YET ANOTHER APPLICATION OF INFERENCE IN COMPUTATIONAL LINGUISTICS [*]

IGOR A. BOLSHAKOV and ALEXANDER GELBUKH

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

Abstract

Texts in natural languages consist of words that are syntactically linked and semantically combinable—like *political party*, *pay attention*, or *brick wall*. Such semantically plausible combinations of two content words, which we hereafter refer to as collocations, are important knowledge in many areas of computational linguistics. We consider a lexical resource that provides such knowledge—a collocation database (CBD). Since such databases cannot be complete under any reasonable compilation procedure, we consider heuristic-based inference mechanisms that predict new plausible collocations based on the ones present in the CDB, with the help of a WordNet-like thesaurus. If *A B* is an available collocation and *B* is 'similar' to *C*, then *A C* is supposedly a collocation of the same category. Also, we touch upon semantically induced morphological categories suiting for such inferences. Several heuristics for filtering out wrong hypotheses are also given and the experience in inferences obtained with CrossLexica CDB is briefly discussed.

Keywords

Collocations, inference rules, enrichment, synonyms, hyperonyms, meronyms.

1 Introduction

Natural languages texts, at least usual ones, contain many syntactically linked and semantically plausible combinations of content words like *to pay attention, proposal concerns, to think of welfare, false promise, political party, brick wall*, etc. We oppose them to senseless combinations like **green ideas* impossible in usual texts, as well as to combinations including non-content words like *is growing* or *she went*, quite usual in texts.

Since the collocations heavily depend on a given language and constitute a great part of usual texts, they are important knowledge in many areas of computational linguistics, e.g., for resolution of lexical and syntactic ambiguities. Hence, creation of the lexical resource of a new type—collocation database (CBD)—proved to be quite topical.

The main performance parameter of a CBD is its completeness expressed in the percentage of covering collocations in arbitrary texts. The efforts for collecting word co-occurrences through a text corpus significantly exceed those for separate words. For compilation of CDB even of a moderate completeness, say, 60 or 70%, it would be necessary to automatically scan through—with further manual control and post-editing—a huge (of many gigabytes) and highly polythematic corpus, at expense of a tremendous labor.

In this paper, we propose a method of replenishment of already existing and rather large CDB by means of automatic inference (generation) of new collocations. Components of these collocations are content words already registered within a CDB, whereas various types of semantic similarity between words are considered as a tool for the inference. Such a similarity can be diagnosed by a WordNet-like thesaurus [6, 9, 10], which can be attached to CDB.

The generalized inference rule is taken of production type well known in Artificial Intelligence. It signifies that, given a content word *A* having semantic similarity of a class **S** to a word *B* and a collocational link **D** of a specific dependency category combining the words *B* and *C,* the inference hypothesis is that *A* and *C* constitute a collocation of the same category **D.** Hence, the general inference rule is very similar to the *modus ponens* of formal logic:

$$(A\ \mathbf{S}\ B)\ \&\ (B\ \mathbf{D}\ C) \Rightarrow (A\ \mathbf{D}\ C). \qquad (1)$$

We use the term ***inference*** for the generation of new collocations, though the rule (1) is only a heuristic, and results of the inference can be sometimes wrong (see later).

When the inference for a word *A* gives rather high percentage of correct collocations with various *C*, the resulting collocations might be stored implic-

[*] Work done under partial support of CONACyT, SNI, and CGEPI-IPN. Mexico.

0-7803-7087-2/01/$10.00 © 2001 IEEE

Igor A. Bolshakov and Alexander Gelbukh. *Yet another application of inference in computational linguistics*. IEEE International Workshop on Natural Language Processing and Knowledge Engineering, NLPKE 2001, IEEE, pp. 1688–1692; www.Gelbukh.com/CV.

itly and be generated at the user's screen only when they are really queried—at runtime.

Otherwise, when the correct collocations with different *C* generated for some *A* are few, they might be incorporated into the immanent part of the given CDB after native speaker's filtering out wrong inferences, thus directly replenishing CDB. In any case, the errors of such inferences are very instructive for further research.

