Heuristics-Based Replenishment of Collocation Databases

Igor A. Bolshakov and Alexander Gelbukh

Center of Computer Research
Nacional Polytecnical Institute
Mexico City, Mexico
{igor,gelbukh}@cic.ipn.mx

Abstract. Collocations are defined as syntactically linked and semantically plausible combinations of content words. Since collocations constitute a bulk of common texts and depend on the language, creation of collocation databases (CBDs) is important. However, manual compilation of such databases is prohibitively expensive. We present heuristics for automatic generation of new Spanish collocations based on those already present in a CBD, with the help of WordNet-like thesaurus: If a word *A* is semantically "similar" to a word *B* and a collocation *B* + *C* is known, then *A* + *C* presumably is a collocation of the same type given certain conditions are met.

1 Introduction

Usual texts contain numerous word combinations like Spanish (Sp.) *prestar atención* 'pay attention'*, pluma fuente* 'fountain-pen'*, promesa falsa* 'false promise'*, partido político* 'political party', etc. We will refer to such syntactically linked and semantically plausible combinations of content words as collocations. We consider as collocations the following groups:

- idiomatic expressions like Sp. *medias tintas* 'half-measures' (literally (lit.) 'half inks') or *campo santo* 'graveyard' (lit. 'saint field'),
- expressions with standard lexical functions by Mel'čuk [5, 6, 10] like Sp. *café cargado* 'strong coffee' (lit. 'loaded coffee') or *prestar atención* 'to pay attention' (lit. 'to lend attention'), and
- free combinations like Sp. *sesión continua* 'continuous session' or *ver al joven* 'to see the lad,' that nevertheless joint only semantically combinable words.

Hence, our term *collocation* does not correspond to a mere co-occurrence of words within a short span of text [8]. It implies syntactic dependencies between the combined content words, immediate or through functional words (usually, prepositions); combined words could be linearly non-contiguous, even remote within a sentence. In this sense, the non-formal treatment of collocations in [3, 7] is just what we mean.

Since collocations usually depend on a given language and cover a bulk of texts, their knowledge is important in many areas of computational linguistics. Hence, creation of corresponding dictionaries [1] and collocation databases (CBDs) [2] is quite topical.

E.M. Ranchhod and N.J. Mamede (Eds.): PorTAL 2002, LNAI 2389, pp. 25-32, 2002.

By various reasons, high completeness of collocation collections (say, for the 95% coverage of collocations in a text) seems unreachable. First of all, the efforts and the necessary amount of statistical data for collecting word co-occurrences from a large text corpus greatly exceed those for separate words. Indeed, the less probable a co-occurrence, the longer and more diversified corpus is needed to guarantee statistically significant results. Hence, for compilation of a CDB of even a moderate coverage, it is necessary to scan through (with further manual control and post-editing) a huge and highly polythematic corpus. With such aggravations, it seems reasonable to try replenishment of existing basic CDBs not only through statistical accumulation and selection but also through heuristic inference of new collocations based on their already available amount.

In this paper we propose a method of replenishment of already rather large collocation collections by means of heuristic generation of new collocations. As components of these new collocations, content words already registered in a CDB are taken, and the semantic and syntactic links between them are analyzed. The general heuristic for such kind of inference can be formulated as follows: if a word *A* is semantically "similar" to a word *B* and a collocation *B* + *C* is already known then *A* + *C* is presumably a collocation of the same type given certain conditions are met. Some initial ideas of such replenishment are given in [2].

Semantic similarity between words is based on the relations used in WordNet-like databases [4, 9], specifically: synonymy, hyperonymy / hyponymy (genus—species relations), meronymy / holonymy (part—whole relations), and one type of semantic derivation. Additionally, one type of semantically induced morphological categories is considered for the inference, namely, grammatical number of nouns. We also describe some counter-heuristics for filtering out wrong or dubious results of the inference.

The rules and considerations proposed below are valid for all European languages, though we consider primarily Spanish examples since our motivation was the development of a corresponding Spanish-oriented CDB.

2 General Inference Rule

Our inference rules are similar to those of formal logic. Let a content word *A* have semantic "similarity" of a class **S** with a word *B* (we denote this as *A* **S** *B*), and a collocational link of a specific dependency category **D** combine the words *B* and *C* (we denote this as *B* **D** *C*). Then the hypothesis is advanced that *A* and *C* constitute a collocation of the same category **D**:

$$(A\ \mathbf{S}\ B)\ \&\ (B\ \mathbf{D}\ C) \Rightarrow (A\ \mathbf{D}\ C).$$

The dependency link **D** can be of several predetermined types and of any direction, i.e., by *B* **D** *C* we mean either a syntactic dependency $B \xrightarrow{\mathbf{D}} C$ or $B \xleftarrow{\mathbf{D}} C$. In both parts of the formula, the direction is assumed the same.

