Spanish Word Embeddings Learned on Word Association Norms

Helena Gómez-Adorno[1][0000−0002−6966−9912], Jorge Reyes-Magaña[2,3][0000−0002−8296−1344], Gemma Bel-Enguix[2][0000−0002−1411−5736], and Gerardo Sierra[2][0000−0002−6724−1090]

[1] Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas, Universidad Nacional Autónoma de México, Ciudad de México, México
helena.gomez@iimas.unam.mx
[2] Instituto de Ingeniería, Universidad Nacional Autónoma de México, Ciudad de México, México
gbele@iingen.unam.mx, gsierram@iingen.unam.mx
[3] Facultad de Matemáticas, Universidad Autónoma de Yucatán, Mérida, Yucatán
jorge.reyes@correo.uady.mx

Abstract. Word embeddings are vector representations of words in an n-dimensional space used for many natural language processing tasks. A large training corpus is needed for learning good quality word embeddings. In this work, we present a method based on the *node2vec* algorithm for learning embeddings based on paths in a graph. We used a collection of Word Association Norms in Spanish to build a graph of word connections. The nodes of the network correspond to the words in the corpus, whereas the edges correspond to a pair of words given in a free association test. We evaluated our word vectors in human annotated benchmarks, achieving better results than those trained on a billion-word corpus such as, word2vec, fasttext, and glove.

Keywords: word vectors · node2vec · word association norms · Spanish

1 Introduction

The representation of words in a vector space is a very active research area in the latest decades. Computational models like the singular value decomposition (SVD) and the latent semantic analysis (LSA) are capable of modeling word vector representations (*word embeddings*) from the term-document matrix. Both methods can reduce a dataset of N dimensions using only the most important features. Recently, *Mikolov et al.* [19] introduced *word2vec* inspired by the distributional hypothesis establishing that words in similar contexts tends to have similar meanings [22]. This method uses a neural network in order to learn vector representations of words by predicting other words in their context. The vector representation of a word obtained by *word2vec* has the awesome capability of conserving linear regularities between words.

In order to build a model of adequate and reliable vector space, capable of capturing semantic similarity and linear regularities of words, large volumes of text are needed. Although *word2vec* is fast and efficient to train, and pre-trained word vectors are usually available online, it is still computationally expensive to process large volumes of data in non-commercial environments, that is, on personal computers.

Free association is an experimental technique commonly used to discover the way in which the human mind structures knowledge [8]. In free association tests, a person is asked to say the first word that comes to mind in response to a given *stimulus* word. The set of lexical relations obtained with these experiments is called Word Association Norms (WAN). These kinds of resources reflect both semantic and episodic contents [6].

In previous work [4] we learn word vectors in English from a graph obtained from a WAN corpus. The vectors learned from this graph were able to map the contents of semantic and episodic memory in vector space. For this purpose, we used the *node2vec* algorithm [14] which is able to learn node mappings to a continuous vector space from the complete network taking into account the neighborhood of the nodes. The algorithm performs biased random paths to explore different neighborhoods in order to capture not only the structural roles of the nodes in the network but also the communities to which they belong to.

In this paper, we extend previous work of learning word vectors in English[4] by learning vector representations of words from a resource that collects words association norms in Spanish. We build two embedding resources of different dimensions, the first one based on *Normas de Asociación Libre en Castellano* [10] (NALC), and the other using the corpus of *Normas de Asociación de Palabras para el Español de México* [2] (NAP). The obtained embeddings from both resources are available on GitHub, the NALC based embeddings[4] and the NAP based embeddings[5].

The rest of the paper is organized as follows. In section 2, we discuss the related work. In Section 3, we present the corpora of Word Association Norms. In section 4, we describe the methodological framework for learning word vectors from WAN's. Section 5, shows the evaluation of the generated vectors, using a word similarity dataset in Spanish. Finally, in section 6 we draw some conclusions and point out to possible directions of future work.