Below, we enumerate firstly several types of collocation most common for European languages, with English examples. This is instantiation of **D**-relation in the rule (1). Then we study the inferences for various types of semantic similarity, thus instantiating **S**-relation in the same rule. Each type is supplied with illustrative examples. Then the heuristic tools for rejecting the main types of wrong inferences are briefly given. At last, we outline an experience with CrossLexica system [3, 4], which are relevant to the topic. CrossLexica is now the unique database that gathers collocations and semantic links, and thus is the only object for immediate application of the method under consideration.

2 Specific types of collocations

There exist several specific relations that instantiate **D**-relation in the rule (1). These relations determine corresponding types of collocations. The collocation types most important in European languages are the following [1, 2, 5, 8, 11].

The direct modificarory relation *HasModifier* joins a given noun, adjective or verb with its modifier, an adjective or an adverb. Thus we have collocation subtypes: (Noun *HasModifier* Adj), (Verb *HasModifier* Adv), (Adj *HasModifier* Adv), and (Adv *HasModifier* Adv). Examples: (*act HasModifier criminal*) and (*prepare HasModifier readily*).

The inverse modificatory relation *IsModifierOf* determines collocations: (Adj *IsModifierOf* Noun), (Adv *IsModifierOf* Verb), (Adv *IsModifierOf* Adj), and (Adv *IsModifierOf* Adv). Examples: (*national IsModifierOf economy*), (*very IsModifierOf quickly*), (*rather IsModifierOf well*).

In (Noun *IsSubjectOf* Verb) collocations, Noun is grammatical subject and Verb is its grammatical predicate. E.g., (*heart IsSubjectOf sink*) reflects the collocation *heart sinks* in a text.

In (Noun *IsNounObjOf* Verb) collocations, Noun is object of Verb: direct, indirect or prepositional one. Examples: *shake hands*, *arrange* (*with*) *enemy*, etc.

In (Verb *IsVerbObjOf* Verb) collocations, one verb in infinitive is subordinated to another verb: *prepare* (*to*) *sleep*.

In (Noun *IsNounComplOf* Noun) collocations, one noun is subordinated to another one: *adjustment* (*of a*) *clock*.

In (Verb *IsVerbCompl* Noun) collocations, Noun rules Verb in infinitive: *readiness* (*to*) *use*.

Government patterns determine a general scheme of collocations, in which a given word rules other words (usually nouns) as its valencies. In CBD, the patterns contain also the lists of collocations for each specific pattern. In the case of verbs, these are instances of their subcategorization frames. For example, the verb *give* has the pattern *who/what gives?* with examples of its dependents *boy, father, government...;* the pattern *what is given?* with examples *hand, money, book...reach*; and the pattern *to whom is given?* with corresponding examples.

GovPattern is the inverse relation to the set of relations *IsSubjectOf, IsNounObjOf, IsVerbObjOf, IsNounComplOf, IsVerbComplOf,* as well as to analogical relations for adjectives and adverbs.

3 Synonymy-based inference

Suppose that the noun *coating* has no collocations in CDB, but it is registered here in to the synonymy group whose member *layer* is supplied by collocations. The inference changes *layer* in its collocations to its synonym lacking the complete characterization. Thus, starting from the collocation *to cover with a layer*, the collocation *to cover with a coating* is inferred.

In mathematical sense, synonymy is **equivalence,** if we ignore all differences between synonyms $\{s_1, ..., s_N\}$ in the group. If the synonym μ has no collocations of the given type **D**, while *Q* other members of the same group do have them, then the lacked collocations for μ can be inferred as intersection of Q collocation sets, i.e. for any *x*,

$$\overset{Q}{\underset{q=1}{\forall}} \left(\left(\mu\ \textbf{\textit{HasSyn}}\ s_q\right) \& \left(s_q\ \mathbf{D}\ x\right)\right) \Rightarrow \left(\mu\ \mathbf{D}\ x\right) \quad . \qquad (2)$$