The term *inference* is used here in a rather figurative way. Namely, the results obtained with this formula are sometimes wrong, so that we have to use counter-heuristics (filters) to minimize the number of errors.

In practice, if the inference for a given word *A* gives a high percentage of correct (acknowledged by a native speaker) collocations with various words *C*, these collocations might be stored implicitly and generated at runtime only when they are really needed. Otherwise, when the number of correct collocations generated for *A* by the given formula is low, the few correct ones can, after the human verification, be incorporated explicitly into the CDB, thus directly replenishing it. In any case, the errors of inferences are a good tool for perfecting the counter-heuristics.

3 Synonymy-Based Inference

Suppose that the Sp. noun *gorra* 'cap' has no collocations in a CDB, however, it forms a part of a synonymy group (synset) {*gorro, gorra, cofia, capillo, casquete, tocado*} and *gorro* 'hat' is supplied with the collocations *ponerse* 'put on' */ quitarse* 'put off' */ llevar* 'wear'... *gorro*. We can conclude that this information can be transferred to all other members of the synset that lack complete characterization, among them to the word *gorra*. Thus, from the collocation *ponerse* (*el*) *gorro* 'put on the hat' we obtain *ponerse* (*la*) *gorra* 'put on the cap'.

To formalize these considerations, we should take into account that synsets can be introduced in various ways.

Synonymy as strict equivalence ignores any semantic differences between the members of a synset. If a word μ has no collocations of a given type **D** while some *Q* other members of the same synset do have them, then the required collocation counterparts (the second content word of the collocation) for μ is inferred as the intersection of the *Q* sets of such counterparts for the other words. I.e., for any *x* it holds

$$\forall_{q=1}^{Q} \left((\mu \; \textbf{\textit{HasSynonym}} \; s_q) \& (s_q \; \mathbf{D} \; x) \right) \Rightarrow (\mu \; \mathbf{D} \; x),$$

where ***HasSynonym*** is a kind of relation denoted earlier as **S**.

Synonymy with dominants supposes that there exists a dominant member within each synset expressing the common concept of the set in the most general and neutral way, like *gorro* 'hat' for the abovementioned synset. Then the inference is conducted in a different way. For any dominant, its collocations must be explicitly stored in the CDB, so that no inference is necessary for it. For a non-dominant word μ, the inference is made as follows:

- If μ belongs to only one synset $\{d, s_1, ... \mu, ..., s_N\}$ with the dominant *d* then any collocation valid for *d* is supposed to be valid for μ. I.e., for any *x* it holds

$$(\mu \; \textbf{\textit{HasDom}} \; d) \& (d \; \mathbf{D} \; x) \Rightarrow (\mu \; \mathbf{D} \; x),$$

  where ***HasDom*** is the relation standing for "has dominant synonym."

- If μ belongs to several synsets with different dominants D_q:

$$\{d_1, s_{11}, ..., \mu, ..., s_{1N_1}\} \; \{d_2, s_{21}, ..., \mu, ...s_{2N_2}\} ..., \{d_k, s_{k1}, ..., \mu, ..., s_{kN_k}\}$$

then those collocations are supposed valid for μ whose analogues (with d_i) are explicitly present in the CDB for all the dominants involved. This means that for any x it holds

$$\forall_{q=1}^{k} \left((\mu \; \textbf{\textit{HasDom}} \; d_q) \& (d_q \; \mathbf{D} \; x) \right) \Rightarrow (\mu \; \mathbf{D} \; x).$$

4 Inference Based on Hyponymy and Hyperonymy

Let the Sp. word *refresco* 'soft drink' have a vast collocation set in a CDB including collocations with verbs *echar* 'to pour' / *embotellar* 'to bottle' / *beber* 'to drink' ... *refresco*. The same information on *Pepsi* may be absent in the CDB, while *Pepsi* is registered (through ***IsA*** relation) as a hyponym of *refresco*. The inference assigns the information about the hyperonym to all its hyponyms that lack relevant type of collocations. Thus, it is inferred that all mentioned verbs are applicable to *Pepsi*.

Here we should bear in mind that the ***IsA*** relations can be structured in various ways.

Monohierarchy presupposes a unique classification hierarchy (tree) relating some content words within the CDB. Thus, a unique hyperonym corresponds to each hyponym in it, except for the uppermost node (the root of the tree) and the words not included in the hierarchy.

