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).