If there is a **dominant** synonym *D* expressing the group concept in a rather general and neural way, then there are two options of inference for non-dominants. If μ belongs only to one group $\{D, s_1, ...\ \mu, ... s_N\}$, any collocation valid for *D* is supposed valid for μ, i.e., for any *x*,

$$\left(\mu\ \textbf{\textit{HasDom}}\ D\right) \& \left(D\ \mathbf{D}\ x\right) \Rightarrow \left(\mu\ \mathbf{D}\ x\right) \qquad (3)$$

If μ belongs to *k* groups with dominants D_q, those collocations are supposed valid for μ whose analogues are registered for all dominants:

$$\overset{k}{\underset{q=1}{\forall}} \left(\left(\mu\ \textbf{\textit{HasDom}}\ D_q\right) \& \left(D_q\ \mathbf{D}\ x\right)\right) \Rightarrow \left(\mu\ \mathbf{D}\ x\right) \qquad (4)$$

4 Hyperonymy-based inference

Let the term *refreshing drink* have the complete collocation set in CDB, with the verbs *to bottle, to drink, to pour...* constituting some type of them. The same data on *Coca Cola* is absent in the CDB. It is only known that it is hyponyms of *refreshing drink*. The inference transfers the information connected with the hyperonym to all its hyponyms lacking same type of collocations. Thus, it is inferred that the mentioned verbs are applicable to *Coca Cola* too.

Case of **monohierarchy** presupposes a unique hyponyms-hyperonym hierarchy uniting content words within given CDB. A unique hyperonym corresponds to each hyponym in it. Suppose that the immediate hyperonym for μ is h_1, while k-th one (k = 2,3...) is h_k. Then the inference by means of hyperonymy attains the first met hyperonym h_k with a non-empty collocation set and assign these collocations to μ: i.e., for any x,

$$\left(\mu\ \boldsymbol{IsA}^k\ h_k\right) \& \left(h_k\ \mathbf{D}\ x\right) \Rightarrow \left(\mu\ \mathbf{D}\ x\right) \qquad (5)$$

Case of **crosshierarchy** presupposes participation of content words in one or more hyperonym-hyponym hierarchies based on different principles of classification. For example, *refrigerator* can participate in the hierarchy of home appliances as well as of electrical devices. Since more then one path can go up from a CDB word μ, the inference procedure is to search widthwise all k-th hyponyms of μ, k = 1, 2,..., until at least one of them has a non-empty collocation set. If there is only one non-empty set at k-th layer, the formula (5) remains valid. Otherwise the intersection of Q non-empty sets is taken

$$\overset{Q}{\underset{q=1}{\forall}} \left(\left(\mu\ \boldsymbol{IsA}^k\ h_{k_q}\right) \& \left(h_{k_q}\ \mathbf{D}\ x\right)\right) \Rightarrow \left(\mu\ \mathbf{D}\ x\right) \qquad (6)$$

5 Meronymy/holohymy-based inference

The meronymy relation (*A HasMero B*) states that *A* has *B* as a part, whereas holonymy (*B HasHolo A*) is inverse relation: *B* is a part of *A*. In simple cases, both *A* and *B* are single words in a given language, like (*clientele HasMero client*) or (*tree HasMero trunk*) in English.

In contradistinction with synonymy and hyperonymy, one can imagine the transferring of collocations in both directions. E.g., the collocations *to serve / satisfy / draw in / lose... a client* are equally applicable to *clientele* and, vice versa, nearly all collocations valid for *clientele* are valid to *client* too. That is the inference rules are, for any *x*,

$$(\mu\ \boldsymbol{HasMero}\ y)\ \&\ (y\ \mathbf{D}\ x) \Rightarrow (\mu\ \mathbf{D}\ x), \qquad (7)$$

$$(\mu\ \boldsymbol{HasHolo}\ y)\ \&\ (y\ \mathbf{D}\ x) \Rightarrow (\mu\ \mathbf{D}\ x). \qquad (8)$$

In fact, not all *x* in the formulas (7) and (8) can be taken, since there exist some complications in the case of meronymy/holonymy. For instance, it is known [10] that meronymy/holonymy can be of five different types: (1) a part proper, like *finger* of *hand*, (2) a portion, like *drop* of *liquid*; (3) a narrower location, like *center* of *city*; (4) a member, like *player* of *team*; (5) a substance the whole is made of, like *stick* of *wood*. Thus, liability of the inference for this type of semantic similarity requires additional studies.