2 Related Work

Semantic networks [25] are graphs relating words [1] used in linguistics and psycholinguistics not only to study the organization of the vocabulary but also to approach the structure of knowledge. Many languages have corpora of WAN. In the past decades, different association lists were elaborated with the collaboration of a large number of volunteers. However, in recent years, the web has

[4] `https://github.com/jocarema/nalc_vectors`
[5] `https://github.com/jocarema/nap_vectors`

become a natural way to get data to build such resources. *Jeux de Mots*[6] provides an example in French [18], whereas the *Small World of Words*[7] contained datasets in 14 languages at the time of writing.

Sinopalnikova and Smrz [24] showed that WATs are comparable to balanced text corpora and can replace them in case of absence of a corpus. The authors presented a methodological framework for building and extending semantic networks with word association thesaurus (WAT), including a comparison of quality and information provided by WAT vs. other language resources.

Borge-Holthoefer & Arenas [6] used free association information for extracting semantic similarity relations with a Random Inheritance Model (RIM). The obtained vectors were compared with LSA-based vector representations and the WAS (word association space) model. Their results indicate that RIM can successfully extract word feature vectors from a free association network.

In a recent work by *De Deyne et al.* [9] the authors introduced a method for learning word vectors from WANs using a spreading activation approach in order to encode a semantic structure from the WAN. The authors used part of the *Small World of Words* network. The word association-based model was compared with a word embeddings model (word2vec) using relatedness and similarity judgments from humans, obtaining an average of 13% of improvement over the word2vec model.

3 Word Association Norms in Spanish

Many languages have compilations of word association norms. In the past decades, some interesting works have been developed with a large number of volunteers. Among the most well-known English resources accessible on the web are the *Edinburgh Associative Thesaurus*[8] (EAT) [17] and the resource of *Nelson et al.*[9] [21].

For Spanish, there are some corpora of free words association, in this work we used two WAN resources in Spanish: a) *Corpus de Normas de Asociación de Palabras para el Español de México* (NAP) [2] and b) *Corpus de Normas de Asociación Libre en Castellano* [10] (NALC).

The NAP corpus was elaborated with a group of 578 native Mexican speakers young adults, 239 men and 339 women, with ages ranging from 18 to 28 years, and with a range of education of at least 11 years. The total number of tokens in the corpus is 65731, with 4704 different words. The authors used 234 stimulus words, all of them common nouns taken from the *MacArthur word compression and production* [16]. It is important to mention that although the stimuli are always nouns, the associated words are free-choice, that is, the informants can relate to the word stimulus with any word regardless of its grammatical category.

[6] `http://www.jeuxdemots.org/`
[7] `https://smallworldofwords.org/`
[8] `http://www.eat.rl.ac.uk/`
[9] `http://web.usf.edu/FreeAssociation`

For each *stimuli* and its *associates*, the authors computed different measures: time, frequency and association strength.

The NALC corpus includes 5819 *stimuli* words and their corresponding *associates* obtained from the free association responses of a sample of 525 subjects for 247 words, of 200 subjects for 664 words and of 100 for the remaining words. In the compilation of association norms, approximately 1500 university students have participated so far. All the subjects had Spanish as their native language and participated voluntarily in the empirical study. The total number of different words in the corpus is 31207.

4 Learning Word Embeddings on Spanish WANs

The graph that represents a given WAN corpus is formally defined as $G = \{V, E, \phi\}$ where:

- $V = \{v_i | i = 1, ..., n\}$ is the finite set of nodes with size n, $V \neq \emptyset$, which corresponds to *stimuli* words along with its *associates*.
- $E = \{(v_i, v_j) | v_i, v_j \in V, 1 \leq i, j \leq n\}$, is the set of edges, which corresponds to the connections between *stimuli* and *associates* words.
- $\phi : E \rightarrow \mathbb{R}$, is a weighting function over the edges.

We performed experiments with directed and non-directed graphs. In the directed graphs, each pair of nodes (v_i, v_j) follows an established order where the initial node v_i corresponds to the *stimulus* word and the final node v_j to an associated word. For the non-directed graph, all the *stimuli* are connected with their correspondent associates without any order of precedence. We evaluated three edges weighting functions:

Time It measures the seconds the participant takes to give an answer for each *stimulus*.

