Spanish word vectors from Wikipedia

Mathias Etcheverry, Dina Wonsever
mathiase@fing.edu.uy, wonsever@fing.edu.uy
Universidad de la Republica
Uruguay

Abstract

Contents analisys from text data requires semantic representations that are difficult to obtain automatically, as they may require large handcrafted knowledge bases or manually annotated examples. Unsupervised autonomous methods for generating semantic representations are of greatest interest in face of huge volumes of text to be exploited in all kinds of applications. In this work we describe the generation and validation of semantic representations in the vector space paradigm for Spanish. The method used is GloVe (Pennington et al., 2014), one of the best performing reported methods , and vectors were trained over Spanish Wikipedia. The learned vectors evaluation is done in terms of word analogy and similarity tasks (Pennington et al., 2014; Baroni et al., 2014; Mikolov et al., 2013a). The vector set and a Spanish version for some widely used semantic relatedness tests are made publicly available.

Keywords: Distributional Semantics, Word Embeddings, Word Analogies

1. Introduction

Learning representations for words from their contexts provides us with new instruments to tackle natural language processing (NLP) tasks.

In many supervised NLP tasks the training data is limited and expensive but plain text is easily accessible thanks to the web. In opposition to NLP systems that considers words as atomic units, representing words as non-supervised learned vectors is a way to use unstructured data to improve results. These methods have a long history in Natural Language Processing, starting from Salton's proposal in the field of Information Retrieval, where a document collection is represented as a two-dimensional array, with the element $a_{i,j}$ indicating the number of times that the word i occurs in document j. The word-document model gave a bag-of-words characterization to the subject contents of documents, in a completely unsupervised way. As words that co-occur in the same document tend to belong to a same subject, this methodology led to the implementation of the distributional hypothesis of lexical similarity: words that occur in the same contexts have similar meanings. From the array with all words i.e., the vocabulary, as rows and documents as columns we now have to consider a *vocabulary x vocabulary* array. And from the idea that similar contexts imply similar meanings we are now facing the fact that from this word-word co-occurrence array we can compute a new representation for words (using a huge volume of text) that "summarizes" in some way its contexts (syntax and semantics included) and that has proved to be very effective in different NLP tasks.

In this paper we present the generation and evaluation of a Spanish word-vector resource, obtained from Spanish Wikipedia text, by means of the method employed in GloVe (Pennington et al., 2014).

In section 2 we present a general background, in section 3 we describe the GloVe method while in section 4 we explain the generation process. We have conducted a first evaluation of the generated resources (section 5) using some standard data sets for word similarity or relatedness that have been adapted to Spanish.

1.1. Word Vectors

Neural network models trained by back-propagation can represent knowledge on hidden units interactions (Rumelhart et al., 1986) and representation of concepts using distributed patterns of activity permits concept generalization (van der Maaten and Hinton, 1986; Pollack, 1990; Elman, 1991). With this in mind, neural approaches become useful for building representations of linguistic units. In NLP, neural networks can be used to build representations of language elements, for example, words. This kind of representation is called *word embedding*.

Different models of neural networks were considered. Feed-forward networks for language models were presented in (Bengio et al., 2003) and (Bengio et al., 2006). A deep network architecture for *multi-task learning* that includes tasks such as part-of-speech tagging, chunking, named entity recognition, and semantic role labeling is presented in (Collobert and Weston, 2008) and (Collobert et al., 2011). Also, recurrent networks for language models are considered in (Mikolov et al., 2010; Mikolov et al., 2011).

Alternatively, term counting combined with matrix factorization methods have been used with success, as in Latent Semantic Analysis (LSA) (Deerwester et al., 1990) and Hyperspace Analogue to Language (HAL) (Lund and Burgess, 1996).

We can refer to term counting approaches as *context-counting* models and to the training based alternatives as *context-predicting* models as suggested in (Baroni et al., 2014) where a comparison between them is done showing that the context-predicting models gives better results.