6 Morphology-based inference

Some morphological categories are semantically induced, i.e. they have their own representation on semantic level. Such categories can be used for inferences too.

In all European languages, nouns have semantically induced category of **number** with singular and plural values. Since these values frequently imply different collocation sets, they should be included into a CDB separately. A version of CDB can contain a collocation subset for only one, more frequent, value. For example, a CDB can contain collocation for *difficulty* but not for *difficulties,* or vice versa.

In such cases, the same set can be assigned to the supplementary value:

$$(\mu\ \boldsymbol{HasSupplNum}\ y)\ \&\ (y\ \mathbf{D}\ x) \Rightarrow (\mu\ \mathbf{D}\ x). \quad (9)$$

Another grammatical category suited for inferences is ***aspect*** of Slavic verbs.

7 Several precautions while inference

To avoid some frequent inference errors, several precaution measures could be taken while inference.

Some types of collocations, e.g., those based on government patterns of verbs, cannon be taken in general as a source of inferences. For example, English verbs *to choose, to select, to pick, to cull, to elect, to opt, to single out* are synonyms, but the government pattern, say, of *to opt* cannon be inferred correctly based on data of any other of its synonyms.

Note that dependencies inverse to government patterns can be freely used for the inferences.

The so-called classifying modifiers that convert a specific notion to its hyponym, e.g., *country* to *European country* or *American country*, should not be used for inferences too, else we can get semantically wrong collocations as *European Argentina.*

In any dictionary, there are labeled words and collocations indicating nonstandard use (special, bookish or obsolete character, colloquialism, vulgarism, etc.) or idiomaticy. Preliminary studies have shown that to use any labeled elements for inferences is incautious. E.g., we cannot derive any collocations from *hot dog*, since **hot poodle* or **hot spaniel* are ridiculously wrong

Prohibitive lists of words are to be compiled, separately for each type of similarities and collocational types. E.g., for transferring of modificatory collocations from plural to singular it is reasonable to exclude plural-oriented modifiers *many, multiple, numerous, various, different, diverse, equal, unequal, of all kinds*... To the inverse direction, the following singular-oriented adjectives are to be excluded: *unique, single, solitary, lonely, individual*...

8 Experience with Crosslexica

The CrossLexica collocation database was developed in the 90s [3, 4] with Russian as the basic language and English only for queries. Its proportions can be characterized by the following statistics of collocations (measured in unilateral links):

Modificatory collocations	615,600
Verbs *vs*. their noun complements	348,400
Nouns *vs*. their predicates (verbs or short-form adjectives)	235,400
Nouns *vs*. their noun objects	216,800
Verbs *vs*. their infinitive objects	21,500
Nouns *vs*. their infinitive complements	10,800
Total	**1,448,500**

It is worth to mention that the mean collocational fertility proved to be a rather constant value. For example, a noun can be object in average of ca. 24 verbs (different aspects are considered as different verbs). This value did not change during five recent years of the version renewal. This shows that the so-called free collocations, which were gathered to CrossLexica, are nevertheless heavily constrained from semantic viewpoint.

CrossLexica also contains semantic relations of WordNet type. Among them, the following are relevant for this paper: synonyms 193,900; holonyms / meronyms 17,300; hyponyms / hyperonyms 8,500; totally 219,700. Synonyms are 39% nouns, 28% are verbs, 22% are adjectives, and 11% are adverbs. The number of unilateral links is counted as $\Sigma_i n_i(n_i-1)$, where n_i counts i-th synonymy group considered with a dominant. Hyponyms and hyperonyms are only nouns and form a crosshierarchy.

For inferences, synonymy, hyperonymy, and morphological number were used. Let us see the results of the inferences for the rather rarely used word *koka-kola* 'Coca Cola' in the earlier version of CrossLexica. The database contains only its hyperonyms: (*Coca Cola IsA*[1] *refreshing drink*), (*refreshing drink IsA*[1] *drink*), so that all collocations inferred are based on *drinks* (*refreshing drink* has no collocations in the current version). The statistics of correct inference were as follows: for modificatory collocations 10%; for predicative ones 93%; for verbal complements 100%, and for substantive complements 94%.

