Unsupervised Learning of P NP P Word Combinations*

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

[1] Faculty of Sciences UNAM Universitary City, Mexico City, Mexico
sngh@fciencias.unam.mx

[2] Center for Computing Research, National Polytechnic Institute, Mexico
gelbukh@cic.ipn.mx; www.Gelbukh.com

We evaluate the possibility to learn, in an unsupervised manner, a list of idiomatic word combinations of the type preposition + noun phrase + preposition (P NP P), namely, such groups with three or more simple forms that behave as a whole lexical unit and have semantic and syntactic properties not deducible from the corresponding properties of each simple form, e.g., *by means of*, *in order to*, *in front of*. We show that idiomatic P NP P combinations have some statistical properties distinct from those of usual idiomatic collocations. In particular, we found that most frequent P NP P trigrams tend to be idiomatic. Of other statistical measures, log-likelihood performs almost as good as frequency for detecting idiomatic expressions of this type, while chi-square and point-wise mutual information perform very poor. We experiment on Spanish material.

1 Introduction

Our goal is to compile, in an unsupervised manner, a list of word combinations of the type preposition + noun phrase + preposition (P NP P) constituted by three or more simple forms (the noun phrase or even a preposition can consist of more than one word) that behave as one lexical unit, with non-compositional semantics. Specifically, such combinations are frequently equivalent to prepositions, i.e., they can be considered as one multiword preposition: e.g., *in order to* is equivalent to *for* (or *to*) and has no relation with *order*; other examples: *in front of* '*before*', *by means of* 'by', etc. Apart from semantic analysis, such a dictionary can be useful in syntactic disambiguation, namely, prepositional phrase attachment: given a compound preposition *in_order_to* is present in the dictionary, the *to* in *John bought flowers in order to please Mary* would not be attached to *bought*.

We experimented with Spanish material. There is no complete dictionary of such word combinations for Spanish. Only a limited number of such combinations are included in common dictionaries, which in addition do not give their variants such as *por vía de* 'by' ('by way of') / *por la vía de* 'by' (literally 'by the way of'), etc.

In this work, we investigate unsupervised corpus-based methods to learn the word combinations of the considered type (P NP P that behave as a single lexical unit; see case 1 in the example below) and the ways to differentiate such idiomatic collocations

* Work partially supported by Mexican Government (CONACyT, SNI, CGPI-IPN, PIFI-IPN).

S. N. Galicia-Haro, A. Gelbukh. Unsupervised Learning of P NP P Word Combinations. Computational Linguistics and Intelligent Text Processing (CICLing-2005), Lecture Notes in Computer Science, N 3406, Springer-Verlag, 2005, pp. 329–332.

from literal combinations (case 2), on the one hand, and from larger idioms (in which such combinations very often participate; case 3), on the other hand:

1. Idiomatic expression: *a fin de* *obtener un ascenso* ‘to obtain a promotion’ (literally ‘at end of’),
2. Free combination: *a fin de* *año obtendrá un ascenso* ‘at the end of the year she will be promoted’,
3. Part of a larger idiom: *a fin de* *cuentas* ‘finally’ (literally ‘at end of accounts’).

In Section 2 we outline our unsupervised method and present the experimental results, which we discuss in Section 3. Finally, in Section 4 we draw some conclusions.

3 Methodology and Results

We used a texts collection obtained from Internet, corresponding to four different Mexican newspapers that daily publish online a considerable part of their complete edition. The texts correspond to diverse sections: economy, politics, sport, etc., from 1998 to 2002. The collection has approximately 60 million tokens; see details in [4].

We extracted all word strings corresponding to PNPP in the following grammar:

PNPP → P NP P
NP → N | D N | V-Inf | D V-Inf

where P stands for preposition, N for noun, D for determinant, and V-inf for infinitive verb (in Spanish, infinitives can be modified by a determinant: *el fumar está prohibido*, literally ‘the to-smoke is prohibited’).

We found 2,590,753 such PNPP string tokens, or 372,074 types, of which 103,009 had frequency higher than 2. Not all of them were idiomatic: e.g., from the 20 more frequent items, the strings *en la ciudad de* ‘in the city of’, *del gobierno de* ‘of the Government of’, *del estado de* ‘of the State of’ are free (literal) combinations.

Then we computed for each extracted item various statistical measures as described below, in order to evaluate whether these characteristics correlate with idiomatic (as a unit, e.g., *in order to*) combinations and discriminate them from free (literal, e.g., *in the head of Mary*) ones.

To compare performance of these measures, we used a list of known Spanish idiomatic word combinations [7], containing 256 cases of the considered type (P NP P). For each particular measure, we ordered the list by this measure and plotted the number of the combinations present in [7] found among the top *k* elements of our extracted list ordered by this particular measure (proportional to recall at top *k* items).

