Assessing Context for Extraction of near Synonyms from Product Reviews in Spanish

Sofía N. Galicia-Haro[1] and Alexander Gelbukh[2]

[1] Faculty of Sciences UNAM, Mexico
sngh@fciencias.unam.mx
[2] Center for Computing Research, National Polytechnic Institute, Mexico
www.Gelbukh.com

Abstract. This paper reports ongoing research on near synonym extraction. The aim of our work is to identify the near synonyms of multiword terms related to an electro domestic product domain. The state of the art approaches for identification of single word synonyms are based on distributional methods. We analyzed for this method different sizes and types of contexts, from a collection of Spanish reviews and from the Web. We present some results and discuss the relations found.

Keywords: semantic similarity, product reviews, vector space model, context

1 Introduction

The motivation of our work is to analyze if the general methods currently applied on scientific and specialized documents to extract single word synonyms or near-synonyms [1] are the methods to obtain semantically related multiword terms (MWTs) appearing in product reviews. We considered that such MWT synonyms correspond to denominative variants as previously [2] defined: different denominations restricted for example to lexicalized forms, with a minimum of consensus among the users of units in a domain, since such conceptualization reflects more properly the degree of informality expressed in the texts we used and the diversity of terms in daily life products domain.

We found this problem when we were working on a collection of washing machine reviews: There are concepts that have several possible term candidates. For example, the Delayed Start function allows the startup of the washing machine program to be delayed for a number of hours. Some of the variants of the Spanish term related to this concept are: *inicio diferido* 'delayed start', *inicio retardado* 'retarded start', *encendido programable* 'programmable switch on', *preselección de inicio* 'start time-preselection', and others including different word order of the same variants.

We supposed that such terms were generated by the authors of the opinion reviews or maybe by the translators of the washing machine instruction manuals from English to Spanish, but we found that these variants also have an origin in the instruction manuals of the manufacturers. Automatically grouping such similar MWTs should be useful

Text, Speech, and Dialogue. TSD 2016. Lecture Notes in Computer Science, vol. 9924, pp. 125–133, doi: 10.1007/978-3-319-45510-5_15. This is a pre-print version.

since there are no glossaries of semantically related terms used by different manufacturers, translated manuals, and users of electro domestic appliances. This type of knowledge should be included in service robots in the future.

For the automatic determination of single word synonyms two main paradigms have been applied: lexicon-based and distributional approach. The former paradigm requires sources of word definitions in general language, i.e. dictionaries or terminology banks are required. The second paradigm relies on the distributional hypothesis of Harris [3]: words with similar meaning tend to occur in similar contexts. The required sources for this approach are corpora. Since sources of word definitions are not available we decided to use the distributional approach. We applied this method first on a collection of washing machine review texts and then on retrieved Web contexts. The work that we describe here is an analysis of the effect of different contexts on MWT synonym extraction.

The paper is organized as follows: in the next section, we present an overview of the method, describing the materials and the details of the method we followed. In Section 3 we describe the context sizes and types we defined and some of their statistics. In Section 4, we present the results and discuss their interpretation. The final section gives the conclusions.

2 The General Method

We followed the well-known assumption that words are more likely to be semantically related if they share the same contexts. Its common implementation in the Vector Space or Word Space Model [4, 5] is based on the computation of a vector for a word w_t. Its dimensions correspond to the close neighbors of w_t, obtained from a corpus.

Seeing that review texts have many linguistic errors we did not consider grammatical relations to select the contexts. We followed other works where close neighbors are computed from one word to the left and one word to the right of w_t to a larger context. Each word neighbor w_i or entry in the vector has a value that measures the tightness between w_i and w_t. In this method, the semantic similarity of two terms is evaluated by applying a similarity measure between their vectors and then making a ranking based on such values. In this work, instead of a word w_t we consider a multiword noun phrase mw_t and its neighbors as single words. We also characterized each MWT mw_t from our collection by a vector computed from its neighbors.