So pour results for modificatory collocations are explained by that the earlier version did not used some abovementioned ideas of filtering out. For example, since *alkogol'naja* and *spirtnaja* modifiers (both 'alcoholic') are classificational, and they were moved to other place in the revised revision; the modifiers *razlichnaja* and *raznoobraznaja* 'various / different' are plural-oriented and were filtered out while inferring on this reason, etc. Thus, the statistics of correct generation of modificatory collocations became much better (83%) in the last version.

An approximate evaluation of few consequent versions has shown that the global portion of inferred collocations was always less than 8% of the total CDB, and more than 3% gave so high percentage of wrong collocations that the generated subcollections were fully revised by hand and then inserted to the CDB as its immanent part.

The process of version revision comes in parallel with detailed characterization of words occurred in texts more and more rarely, and some potential 'clients' of the inference thus disappear. In this way, the total portion of inferred collocations diminishes. However, the expansion and the perfection of the synonyms and the crosshierarchy act to the opposite direction, thus conserving the inferred part and necessity in the inferences.

9 Conclusions

A method is developed of generating new collocations based on an available—already large—collocation database and a set of semantic relations concerning one component of a source collocation. In the target collocations, one component is changed to semantically similar one. Semantic similarity is supposed to be determined by synonymy, hyperonymy, holonymy, and semantically induced morphological categories.

The enrichment is performed by means of inference rules similar to deduction formulas of mathematical logic. With any semantic similarity including generic terms, the inference rules remain nevertheless heuristics, and the 100-percent correctness of results remained unreachable.

In order to improve results, several precautionary heuristics are proposed, i.e., prohibitive subtypes and word lists are introduced. In the prototype system

CrossLexica, generated collocations are always given at the screen with marks of their tentativeness.

On the contrary, the inferences proved to be quite opportune for semiautomatic replenishing of the database with collocation containing infrequent words not yet fully described in the current version of the database.

In fact, computational linguistics could not manage the replenishing of collocation databases without the automatic or semi-automatic generation of new collocations, even if the use of inferred collocations is rather marginal.

References

1. Apresjan, Ju. Systematic lexicography. Oxford Univ. Press, 2000.
2. Benson, M., et al. The BBI Combinatory Dictionary of English. John Benjamin Publ., Amsterdam, Philadelphia, 1989.
3. Bolshakov, I.A. Multifunction thesaurus for Russian word processing. Proceedings of 4th Conference on Applied Natural language Processing, Stuttgart, 13-15 October 1994, p. 200-202.
4. Bol'shakov, I.A. Multifunctional thesaurus for computerized preparation of Russian texts. Automatic Documentation and Mathematical Linguistics. Allerton Press Inc. Vol. 28, No. 1, 1994, p. 13-28.
5. Calzolari, N., R. Bindi. Acquisition of Lexical Information from a Large Textual Italian Corpus. Proc. of COLING-90, Helsinki, 1990.
6. Fellbaum, Ch. (Ed.) WordNet: An Electronic Lexical Database. MIT Press, Cambridge / London, 1998.
7. Mel'cuk, I. Dependency Syntax: Theory and Practice. SUNY Press, NY, 1988.
8. Mel'cuk, I., Zholkovsky, A. The explanatory combinatorial diccionary. In: M. Evens (Ed.). Relational models of lexicon. p.41-74. Cambridge Univ.Press, NY, 1988.
9. The Spanish WordNet. Version 1.0, July 1999. EuroWordNet, LE2-4003 & LE4-8328. CD ROM.
10. Vossen, P. (Ed.). EuroWordNet General Document. Vers. 3 final. 2000, www.hum.uva.nl/~ewn.
11. Wanner, L. (Ed.) Lexical Functions in Lexicography and Natural Language Processing. Studies in Language Companion Series N. 31. Benjamin Publ., Amsterdam, Philadelphia, 1996.