Rodrigues, E., Soares, C., Oliveira, E.: Learning word embeddings from the portuguese twitter stream: A study of some practical aspects. In: Portuguese Conference on Artificial Intelligence, pp. 880–891 (2017)
20. Sardinha, T. B.: Linguística de corpus: histórico e problemática. Delta 16(2), 323–367 (2000)
21. Sardinha, T. B.: Tamanho de corpus. The ESPecialist 23(2), 103–122 (2002)
22. Sure, Y., Staab, S., Studer, R.: On-to-knowledge methodology (otkm). In: Handbook on Ontologies, International Handbooks on Information Systems, pp. 117–132 (2003)
23. Vargas, F., Pardo, T.: Aspect clustering methods for sentiment analysis. In: 13th International Conference on the Computational Processing of Portuguese (PROPOR), pp. 365–374 (2018)
24. de Castro Sonnenfeld Vilela, P.: Classificação de Sentimento para Notícias sobre a Petrobras no Mercado Financeiro. Ph.D. thesis, PUC-Rio (2011)
25. Řehůřek, R., Sojka, P.: Software framework for topic modelling with large corpora. In: Workshop on New Challenges for NLP Frameworks, pp. 45–50 (2010)
26. Wiebe, J., Wilson, T., Bruce, R., Bell, M., Martin, M.: Learning subjective language. Computational linguistics 30(3), 277–308 (2004)