2.1 Multiword Terms

In this work a MWT is a noun phrase of several words including prepositions and articles. They were obtained by the following patterns:

Noun [Noun| Adjective]

Noun Preposition [Article] Noun [Adjective]

Noun Preposition Num Noun

These patterns covered all the names of the functions and the noun phrases in the product characteristics section of the washing machine manuals that we collected. The

patterns include prepositions and articles since in colloquial Spanish it is usual to include articles. For example: *bloqueo infantil* 'child lock', *bloqueo para niños* 'child lock', *seguro para niños* 'child safety', *seguridad para niños* 'child safety', *apertura a prueba de niños* 'child-proof opening system', *sistema de bloqueo infantil* 'child lock system', *bloqueo para los niños* 'child lock' are denominative variants for the same concept: the door lock that prevents children putting their hand into the washing machine while it is working.

2.2 Corpora

Corpus of Product Reviews

We used a collection of review texts compiled in a previous study [6], named here Corpus of Product Reviews (CPR), comprising 2800 reviews extracted from the *ciao.es* website. This site has product reviews in Spanish for diverse electro domestic appliances. The collection was automatically compiled from the washing machine section and it was tagged using Freeling [7].

We wrote a program that executed a sequence of pre-processing steps. From the raw text corpus, the first step split the CPR texts in sentences. Sahlgren [8] stated that the semantic properties of the word space model were determined by the choice of context and that the more linguistically justified definition of context in which to collect syntagmatic information should be a clause or a sentence. Then the program extracted the sentences where the MWT candidates appeared, based on the POS patterns described above. Normalization was not considered so a word with the first letter in upper case was different from that with all letters in lowercase. We did not consider applying any spell-checking correction since Freeling gave the correct tag in many spelling mistakes. In the last step the sentences including MWTs with frequency higher than one were selected since the quantity of examples for each multiword phrase was low. The preprocessing resulted in 34871 sentences for 5422 MWTs.

We selected 112 noun phrases that corresponded to 97 MWTs associated with the washing machine functions. Since we did not normalize, *botón de inicio* and *boton de inicio* are different MWTs in our work. We considered that each example was related to a different term. One reason is that we had few examples for most of the MWTs and we could use their results as a base line. Also, we noticed that MWTs may have a distinct behavior since manufacturers, translators and users made them rare cases for the quantity of terms they used for a single concept.

Table 1.

#Examples	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**13**	**14**
#MUTs	45	20	13	2	6	4	1	2	1	1	2	1
#Examples	**15**	**16**	**18**	**19**	**21**	**22**	**24**	**29**	**33**	**34**	**54**	**147**
#MWTs	1	1	1	1	2	2	1	1	1	1	1	1

Table 1 shows the distribution of the quantity of examples in the collection for each MWT. For example, in the first row, the first column shows that there were 2 examples

for each one of 45 terms and the last column shows that there were 14 examples for only one term.

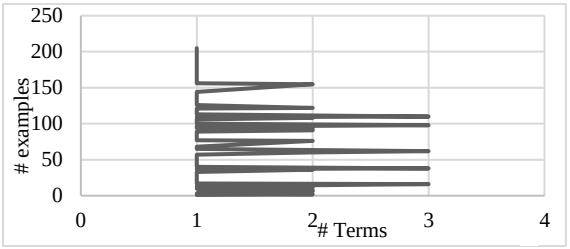
Corpus Obtained from the Web

Because of the small number of contexts in the CPR for most of the selected MWT's, we decided to acquire contexts for them from the Web. Two problems inherent in Web searching we tried to avoid: different domain for the MWT's, and incomplete context in the snippets.

We wrote a program to obtain context examples from the Web searching for each one of the selected 97 MWTs, using the Google's asterisk facility [9]. The word *lavadora* 'washing machine' was incorporated in the search to try to limit the context to such domain. For example, for the *inicio diferido* 'delayed start' term the search was launched with the string: "* * *inicio diferido* * *" *lavadora*, where the asterisks substitute for the possible sequences of words around the MWT. The Google search engine tool was limited to the Spanish language. Google returned a quantity of hits where each snippet most probably there would have a string of different words, then the MWT followed by another sequence of words. The program retrieved a maximum of 500 hits for each MWT. The total number of retrieved web contexts were 25013, from them 7251 were useful, mainly due to very short contexts and missing words corresponding to the search keywords in context in the snippets.