Improved results with a neural approach were introduced in (Mikolov et al., 2013a) where Continuous Bag of Words (cbow) and skip-gram models are presented. Mikolov also introduces the word analogy task, observing that analogue related words hold vector offsets in representations (e.g. $king - queen \approx man - woman$) and achieves state-of-the-art results on its own introduced benchmark.

GloVe (for *Glo*bal *Ve*ctors) was introduced in (Pennington et al., 2014), reporting better results with less computing effort, even on smaller corpora. These results encouraged us to choose it to build the representations. GloVe method can be seen as a combination of counting and predictive models since it consists of resolving a least square problem that includes context-counting information in its formulation.

1.2. Applications

Application range is unlimited. Just to name some application examples, many articles that consider context-counting based word vectors for tasks including word clustering, word sense disambiguation and semantic role labeling, are cited in (Turney and Pantel, 2010).

Using a neural network approach and context-predicting based word vectors, great performance is achieved in (Socher et al., 2011) for paraphrase detection, in (Socher et al., 2013a) for parsing and in (Socher et al., 2013b) for sentiment analysis, achieving state-of-art results in paraphrase detection and sentiment analysis.

1.3. Word Vectors Evaluation

The main ways to evaluate the quality of word vectors used in latest works (Pennington et al., 2014; Mikolov et al., 2013a) are similarity and analogy tasks. Word similarity rests upon the fact that similar words should have similar representations (close vectors). Word analogy task evaluates whether analogy relations between words can be calculated in terms of offset vectors. We will return to these evaluation tasks in section 3..

Despite the fact that analogy and similarity tasks are the most commonly used, other tasks were considered for evaluation, such as synonym detection, concept categorization and selectional preferences , as explained in (Baroni et al., 2014).

1.4. Word Vectors in Spanish

As far as we know, the only work on word representations that considers Spanish is Polyglot (Al-Rfou et al., 2013). It generates representations from Wikipedia sites for more than one hundred languages with machine translation in mind but no evaluation progress is reported.

2. GloVe Method

GloVe (Global Vectors), introduced by Pennington in (Pennington et al., 2014), is a non-supervised approach for computing word's representations that gives state-of-the-art results in similarity and analogy tasks.

This article does not contain a detailed description of the method, however, general aspects are described.

The method counts words co-occurrences in a corpus and uses them to formulate a least square problem.

Let's denote by X the word-word co-occurrence matrix, where X_{ij} is the number of times that j occurs in the context of i [1]. Then we can define

$$P_{ij} = P(j|i) = \frac{X_{ij}}{\sum_k X_{ik}} \tag{1}$$

as the probability that the word j appears in the context of word i.

The main fact of the method is how $\frac{P_{ik}}{P_{jk}}$ behaves, that can be summarized as follows:

$$\frac{P_{ik}}{P_{jk}} \begin{cases} > 1 & \text{k more related to i} \\ < 1 & \text{k more related to j} \\ \approx 1 & \text{k equally related to i and j} \end{cases}$$

Then, consider

$$w_i^T \tilde{w}_k = \log P_{ik}. \tag{2}$$

Two representations, w_i and $\tilde{w}_k$, are considered to reflect that one word is considered as context. Note that $P(k|i)$ is not necessarily the same as $P(i|k)$.[2]

Equation (2) means that

$$(w_i - w_j)^T \tilde{w}_k = \log \frac{P_{ik}}{P_{jk}}, \tag{3}$$

relating vector difference and product with the quotient of probabilities mentioned before.

From equations (1) and (2), we have that

$$w_i^T \tilde{w}_k = \log X_{ik} - \log \sum_j X_{ij}. \tag{4}$$

Note that $\log \sum_j X_{ij}$ is independent of k, so it can be absorbed into a bias b_i and to keep symmetry an additional bias $\tilde{b}_k$ is added resulting in

$$w_i^T \tilde{w}_k + b_i + \tilde{b}_k = \log X_{ik}. \tag{5}$$

Finally, equation (5) is casted to the following weighted least squares regression model

$$J = \sum_{i,k=1}^{V} f(X_{ik})(w_i^T \tilde{w}_k + b_i + \tilde{b}_k - \log X_{ik})^2,$$

where V is the vocabulary size and f a weighting function with convenient properties, such as, vanish in 0 and not overweight rare or frequent co-occurrences.