3.1 Statistical Measures Considered

Various statistical measures to identify lexical associations between words from a corpus have been suggested in literature [3]. We applied the NSP statistical package [1] to obtain the following measures:

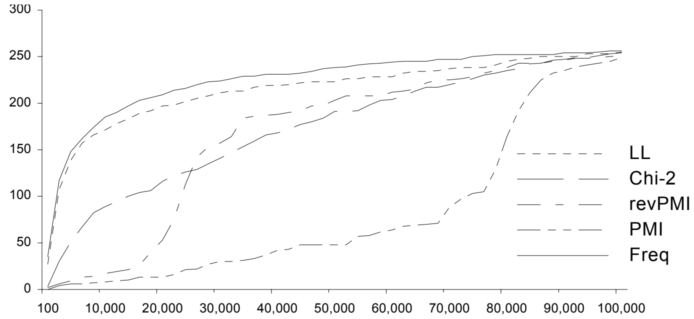

Fig. 1. Results for groups appearing more than 2 times.

– Frequency (Freq),
– Point-wise mutual information (PMI),
– Log-likelihood (LL),
– Pearson measure (χ^2, or Chi-2).

Since there is no generally accepted definition of the latter three measures for three elements, we applied pair-wise measures to [P NP] and [P] of the whole combination P NP P: e.g., the PMI assigned to *a fin de* was the PMI of the two strings (1) *a fin* and (2) *de*. Evaluation of other possible ways of calculating the dependency between the three elements was left for our future work.

We only considered the strings with frequency greater than 2, since statistical measures of this type are unreliable on sparse data. The results are shown in Figure 1.

3.2 Discussion

The best measure proved to be simple frequency, and log-likelihood shows nearly the same performance. On the first 100 items the precision obtained with frequency measure was 50% (which is 20% recall on the list [7]).

However, manual inspection of the results revealed among the top elements new idiomatic collocations, such as *por la vía de* 'by way of' (literally 'by the way of'), while [7] only contains *por vía de* 'by way of'. Thus, real precision of our method is better than that measured by comparison with the list [7], and the method allows detection of new combinations.

PMI gives very poor results, which is in accordance with the general opinion that PMI is not a good measure of dependency (though it is a good measure of independency) [6]. What is more, PMI seems to have inverse effect: it tends to group the idiomatic examples nearer the end of the list. With this, ordering the list in *reverse* order by PMI (revPMI in Figure 1) gives better results. However, such order is much worse than Freq and LL orders: it groups most of the combinations in question around the positions from 25,000 to 35,000 in the list of 100,000. Pearson measure also shows much worse performance than LL. For a detailed comparison of the log-likelihood and chi-squared statistics, see [8].

A possible explanation for the fact that simple frequency performs in our case better than statistical dependency measures is as follows. A usual collocation is often a new term formed out of existing words, so it is more specific, and of more restricted use, then each of the two words separately. However, in our case the idiomatic P NP P combinations are equivalent to functional words—prepositions—which are of much more frequent use than the corresponding NP used in its literal meaning. Also, [5] suggests that frequent word chains of some specific POSs (such as N P N) tend to be terminological; our study can be considered as a particular case of this method.

On the other hand, [2] reports that idiomatic expressions combine with a more restricted number of neighboring words than free combinations. However, we observed the opposite effect: e.g., *a fin de* (literally 'at end of') in the sense 'in order to' combines with nearly any verb, while in its literal sense nearly only with nouns with semantics of time: *a fin de semana* 'at the end of the week'. The explanation for this fact can be the same as the one suggested in the previous paragraph.

Conclusions

Idiomatic word combinations of P NP P type, usually functioning as compound prepositions, have statistical properties distinct from those of usual idiomatic collocations. In particular, they combine with a greater number of words than usual idioms. For their unsupervised learning from a corpus, a simple frequency measure performs better than other statistical dependence measures. In particular, among most frequent P NP P word chains, about 50% are idiomatic. Inspection of the most frequent chains of this type permits to detect idiomatic combinations not present in existing dictionaries.

References

1. Banerjee, S., and Pedersen, T. The Design, Implementation, and Use of the Ngram Statistics Package. In: A. Gelbukh (ed.). Computational Linguistics and Intelligent Text Processing (CICLing-2003), Lecture Notes in Computer Science, N 2588, Springer, 2003; see http://www.d.umn.edu/~tpederse/nsp.html.
2. Degand, L., and Bestgen, Y. Towards automatic retrieval of idioms in French newspaper corpora. *Literary and Linguistic Computing* 18 (3), (2003) 249-259.
3. Evert, S., and Krenn, B. Methods for the Qualitative Evaluation of Lexical Association In *Proc. ACL-2001*, pp 188-195.
4. Galicia-Haro, S. N. Using Electronic Texts for an Annotated Corpus Building. In: *4th Mexican International Conference on Computer Science, ENC-2003*, Mexico, pp. 26–33.
5. Justeson, J. S., S. M. Katz. Technical Terminology: Some Linguistic properties and an algorithm for identification in text. Natural Language Engineering 1:9–27, 1995.
6. Manning, C. D. & Schutze, H. *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge, Massachusetts, 1999.
7. Nañez Fernández, E. *Diccionario de construcciones sintácticas del español. Preposiciones*. Editorial de la Universidad Autónoma de Madrid, 1995.
8. Rayson, P., Berridge, D., Francis, B. Extending the Cochran rule for the comparison of word frequencies between corpora. In Vol. II of Purnelle G. *et al*. (eds.) *Le poids des mots*: Proc. of 7th International Conf. on Statistical analysis of textual data (JADT 2004), Belgium, 2004, Presses universitaires de Louvain, pp. 926–936.