Figure 1 shows the distribution of the quantity of examples obtained from the Web. The horizontal axis corresponds to the number of multiword terms (from 1 to 3). The vertical axis measures the number of examples. For example, 3 terms have 100 examples, only unique terms have more than 160 examples.

Fig. 1. Distributional quantity of examples obtained from the Web



2.3 Measures for the Distributional Method

In this section we describe the measures that we considered for the tightness between the neighbors and the MWT, and for the semantic similarity of two term vectors.

There are different measures that have been applied in the distributional method, for example Ferret [10] evaluated three measures for neighbor tightness: T-test, Tf-Idf, and Pointwise Mutual Information (PMI), and six measures for term similarity: Cosine, Jaccard, Jaccard†, Dice, Dice†, Lin. He found that Cosine measure with PMI gave the best results. Hazem and Daille [11] also applied diverse measures and they found the results

were more contrastive for the Spanish corpus in comparison with a French and an English corpus. The PMI and Cosine measures performed the best for the Spanish corpus. Based on such works we decided to use PMI and Cosine measures.

The general method was applied in three steps. In the first step for each MWT mw_i its vector was built gathering all the words that co-occurred with mw_i within each specific window that we detail in the next section. In the second step the program computed the value based on PMI [12] that measures the tightness of the mw_i with each word in its specific window. This value was computed by the following equation:

$$\text{PMI}\,(mw_i, w_j) = \log \frac{\text{P}(mw_i, w_j)}{\text{P}(mw_i)\text{P}(w_j)}$$

Where

$$\text{P}(mw_i) = \text{P}(w_{pmi\_1})\ P(w_{pmi\_2})\ \square\ \text{P}(w_{pmi\_n})$$

P(mw_i, w_j) is the coocurrence of mw_i with the word w_j appearing in the window of mw_i and P(w_i) is the ocurrence of w_i in the collection. The formula for P(mw_i) is the model of independence of McInnes [13] for n-gram of any size *n*. We considered the lemma of each word to group those that differ in gender and number, for example: *lavadora* and *lavadoras* were gathered and represented by the *lavadora* lemma. For the corpus obtained from the Web, P(mw_i, w_j) and P(w_i) corresponded to the number of hits retrieved by the Google search engine in the same sense described above.

In the third step the cosine similarity [14] was computed for each pair of vectors v_k and v_l by the following equation:

$$\text{cosine}_{v_l}^{v_k} = \frac{\sum_i \text{PMI}(mw_i, l)\text{PMI}(mw_i, k)}{\sqrt{\sum_i \text{PMI}(mw_i, l)^2}\ \sqrt{\sum_i \text{PMI}(mw_i, k)^2}}$$

The candidate denominative variants of the MWT mw_i are the MWTs best ranked following their cosine value.

3 Contexts

We experimented on the selection of the quantity of word neighbors of the MWT, i.e. on the context size. But we also experimented with the restrictions imposed by the review authors according to the punctuation marks they included. Finally, we experimented delimiting the context by means of eliminating specific kind of words.

3.1 Context Sizes

Different window sizes have been defined in the distributional method. For example, Ferret [10] analyzed a measure that performed well on an extended TOEFL test, it was applied for synonym extraction. The measures were tested with window sizes between 1 and 5 words. He found the best results for the window size of 1 word on a corpus made of around 380 million words from news articles.

Rosner and Sultana [15] investigated methods for extending dictionaries using non-aligned corpora by finding translations through context similarity. The contexts were converted into vectors and then compared using the cosine measure. They used news

text as the main source of comparable text for English and Maltese. The authors tested different window sizes from 1 to 5 words, and the window size of 3 was found to be the optimal.

Hazem and Daille [11] applied a 7-window size. Their experiments were carried out on a French/English/Spanish specialized corpus from the domain of wind energy of 400,000 words. Their work was devoted to extracting synonyms of MWTs by means of a semi-compositional method.

Seeing that the best size for the window differ from one work to another we decided to use two window sizes: 12 words around the MWT, named CT12, and 6 words around the MWT, named CT6, for the contexts obtained from the CPR. Regarding the contexts extracted from the Web, since we used the snippets retrieved by the Google search engine we did not consider experimenting on sizes for no-clear-cut contexts.