Frequency It establishes the number of occurrences of each of the associated words with a *stimulus*. In this work we use the inverse frequency (IF):

$$IF = \Sigma F - F$$

where F the frequency of a given associated word, and ΣF is the sum of the frequencies of the words connected to the same *stimulus*

Association Strength Establishes a relation between the frequency and the number of responses for each *stimulus*. It can be calculated as follows:

$$AS_W = \frac{AW * 100}{\Sigma F}$$

where AW is the frequency of a given word associated with a *stimulus*, and ΣF the sum of the frequencies of the words connected the same *stimulus* (the total number of answers). We also used the inverse of the association strength (IAS):

$$IAS = 1 - \frac{F}{\Sigma F}$$

The NAP corpus provides the three weighting functions, however for the NALC corpus only the association strength is available. Thus, in our evaluation we only report results using the association strength for the NALC corpus.

4.1 Node2vec

Node2vec [14] finds a mapping $f : V \rightarrow \mathbb{R}^d$ that transforms the nodes of a graph into vectors of d-dimensions. It defines a neighborhood in a network $N_s(u) \subset V$ for each node $u \in V$ through a S sampling strategy. The goal of the algorithm is to maximize the probability of observing subsequent nodes on a random path of a fixed length.

The sampling strategy designed in *node2vec* allows it to explore neighborhoods with skewed random paths. The parameters p and q control the change between the breadth-first search (BFS) and depth-first search (DFS) in the graph. Thus, choosing an adequate balance allows preserving both the structure of the community and the equivalence between structural nodes in the new vector space.

In this work, we used the implementation of the project *node2vec*, which is available on the web[10] with default values for all parameters. We also examined the quality of vectors with a different number of dimensions.

5 Spanish Word Embeddings Evaluation

There are several evaluation methods for unsupervised word embeddings methodologies [23], which are categorized as extrinsic and intrinsic. In the extrinsic evaluation, the quality of the word vectors is evaluated by the improvement of performance in a given natural language processing tasks (PLN) [12, 13]. Intrinsic evaluation measures the ability of word vectors to capture syntactic or semantic relationships [3].

The hypothesis of the intrinsic evaluation is that similar words should have similar representations. So, we first performed a visualization of a sample of words using the T-SNE projection of the word vectors in a two-dimensional vector space. Figure 1 shows how the words that are related to each other are grouped. We show the word vectors obtained from graphs with the three weighting functions using the NAP corpus only. It is observed that in all cases the vectors illustrate some interesting phenomena. For example, when frequency is taken as weight (the graph below), the word *pájaro* (bird) is drawn very close to *avión* (plane). From this, it is inferred that the feature "fly" is more representative than "animal" for the model. For its part, the word *caballo* (horse), is represented closer to *camioneta* (truck) than to other animals, focusing more on its status as "transportation".

In addition, we evaluated the ability of word vectors to capture semantic relationships through a word similarity task. Specifically, we used two widely