3. Experiments in Spanish

Experiments presented are centered in word analogy and similarity tasks as in (Pennington et al., 2014).

There weren't many data available for evaluation purposes in Spanish so we generated these data based on existing English resources. These test sets are available for future use and comparisons. The generated data for analogy and similarity tasks are described in 3.2. and 3.3. respectively.

[1] Context means a fixed size, symmetric or asymmetric, window of i. Context size and context symmetry are method hyper parameters.

[2] This means that two word representations will be learned. The final representation is the sum of both $(w_i + \tilde{w}_i)$, argued by the fact that combining neural network results sometimes leads to better performance.

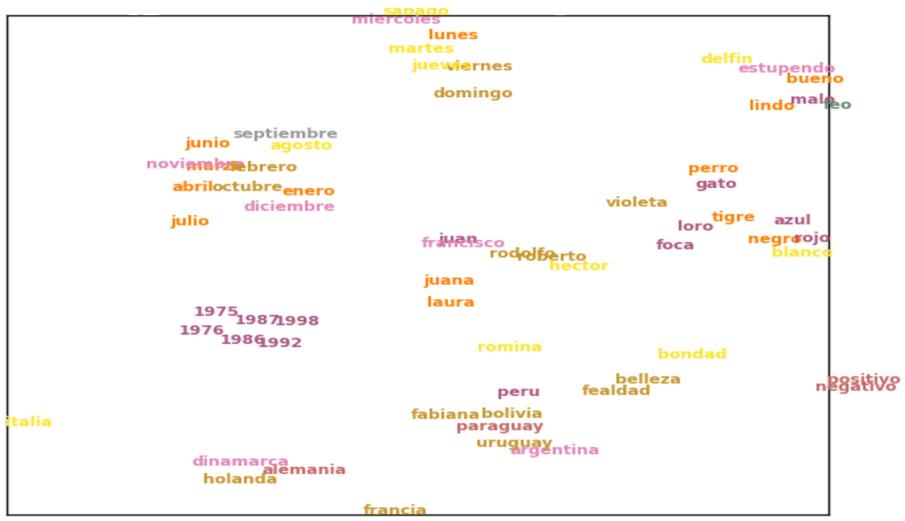

Figure 1: Sample of words representations visualization. Dimension is reduced using t-sne and vectors considered are 150-dimensioned. Note how related words are clustered

3.1. Corpora and training

The source of our corpus is a Spanish Wikipedia dump [3].
We use the Wikipedia pre-processing perl script available in Matt Mahoney's page [4] for cleaning the boilerplate extended to considerate accented characters used in Spanish. We use the corpus in lower case, losing the distinction of proper nouns (e.g. apple -the fruit- and Apple -the company-), but avoiding any error introduced by letter case disambiguation.
The training was performed using C implementation available on GloVe's site . The training time for the most expensive experiments was less than one day on an actual dual core pc.

3.2. Word Analogies

The *analogy task* (Mikolov et al., 2013b) consists in, given two pairs of related words holding the same relation, infer one of them, knowing that the offset vectors of the related words is preserved.
Consider a related to b in the same relation that c is related to d (usual notation used is a:b::c:d) then vectors offset is preserved, meaning that $a - b \approx c - d$. So, if we consider a:b::c:$?$, we can infer that the answer is b retrieving the closest vector to $c + b - a$.
For example, consider that *montevideo*:*uruguay*::*londres*:*inglaterra* is our analogy. We would read it as "*montevideo* is to *uruguay* as *londres* is to *inglaterra*". They hold the same relation of being capital city and, because offset vectors are preserved , the vector *uruguay* $-$ *montevideo* $+$ *londres* should be close to the vector of *inglaterra*.
Word analogies can be syntactic or semantic. Syntactic analogies are those based on a syntactic feature. For example, *run*:*running*::*grow*:*growing* holds the relation infinitive-gerund. Semantic analogies are based on words meaning as in the example of capital cities.
We translated the analogies dataset used in (Pennington et al., 2014), composed of twenty thousand questions, to evaluate the trained vectors.
The results are shown in Table 1. The score is the percentage of correct answers, considering that a result is correct when the expected word is contained in the five closest ones.