Suppose the relation $\textbf{\textit{IsA}}^1$ (^{1} stands for 'one-step') gives the immediate (nearest) hyperonym h_1 for the word μ. The word h_1 is unique (if it exists) within a monohierarchy. If the collocation set for h_1 is empty, then a hyperonym h_k (of a higher level) is used, where $\textbf{\textit{IsA}}^k$ and h_k are defined as

$$(\mu \; \textbf{\textit{IsA}}^k \; h_k) = (\mu \; \textbf{\textit{IsA}}^1 \; h_1) \; \& \; (h_1 \; \textbf{\textit{IsA}}^1 \; h_2) \; \& \; ... \; \& \; (h_{k-1} \; \textbf{\textit{IsA}}^1 \; h_k).$$

Inference by means of hyperonymy determines the first (i.e., with the minimal k) hyperonym h_k that has non-empty collocation set of the required type and assigns these collocations to μ. I.e., for any x it holds

$$(\mu \; \textbf{\textit{IsA}}^k \; h_k) \& (h_k \; \mathbf{D} \; x) \Rightarrow (\mu \; \mathbf{D} \; x).$$

For example, if the relevant collocation set for *refresco* is empty, while for *bebida* 'a drink' is nonempty then *Pepsi* is characterized by the information borrowed from this two-step-up hyperonym.

Crosshierarchy permits content words to participate in more than one hyperonym—hyponym hierarchy based on different principles of classification. The whole structure is a directed acyclic graph; a word (except for uppermost nodes of partial hierarchies and unclassified words) can have several hyperonyms. There are various cases of replenishment in this case.

Since more then one path can go up from a word μ, we propose the following inference procedure. All k-step-up hyponyms of μ, k = 1, 2, ..., are searched width-first, until at least one of them has a non-empty collocation set. If there is only one non-empty set at such k-th layer, the formula given above remains valid. Otherwise the intersection of all non-empty sets is used. To represent this mathematically, we enu-

merate all hyponyms with non-empty collocation sets of k-th layer as q = 1, 2, ..., Q. Then for any x it holds

$$\forall_{q=1}^{Q} \left(\left(\mu \; \boldsymbol{IsA}^{k} \; h_{kq} \right) \& \left(h_{kq} \; \mathbf{D} \; x \right) \right) \Rightarrow \left(\mu \; \mathbf{D} \; x \right).$$

The width-first search excludes the cases when a collocation set of k-th hyperonym is taken while m-th hyperonym has non-empty set, $m < k$.

5 Inference Based on Holohymy and Meronymy

The meronymy relation (x ***HasMero*** y) holds when x has y as a part; holonymy (y ***HasHolo*** x) is the inverse relation holding when x includes y. In simple cases, both x and y are single words, e.g., Sp. (*clientela* 'clientage' ***HasMero*** *cliente* 'client') or (*árbol* 'tree' ***HasMero*** *tronco* 'trunk').

Unlike synonymy and hyperonymy, both directions of collocation transfer are possible. E.g., Sp. collocations *atender* 'to attend (to)' / *satisfacer* 'to satisfy' / *atraer* 'to attract' / *perder* 'to lose' ... *al cliente* are equally applicable to *clientela* and vice versa. The general inference rule is as follows: for any x it holds

$$(\mu \; \boldsymbol{HasMero} \; y) \; \& \; (y \; \mathbf{D} \; x) \Rightarrow (\mu \; \mathbf{D} \; x),$$
$$(\mu \; \boldsymbol{HasHolo} \; y) \; \& \; (y \; \mathbf{D} \; x) \Rightarrow (\mu \; \mathbf{D} \; x).$$

In fact, not all combinations of μ and y can be used in these formulas. The complications are implied by the fact that meronymy / holonymy can be of five different types [9]: a part proper: *dedo* 'finger' of *mano* 'hand'; a portion: *gota* 'drop' of *líquido* 'liquid'; a narrower location: *centro* 'center' of *ciudad* 'city'; a member: *jugador* 'player' of *equipo* 'team'; and a substance the whole is made of: *madera* 'wood' for *bastón* 'stick'. Not all these types are equally suitable for bi-directional inferences. The membership type seems the most facilitating, while the made-of type the least facilitating. In any case, further research is necessary to develop the filters to eliminate wrong combinations.