3.2 Context Types

Sahlgren [8] considered that clauses and sentences or at least the functional equivalent to such entities seem to be linguistic universals, i.e. some sequence delimited by some kind of delimiter. We followed this idea considering that delimiters in the CPR texts should be taken into account to restrict the window size since users used punctuation marks with a specific purpose. We proposed to reduce the contexts according to the following punctuation marks: points, quotes, parenthesis, exclamation mark, slash, semicolon, and hyphen as delimiters of the left and right contexts.

We wrote a program to obtain two reduced contexts from the previous CT12 and CT6, delimited by the indicated punctuation marks and named CR12 and CR6 respectively. Table 2 shows the cosine values obtained for some MWTs from the CPR contexts for the various sizes and types described above. We could observe that the highest values corresponded to the taxonomic relation between *sistema de bloqueo* 'lock system' and *sistema de bloqueo infantil* 'child lock system' while the lowest values corresponded to the first row for another semantic relation across taxonomic relation links between *seguro para niños* 'child safety locks' and *sistemas de seguridad* 'security systems'. The rest of the MWTs corresponded to near synonyms or denominative variants.

Table 2. Cosine values for some MWTs obtained from CPR contexts

Multiword Terms	**CT12**	**CT6**	**CR12**	**CR6**
seguro para niños VS *sistemas de seguridad*	0.5000	0.4376	0.3574	0.2989
boton de on VS *botón de inicio*	0.5645	0.8629	0.4823	0.5215
tiempo restante VS *tiempo remanente*	0.85438	0.79748	0.89059	0.5139
programación de fin VS *fin diferido*	0.9473	**0.8929**	0.6349	0.6349
bloqueo de seguridad VS *sistema de bloqueo*	0.9485	0.8825	0.8752	0.7812
sistema de bloqueo VS *sistema de bloqueo infantil*	**0.9762**	0.8428	**0.9781**	**0.7888**

For the Web contexts we defined 3 types of context delimitation. We wrote a program to obtain the context delimited by the indicated punctuation marks and by deleting

the following function words: determinants, pronouns, numbers, conjunctions, prepositions and auxiliary verbs, named MW1. We also obtained another context delimited as MW1 and reduced additionally by deleting adverbs from them, named MW2. We supposed that attributes could not be useful for context similarity. The third context type named MW3 was delimited by deleting in addition to the previous ones the short words (1-2 letters) with unknown POS.

Table 3. Cosine values for some MWTs obtained from Web contexts

Multiword Terms	MW1	MW2	MW3
boton de on VS *botón de inicio*	0.1462	0.1477	0.1409
seguro para niños VS *sistemas de seguridad*	0.1553	0.1511	0.1779
programación de fin VS *fin diferido*	0.2357	0.2418	0.2569
bloqueo de seguridad VS *sistema de bloqueo*	0.2715	0.2639	0.2560
tiempo restante VS *tiempo remanente*	0.3146	0.3079	0.2940
sistema de bloqueo VS *sistema de bloqueo infantil*	**0.3369**	**0.3384**	**0.3285**

Table 3 shows the cosine values obtained for some MWTs according to their Web contexts. We could observe that the cosine values are lower than those obtained from the CPR collection since the number of occurrences and co-occurrences were taken from the total hits reported by the Google search engine.

4 Results and Discussion

One method for evaluating the performance of an extraction system is to compare the similarity scores assigned by the system to the results given by human judges. Since we do not have such a golden standard we manually analyzed the first 100 top results for each one of the several context types we defined. Two students that manually analyzed the first 100 top results for each kind of context were required to search in the Web to clarify the specific meaning of many MWTs since initially their agreement rate (kappa statistic [16]) was 69. The precision for the 100 top values of similarity is shown in Table 4.

Table 4. Precision for top 100 results

Context	CT12	CR6	CR12	CT6	MW1	MW2	MW3
Precision	0.33	0.43	0.45	0.48	0.55	0.60	0.60

We observe that we obtained some differences in results among the various contexts applied to the collection of product reviews. The delimitation by punctuation marks was more useful on the 12-word window increasing their results by 12%. This delimitation has an adverse effect on the 6-word context where the complete context scored 5% percent higher than its reduced counterpart.