[10] `http://snap.stanford.edu/node2vec/`

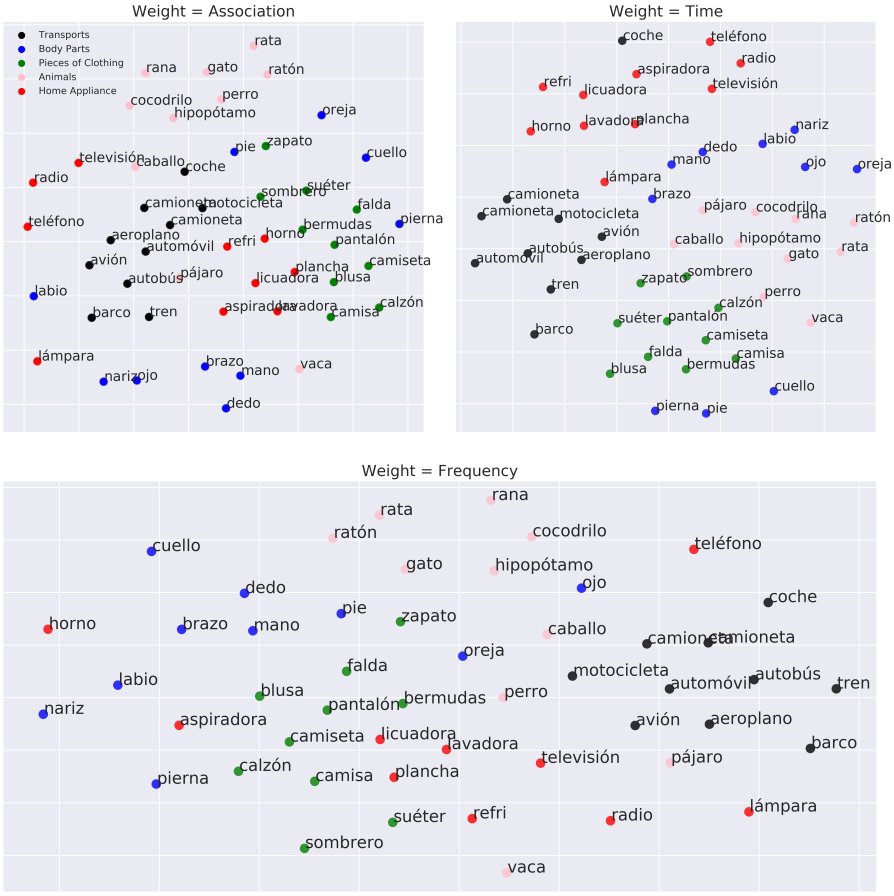

Fig. 1. Projection of the word vectors in 5 semantic groups (of ten words each).

known corpora: a) the corpus *WordSim-353* [11] composed of pairs of terms semantically related with similarity scores given by humans and b) the MC-30 [20] benchmark containing 30 word pairs. Both datasets in its Spanish version [11] [15].

We calculated the cosine similarity between the vectors of word pairs contained in the above mentioned datasets and compare it with the similarity given by humans using the Spearman correlation. To deal with the non-inclusion of every word of the testing data sets in our NALC word association norms, we introduced the concept of overlap in the experiments and calculated the total number of common words between the lists that are being compared. The others are excluded from the evaluation. In principle, having large overlaps is a positive feature this approach. Tables 1 and 2 present the Spearman corre-

[11] `http://web.eecs.umich.edu/~mihalcea/downloads.html`

lation, of the similarity given by human taggers, with the similarity obtained with word vectors (learned from NAP and NALC separately). We also report different dimensions of word vectors learned on the non-directed graphs with different weighting functions. We also report the overlap, which is the number of words that can be found in in both, the given WAN corpus (NAP or NALC) and the evaluation dataset (ES-WS-53 or MC-30).

Table 1. Spearman rank order correlations between Spanish WAN embeddings (based on cosine similarity) and the ES-WS-353 dataset.

	NAP Overlap 140		NALC Overlap 322	
Dimension	Inv. Frequency	Inv. Association	Time	Inv. Association
300	0.489	0.463	0.461	0.650
200	0.454	0.456	0.491	0.641
128	0.503	0.463	0.450	0.659
100	0.471	0.478	0.495	**0.664**
50	**0.523**	**0.503**	0.503	0.626
25	0.484	0.478	**0.572**	0.611

Table 2. Spearman rank order correlations between Spanish WAN embeddings (based on cosine similarity) and the MC-30 dataset

	NAP Overlap 11		NALC Overlap 27	
Dimension	Inv. Frequency	Inv. Association	Time	Inv. Asociation
300	0.305	**0.563**	0.545	0.837
200	0.468	0.381	0.263	**0.844**
128	**0.545**	0.272	0.300	0.767
100	0.336	0.418	0.372	0.806
50	0.527	0.509	0.272	0.814
25	0.454	0.400	**0.563**	0.788

It can be observed that the word embeddings obtained from the NALC corpus achieved better correlation with the human similarities than the embeddings obtained from the NAP corpus in both datasets, ES-WS-53 and MC-30. The difference in the results can be explained by the size of the vocabulary in both WANs, the NALC corpus has higher overlap with both evaluation datasets than the NAP corpus.

In order to test and compare the quality of the Spanish word vectors, we also performed the experiments with pre-trained Spanish vectors[12]. We selected three word embeddings models: word2vec[13], gloVe[14], and fasttext[15].