6 Inference Based on Semantic Derivation

Semantic derivation is the formation of new lexemes from other lexemes close in meaning. There exists at least one kind of semantic derivation in European languages that excellently suits for the inferences, namely derivation of adverbs from a corresponding adjective: Sp. *bueno* 'good' → *bien* 'well', *corto* 'short' → *cortamente* 'shortly'. The modifiers for an adjective and an adverb constituting a derivative pair are the same:

$$(\mu \; \boldsymbol{IsDerivFrom} \; y) \; \& \; (y \; \boldsymbol{HasModif} \; x) \Rightarrow (\mu \; \boldsymbol{HasModif} \; x).$$

E.g., Sp. (*bien* ***IsDerivFrom*** *bueno*) & (*bueno* ***HasModif*** *muy* 'very') ⇒ (*bien* ***HasModif*** *muy*).

7 Morphology-Based Inference

Some morphological categories are semantically induced and have explicit expression at the semantic level of text representation. The most common of these categories is grammatical number that in most languages has two values: singular and plural. Since different number values frequently imply different collocation sets, singular and plural words should be included into the CDB as separate headwords, as it has been proposed in [2]. However, if the CDB contains a collocation set of a given relation **D** for only one (supposedly more frequently used) value of number while the collocations for its counterpart are still absent in the CDB then the same set can be transferred to the word of the complementary value of number. I.e. for any *x* it holds

$$(\mu\ \textbf{\textit{HasComplementaryNumber}}\ y)\ \&\ (y\ \mathbf{D}\ x) \Rightarrow (\mu\ \mathbf{D}\ x).$$

E.g., modifiers of Sp. *información* 'information' are applicable for its more rarely used plural *informaciones*.

However, some specific modifiers frequently bring in wrong hypotheses. To filter them out, it should be taken into account that, at semantic level, singular of a noun *N* is usually opposed to plural through the predicates ***Single***(*N*) *vs.* ***MoreThanOne***(*N*). Hence, we can introduce a prohibitive list of words with a semantic element of singularity / uniqueness, thus combinable mainly with singular, and, vice versa, a list of words with a semantic element of collectiveness / multiplicity, thus combinable mainly with plural. For ***HasModif*** relation in Spanish, following are mainly plural-oriented modifiers: *muchos* 'many', *múltiples* 'multiple', *numerosos* 'numerous', *varios* 'several', *diferentes* 'different', *diversos* 'various', *distintos* 'distinct (ones)', *idénticos* 'identical', *iguales* 'equal (ones)', *desiguales* 'unequal (ones)', *de todos tipos* 'of all kinds', etc. Singular-oriented modifiers are fewer: *único* 'unique', *singular* 'single', *solitario* 'alone', *individual* 'individual', etc. The counter-heuristic (filtering) rule consists in the use of the corresponding heuristic rule only if *x* does not belong to the proper prohibitive list.

8 Filtering Out Wrong Hypotheses

We have already mentioned some counter-heuristics we use to filter out inference errors. Following are other filters we use.

Do not consider lexeme-specific syntactic dependencies D. The most error-prone of such dependencies is ***HasValency*** of verbs (= subcategorization frame). To illustrate this, consider the Spanish synset {*elegir* 'to choose, to elect', *seleccionar* 'to select', *designar* 'to assign', *optar* 'to opt', *preferir* 'to prefer', *escoger* 'to choose'}. All its members except for *optar* subcategorize for the target of selection as direct object (*elegir / seleccionar / designar / preferir / escoger algo* 'smth.'), while *optar* uses a prepositional phrase for this purpose: *optar por algo* 'to opt for smth.'. Each of them, except for *elegir* and *designar*, can introduce a prepositional complement with *entre* 'between' for options of the selection. Thus, collocations including ***HasValency*** relation, say, of the verb *optar* cannot be inferred correctly based on the properties of the other members of the synset.

Meanwhile, the dependencies inverse to ***HasValency*** can be freely used for the inferences involving nouns. Indeed, if the relation ***HasValency*** gives for *país* 'country' the collocations *traspasar* 'to cross' / *visitar* 'to visit' / *arruinar* 'to destroy' ... *país* then all these verbs form correct collocations with any specific country name (e.g., *Portugal*).

Ignore classifying modifiers. Some inferences for modificatory collocations also give wrong results. For example, *bayas* 'berries' as a hyperonym can have nearly any color, smell, and taste, while its hyponym *arándanos* 'blueberries' are scarcely *amarillas* 'yellow'.

Among modifiers, one rather broad class is most error-prone. Consider the wrong inference:

(*Colombia* ***IsA*** *país*) & (*país* ***HasModif*** *europeo*) ⇒ *(*Colombia* ***HasModif*** *europea*)

'Colombia is a country' & 'European country' ⇒ 'European Colombia'. To exclude such cases, we ignore in inferences the so-called classifying modifiers. These modifiers convert a specific noun to its hyponym, e.g., *país* 'country' → *país europeo* 'European' / *americano* 'American' / *africano* 'African'... As to other modifiers for *país*, they frequently give correct results: *agrícola* 'agricultural' / *bella* 'beautiful' / *gran* 'great' ... *Colombia*.