Regarding the results obtained for the 3 context types defined for Web contexts, we observe that the precision did not change for the MW2 and MW3 context types and that

short words elimination had no effect on results. The MW2 type obtained 5% better results than the MW1 type where elimination of adverbs was the only difference between them. The attributes elimination has more sense if applied to product reviews since the texts include personal experiences, personal thoughts, opinions about anything, etc. but we wanted to analyze their effect on the Web contexts.

Despite the 60% precision obtained for the better results, we obtained several MWT groups related to a concept, we show two of such groups:

delayed start: *comienzo retrasado, marcha diferida, programación diferida, preselección de fin, función de inicio, inicio diferido, retardo horario, tiempo diferido*

on/off button: *botón de arranque, botón inicio, botón de encendido, tecla de encendido, botón de inicio*

5 Conclusions

As Sahlgren [8] stated, the distributional models are not only grounded in empirical observation, but they also rest on a solid theoretical foundation. Despite the lower quantity of examples used in this work we concluded that the results are useful according to the task complexity. We present in this work experiments to analyze the adequacy of several kind of contexts to extract denominative variants of MWTs applying the distributional method first to a collection of Spanish reviews for washing machines and then to contexts retrieved from the Web for the MWTs obtained from such product reviews.

We manually tested the results for multiword terms associated to different concepts and the best results were obtained for the Web contexts delimited by punctuation marks, function words and attributes.

Acknowledgments. The second author recognizes the support of the Instituto Politécnico Nacional, grants SIP-20161958 and SIP-20162064.

6 References

1. Edmonds, P. and G. Hirst: Near-synonymy and lexical choice. Computational Linguistics 28(2), 105–144 (2002)
2. Freixa, J.: Causes of denominative variation in terminology: A typology proposal. Terminology 12(1), 51–77 (2006)
3. Harris, Z.: Distributional structure. Word 10 (23), 146–162 (1954)
4. Schütze, H.: Dimensions of meaning. Proceedings of the ACM/IEEE conference on Supercomputing. IEEE Computer Society Press pp. 787-796 (1992)
5. Schütze, H.: Automatic word sense discrimination. Computational Linguistics 24(1), 97-123 (1998)
6. Galicia-Haro, S. N. and A. F. Gelbukh: Extraction of Semantic Relations from Opinion Reviews in Spanish. MICAI 2014 Proceedings in LNCS (8856) pp.175-190 (2014)
7. Padró, L. and E. Stanilovsky: Freeling 3.0: Towards wider multilinguality. In Proceedings of the Language Resources and Evaluation Conference LREC 2012. ELRA (2012)

8. Sahlgren, M.: The Word-Space Model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces. Ph.D. thesis, Department of Linguistics, Stockholm University (2006)
9. Gelbukh, A.F., Bolshakov, I. A.: Internet, a true friend of translator: the Google wildcard operator. International Journal of Translation 18(1-2), 41–48 (2006)
10. Ferret, O.: Testing Semantic Similarity Measures for Extracting Synonyms from a Corpus. Proceedings of the 7th International Conference on Language Resources and Evaluation LREC 2010. pp. 3338-3343 (2010)
11. Hazem, A. and B. Daille: Semi-compositional Method for Synonym Extraction of Multi-Word Terms. Proceedings of the 9th International Conference on Language Resources and Evaluation LREC2014, pp.1202-1207 (2014)
12. Manning, C. and H. Schütze: Foundations of Statistical Natural Language Processing. MIT Press, Cambridge, MA. (1999)
13. McInnes, B. T.: Extending the log-likelihood measure to improve collocation identification. Master thesis, University of Minnesota (2004)
14. Salton, G. and M. E. Lesk: Computer evaluation of indexing and text processing. Journal of the Association for Computational Machinery, 15(1), 8–36 (1968)
15. Rosner, M. and K. Sultana: Automatic Methods for the Extension of a Bilingual Dictionary using Comparable Corpora. Proceedings of the 9th International Conference on Language Resources and Evaluation LREC2014, pp. 3790- 3797 (2014)
16. Manning, C. D., P. Raghavan and H. Schütze: Introduction to Information Retrieval. Cambridge University Press. (2008)