Dataset	**Dimension**				
semantic	**25**	**50**	**100**	**150**	**200**
capital-comm	40.4	65.1	72.5	74.4	75.4
capital-world	21.3	40.3	51.3	53.2	51.8
city-in-state	25.6	42.8	52.6	57.1	59.0
currency	0.3	0.7	0.7	0.7	0.6
family	62.6	78.0	79.6	81.8	80.1
syntactic	**25**	**50**	**100**	**150**	**200**
adj-to-adv	4.5	6.0	8.9	9.7	8.3
opposite	4.0	7.6	8.5	10.1	11.7
comparative	-	-	-	-	-
superlative	-	-	-	-	-
present-part	21.9	29.0	37.1	35.7	32.9
nation-adj	44.0	68.3	81.8	86.0	86.6
past-tense	12.3	21.4	26.9	27.5	27.7
plural	13.5	22.7	30.9	33.0	36.5
plural-verbs	26.9	39.8	47.5	45.7	43.1

Table 1: Word analogy task results for each dataset using different dimensions. Analogies are grouped in syntactic and semantic keeping the same classification between syntactic and semantic. The comparative and superlative sets do not apply in Spanish.

3.3. Word Similarities

Word vectors should generalize particular cases in NLP tasks, so similar words should have similar representations.

[3] `http://dumps.wikimedia.org/eswiki/20150228/eswiki-20150228-pages-articles.xml.bz2`

[4] `http://mattmahoney.net/dc/textdata.html`

To evaluate similarity we use some similarity tasks and visualize a sample of words in a reduced dimension space using t-sne (van der Maaten and Hinton, 2008). In the visualizations it can be seen how related words are clustered (see figure 1).

In similarity test sets for Spanish we use a translation of *WordSim-353* (Finkelstein et al., 2002) and MC30 (Miller and Charles, 1991) provided by (Hassan and Mihalcea, 2009) with some spelling corrections. Also, we manually translate SimLex-999 (Hill et al., 2014).

For the translation of SimLex-999 we considered similarity scores to disambiguate words senses to obtain a better quality translation, leaving as blank complicated cases. We found fifty five cases difficult to translate from the 999 contained in the test set.

SimLex-999 is focused on similarity relations, however, an association score is also included (from the University of South Florida Free Association Database). We evaluate vectors considering both scores. As expected, better results were reached with the association score.

We present Spearman rank correlation for each test set in Table 2. The results are mostly under state-of-the-art values for English, probably due to the reduced size of our corpus. The numbers obtained in MC30 are significantly better than in the others datasets but note that it is by far the smallest dataset and more common words are considered. In fact, *WordSim-353* is an extension of MC30.

Dim	WS353	MC30	SL999a	SL999
25	19.9	64.6	14.7	11.7
50	26.7	67.6	18.8	16.0
100	28.8	67.0	23.7	19.3
150	30.5	65.5	25.5	20.0
200	30.5	64.2	26.0	20.7
250	30.5	61.6	27.2	21.3

Table 2: Similarity results using Spearman rank correlation on many word similarity datasets. Results for different dimensions are reported. Note that SimLex-999 (SL999) measures similitude rather than relatedness. SL999a is the SimLex-999 set considering the Assoc(USF) score that reflects association rather than similarity

4. Conclusion

This work is a resource contribution for the Spanish language in terms of the trained vectors and the translated versions of two test sets for evaluation purposes. Results do not attain yet similar measures for English, and we propose to train vectors with larger corpora and evaluate them emphasizing in syntactic analogies, our worst results till now. We also propose to improve Spanish test sets and reannotate and validate SimLex-999 scores for the Spanish version. Finally, we encourage the use of the learned word vector representations in challenging NLP tasks.