Table 3 shows the Spearman rank order correlation between the cosine similarity obtained with word vectors pre-trained in large corpora and the similarity of humans (obtained from *WordSim-353*) and *MC-30* datasets) in comparison with the correlation between NAP embeddings and the humans rated similarities. In the same way, Table 4 shows the same comparison with pre-trained word vectors and the NALC based embeddings.

The highest correlation value was obtained with the vectors trained with the fasttext [5] model. The vectors trained on the Wikipedia in Spanish obtained the best results among the pre-trained models. Our method outperformed the results obtained by the pre-trained vectors when the vectors were learned on the NALC corpus in both evaluation datasets, ES-WS-353 and MC-30.

Table 3. Spearman rank order correlation comparison of NAP embeddings and pre-trained word vectors with the evaluation datasets.

Source	Vector size	MC-30 (Overlap 11)	ES-WS-353 (Overlap 140)
Fasttext-sbwc	300	0.881	0.639
Fasttext-wiki	300	**0.936**	**0.701**
Glove-sbwc	300	0.827	0.532
Word2vec-sbwc	300	0.890	0.634
n2v-Inverse Association	300	0.563	0.463
n2v-Inverse Frequency	300	0.305	0.489
n2v-Time	25	0.563	0.572

6 Conclusions and Future Work

We introduced a method for learning Spanish word embeddings from a Corpus of Word Association Norms. For learning the word vectors, we applied the *node2vec* algorithm on the graph of two WAN corpora, NAP and NALC.

We employ weighting functions on the edges of the graph taking into account three different criteria: time, inverse frequency and inverse associative strength. The best results have been obtained with the association strength, however, the time weighting function also achieved high results. Words with a higher

[12] `https://github.com/uchile-nlp/spanish-word-embeddings`
[13] `https://code.google.com/archive/p/word2vec/`
[14] `https://nlp.stanford.edu/projects/glove/`
[15] `https://github.com/facebookresearch/fastText/blob/master/pretrained-vectors.md`

Table 4. Spearman rank order correlation comparison of NALC embeddings and pre-trained word vectors with the evaluation datasets.

Source	Vector size	MC-30 (Overlap 27)	ES-WS-353 (Overlap 322)
Fasttext-sbwc	300	0.762	0.613
Fasttext-wiki	300	0.793	0.624
Glove-sbwc	300	0.707	0.482
Word2vec-sbwc	300	0.795	0.624
n2v-Inverse Association	300	0.837	0.650
n2v-Inverse Association	200	**0.844**	**0.664**

association strength usually have a shorter formulation time, which leads to the algorithm to connect more related words in a neighborhood because the node2vec algorithm looks for shorter paths in the graphs.

The results we obtained using the NALC corpus are higher than those obtained with pre-trained word embeddings trained on large corpora. The performance even improves the results achieved with the vectors trained on the Spanish billion words corpus [7]. However, some simple strategies would help improve our results. Some of them would be to adjust the parameters of the algorithm and adapt the system to different types of neighborhoods for the nodes, which could produce different configurations of the vectors. In future work we will perform an extrinsic evaluation these Spanish word vectors, i.e. in some Natural Language Processing task [4].

The evaluations carried out with the vectors learned on the NAP corpus also showed promising results with respect to the similarity and relational indexes. However, due to the low vocabulary length, the results were lower than those obtained on pre-trained embeddings. As future work, we plan to solve this problem by automatically generate word association norms between pairs of words retrieved from a medium-sized corpus. With this process, we will build a new resource that can account for syntactic, semantic and cognitive connections between words.

Acknowledgments

This work was partially supported by the following projects: Conacyt FC-2016-01-2225 and PAPIIT IA401219, IN403016, AG400119.