Unfortunately, even such filtering is rarely sufficient. E.g., the modifiers *del sur* 'Southern' or *del norte* 'Northern' change their meaning while moving from *país* to a specific country name, compare *país del norte* 'Northern country' and *Portugal del norte* 'Northern Portugal'. All such cases are to be dealt with using specific word lists.

Ignore marked words and collocations. In some printed and electronic dictionaries oriented to humans, the number of various usage marks reaches several tens. All these labels can be divided into two large groups:

- ***Scope of usage***, such as *special*, *obsolete*, *vulgar*, etc., and
- ***Idiomacity***, covering cases of figurative and direct interpretation (Sp. *rayo de luz* 'good idea' or 'light beam', lit. ray of light) as well as of idiomatic use only (Sp. *vara alta* 'power, high influence', lit. 'high stick').

We propose to ignore all marked items in inference, since they more frequently than not lead to wrong or dubious results. For example, Sp. *vara* 'stick' is dominant of the synset {*vara*, *varejón* 'long stick', *varal, garrocha, palo* 'thick stick'}, but we cannot correctly semantically interpret the results of the inference (*palo* ***HasDom*** *vara*) & (*vara* ***HasModif*** *alta*)$_{\text{idiom}}$ ⇒ *(*palo* ***HasModif*** *alto*)$_{\text{idiom}}$.

9 Conclusions

A heuristic method is developed for inferring new collocations basing on their available amount and a WordNet-like thesaurus. The main idea of the method consists in the possibility, except for some cases, to substitute a word in the existing collocation

with a semantically "similar" one. We have discussed what we mean by similarity and what exceptions there are.

The types of similarity considered are synonymy, hyperonymy / holonymy, holonymy / meronymy, one kind of semantic derivation, and one semantically induced morphological category: number of nouns.

To exclude the most frequent errors, some counter-heuristics have been proposed in the form of prohibitive lists or categories of content words, lexeme-specific syntactic relations used in collocations, and the labeled dictionary items.

Using the proposed inference procedure, runtime or off-line replenishment of large collocation databases can be performed. All introduced rules are well suitable to develop a Spanish collocation DB that is now under development (ca. 10,000 collocations and semantic links). In the future, we plan to reach amounts comparable to those reported for the analogous Russian database [2]. Using of inference will facilitate this task.

References

1. Benson, M., E. Benson, and R. Ilson. *The BBI Combinatory Dictionary of English*. John Benjamin, Amsterdam / Philadelphia, 1989.
2. Bolshakov, I. A., A. Gelbukh. *A Very Large Database of Collocations and Semantic Links*. In: Mokrane Bouzeghoub et al. (eds.) *Natural Language Processing and Information Systems.* 5th International Conference on Applications NLDB-2000, Versailles, France, June 2000. Lecture Notes in Computer Science No. 1959, Springer, 2001, p. 103-114.
3. Calzolari, N., R. Bindi. *Acquisition of Lexical Information from a Large Textual Italian Corpus*. Proc. of COLING-90, Helsinki, 1990.
4. Fellbaum, Ch. (ed.) *WordNet: An Electronic Lexical Database.* MIT Press, Cambridge, London, 1998.
5. Mel'čuk, Igor. *Fraseología y diccionario en la lingüística moderna.* In: I. Uzcanga Vivar et al. (eds.) *Presencia y renovación de la lingüística francesa*. Salamanca: Ediciones Universidad, 2001, p. 267-310.
6. Mel'čuk, I., A. Zholkovsky. *The explanatory combinatorial dictionary*. In: M. Evens (ed.) *Relational models of lexicon*. Cambridge University Press. Cambridge. England, 1988, p. 41-74.
7. Satoshi Sekine et al. *Automatic Learning for Semantic Collocation.* Proc. 3rd Conf. Applied Natural Language Processing, Trento, Italy, 1992, p. 104-110.Smadja, F. *Retreiving collocations from text: Xtract*. Computational Linguistics. Vol. 19, No. 1, 1991, p. 143-177.
8. Smadja, F. *Retreiving collocations from text: Xtract*. Computational Linguistics. Vol. 19, No. 1, 1991, p. 143-177.
9. Vossen, P. (ed.). *EuroWordNet General Document*. Vers. 3 final. 2000, www.hum.uva.nl/~ewn.
10. Wanner, Leo (ed.). *Lexical Functions in Lexicography and Natural Language Processing*. Studies in Language Companion Series, ser.31. John Benjamin, Amsterdam/ Philadelphia, 1996.