5. Bibliographical References

Al-Rfou, R., Perozzi, B., and Skiena, S. (2013). Polyglot: Distributed word representations for multilingual nlp. *In Proceedings Seventeenth Conference on Computational Natural Language Learning (CoNLL).*

Baroni, M., Dinu, G., and Kruszewski, G. (2014). Dont count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. *Association for Computational Linguistics (ACL).*

Bengio, Y., Ducharme, R., Vincent, P., and Jauvin, C. (2003). A neural probabilistic language model. *Journal of Machine Learning Reseach, 3(6).*

Bengio, Y., Schwenk, H., Sencal, J., Morin, F., and Gauvain, J. (2006). Neural probabilistic language models. *Innovations in Machine Learning, pages 137186.*

Collobert, R. and Weston, J. (2008). A unified architecture for natural language processing: deep neural networks with multitask learning. *In Proceedings of ICML, pages 160167.*

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., and Kuksa, P. (2011). Natural language processing (almost) from scratch. *JMLR, 12:24932537.*

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science, 41.*

der Maaten, L. V. and Hinton, G. (2008). Visualizing high-dimensional data using t-sne. *Journal of Machine Learning Research 9(Nov):2579-2605.*

Elman, J. (1991). Distributed representations, simple recurrent networks, and grammatical structure. *Machine learning, Vol. 7(2), pages 195225.*

Hassan, S. and Mihalcea, R. (2009). Cross-lingual semantic relatedness using encyclopedic knowledge. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP).*

Hill, F., Reichart, R., and Korhonen, A. (2014). Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Preprint pubslished on arXiv. arXiv:1408.3456.*

Hinton, G. (1986). Learning distributed representations of concepts. *In Proceedings of the eighth annual conference of the cognitive science society, pages 112. Amherst, MA.*

Lund, K. and Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instrumentation, and Computers, Vol. 28(2), pages 203208.*

Mikolov, T., Karafiat, M., Cernocky, J., and Khudanpur, S. (2010). Recurrent neural network based language model. *In Proceedings of Interspeech.*

Mikolov, T., Kombrink, S., Burget, L., Cernocky, J., , and Khudanpur, S. (2011). Extensions of recurrent neural network based language model. *In Proceedings of ICASSP.*

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient estimation of word representations in vector space. *In Proceedings of Workshop at ICLR.*

Mikolov, T., Yih, W., and Zweig, G. (2013b). Linguistic regularities in continuous space word representations. *In Proceedings of NAACL HLT.*

Miller, G. A. and Charles, W. G. (1991). Contextual correlates of semantic similarity. *Language and Cognitive Processes, 6(1), 1-28.*

Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. *Conference on Empirical Methods in Natural Language Processing (EMNLP).*

Pollack, J. (1990). Recursive distributed representations. *Artificial Intelligence, Vol. 46(1), pages 77105.*

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning internal representations by backpropagating errors. *Nature, Vol. 323, pages 533-536.*

Socher, R., Huang, E. H., Pennington, J., Ng, A. Y., and Manning, C. D. (2011). Dynamic pooling and unfolding recursive autoencoders for paraphrase detection. *Advances in Neural Information Processing Systems (NIPS).*

Socher, R., Bauer, J., Manning, C. D., and Ng, A. Y. (2013a). Parsing with compositional vector grammars. *Association for Computational Linguistics 2013 Conference (ACL).*

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C., Ng, A., and Potts, C. (2013b). Recursive deep models for semantic compositionality over a sentiment treebank. *Conference on Empirical Methods in Natural Language Processing (EMNLP).*

Turney, P. D. and Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research, 37:141188.*

Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., and Wolfman, G. (2002). Placing search in context: The concept revisited. *ACM Transactions on Information Systems, 20(1), 116-131.*