References

1. Aitchison, J.: Words in the mind: An introduction to the mental lexicon. John Wiley & Sons (2012)
2. Arias-Trejo, N., Barrón-Martínez, J.B., Alderete, R.H.L., Aguirre, F.A.R.: Corpus de normas de asociación de palabras para el espaol de Mxico [NAP]. Universidad Nacional Autnoma de Mxico (2015)

3. Baroni, M., Dinu, G., Kruszewski, G.: Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). vol. 1, pp. 238–247 (2014), http://www.aclweb.org/anthology/P14-1023
4. Bel-Enguix, G., Gómez-Adorno, H., Reyes-Magaña, J., Sierra, G.: Wan2vec: Embeddings learned on word association norms. Semantic Web (2019)
5. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. Computing Research Repository **arXiv:1607.04606** (2016). https://doi.org/10.1162/tacl_a_00051, https://arxiv.org/abs/1607.04606
6. Borge-Holthoefer, J., Arenas, A.: Navigating word association norms to extract semantic information. In: Proceedings of the 31st Annual Conference of the Cognitive Science Society (2009)
7. Cardellino, C.: Spanish Billion Words Corpus and Embeddings (March 2016), http://crscardellino.github.io/SBWCE/
8. De Deyne, S., Navarro, D.J., Storms, G.: Associative strength and semantic activation in the mental lexicon: Evidence from continued word associations. In: Proceedings of the 35th Annual Conference of the Cognitive Science Society. Cognitive Science Society (2013)
9. De Deyne, S., Perfors, A., Navarro, D.J.: Predicting human similarity judgments with distributional models: The value of word associations. In: Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers. pp. 1861–1870 (2016). https://doi.org/10.24963/ijcai.2017/671
10. Fernandez, A., Díez, E., Alonso, M.: Normas de asociación libre en castellano de la universidad de salamanca (2010), http://inico.usal.es/usuarios/gimc/normas/index_nal.asp
11. Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., Ruppin, E.: Placing search in context: The concept revisited. In: Proceedings of the 10th International Conference on World Wide Web. pp. 406–414. ACM (2001)
12. Gómez-Adorno, H., Markov, I., Sidorov, G., Posadas-Durán, J., Sanchez-Perez, M.A., Chanona-Hernandez, L.: Improving feature representation based on a neural network for author profiling in social media texts. Computational Intelligence and Neuroscience **2016**, 13 pages (2016)
13. Gómez-Adorno, H., Posadas-Durán, J.P., Sidorov, G., Pinto, D.: Document embeddings learned on various types of n-grams for cross-topic authorship attribution. Computing pp. 1–16 (2018)
14. Grover, A., Leskovec, J.: node2vec: Scalable feature learning for networks. In: Proceedings of the 22nd ACM International Conference on Knowledge Discovery and Data Mining. pp. 855–864. ACM (2016)
15. Hassan, S., Mihalcea, R.: Cross-lingual semantic relatedness using encyclopedic knowledge. In: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3-Volume 3. pp. 1192–1201. Association for Computational Linguistics (2009)
16. Jackson-Maldonado, D., Thal, D., Fenson, L., Marchman, V., Newton, T., Conboy, B.: Macarthur inventarios del desarrollo de habilidades comunicativas (inventarios): Users guide and technical manual. Baltimore, MD: Brookes (2003)
17. Kiss, G., Armstrong, C., Milroy, R., Piper, J.: An associative thesaurus of English and its computer analysis. Edinburgh University Press, Edinburgh (1973)
18. Lafourcade, M.: Making people play for lexical acquisition. In Proceedings of the th SNLP 2007, Pattaya, Thaland **7**, 13–15 (December 2007)

19. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. Computing Research Repository **arXiv:1301.3781** (2013), `https://arxiv.org/abs/1301.3781`
20. Miller, G., Charlees, W.: Contextual correlates of semantic similarity. Language and cognitive processes **6**(1), 1–28 (1991). https://doi.org/10.1080/01690969108406936
21. Nelson, D.L., McEvoy, C.L., Schreiber, T.A.: Word association rhyme and word fragment norms. The University of South Florida (1998)
22. Sahlgren, M.: The distributional hypothesis. Italian Journal of Disability Studies **20**, 33–53 (2008)
23. Schnabel, T., Labutov, I., Mimno, D., Joachims, T.: Evaluation methods for unsupervised word embeddings. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. pp. 298–307 (2015)
24. Sinopalnikova, A., Smrz, P.: Word association thesaurus as a resource for extending semantic networks. pp. 267–273 (2004)
25. Sowa, J.F.: Conceptual graphs as a universal knowledge representation. Computers & Mathematics with Applications **23**(2), 75–93 (1992). https://doi.org/10.1016/0898-1221(92)90137-7