Extending WordNet with Syntagmatic Information

Luisa Bentivogli and Emanuele Pianta

ITC-irst, Via Sommarive 18
38050 Povo - Trento, Italy
{bentivo,pianta}@itc.it

Abstract. In this paper we present a proposal to extend WordNet-like lexical databases by adding information about the co-occurrence of word meanings in texts. More specifically we propose to add *phrasets*, i.e. sets of free combinations of words which are recurrently used to express a concept (let's call them *Recurrent Free Phrases*). Phrasets are a useful source of information for different NLP tasks, and particularly in a multilingual environment to manage lexical gaps. At least a part of recurrent free phrases can also be represented through a new set of *syntagmantic* (lexical and semantic) WordNet relations.

1 Introduction

Most lexical information encoded in WordNet has a paradigmatic nature, that is if we take a word from a sentence in a real text, and consider which semantic and lexical relations are coded in WordNet with regard to that word, we will see that all relations hold between that word and other words that most probably do not occur in the same sentence or text. In Saussurean terms [5], paradigmatic relations occur *in absentia*, i.e. they hold with words that could in principle *substitute* each other rather than co-occur. On the other side, syntagmatic relations are *in praesentia*: they hold between words co-occurring in the same text. Syntactic relations are the best known kind of syntagmatic relations between words, whereas selectional restrictions between a verb and its arguments are a typical example of semantic syntagmatic relations [4].

As a matter of fact, information about the co-occurrence of words is not completely missing in WordNet. One can find such information in synonyms formed by more than one word and in gloss examples. However, WordNet includes only more-than-one-word synonyms that are elementary lexical meaning units, so they give information about the co-occurrence of words but not about the co-occurrence of meanings (as a more-than-one-word synonym involves only one meaning). On the other side, the information about the co-occurrence of words encoded in examples is not made explicit, and is out of the WordNet relational model.

In spite of the secondary role that syntagmatic relations play in WordNet, they are as relevant as paradigmatic relations both from a lexicographic and computational point of view. To have an idea of their lexicographic relevance, one only need to have a look at the space that examples of usage take in any dictionary entry, and it is every language learner's experience that an example of usage can be more useful than any

definition to the comprehension of a word meaning. From a computational point of view, information about the co-occurrence of words is the most crucial, and in many cases, the only kind of information which is used for many NLP tasks. This is more and more true given the increasing role of statistics oriented, corpus based methods. In fact, co-occurrence is the most simple and effective kind of information that can be extracted from texts. A distinction needs to be done here between co-occurrence of words and co-occurrence of meanings. Whereas the former kind of information is indeed easily available in texts, the latter is much harder to be extracted, as it requires the disambiguation of texts. For this reason the encoding of information about the co-occurrence of meanings in a lexical resource as WordNet could be highly beneficial to the NLP community.

In the rest of this paper we will constrain the type of meaning co-occurrence information that we think should be encoded in WordNet. More specifically, in Section 2 we will concentrate on a set of expressions that we call Recurrent Free Phrases (RFPs). Then, in Sections 3 and 4 we will present two strategies to encode RFPs in WordNet. The first is based on a new data structure called *phraset*; the second is based on a new set of lexical and semantic relations. Finally, in Section 5 we will see that both dictionaries and corpora are useful sources of RFPs.

2 Recurrent Free Phrases

Following the Princeton WordNet model, synsets can include both single words and *multiword expressions*. See [3] and [10] for a recent discussion on the linguistic status of multiword expressions. More specifically WordNet includes *idioms*, that is relatively frozen combinations of words whose meaning cannot be built compositionally, and *restricted collocation*, that is combinations of words that combine compositionally but show a kind of semantic cohesion which considerably limit the substitution of the component words with synonyms. Multiword expressions must be distinguished from free combinations of words [2, 7]. A *free combination* is a combination of words following only the general rules of syntax: the word meanings combine compositionally and can be substituted by synonyms. Whereas multiword expressions, along with single words, are elementary lexical units [4], free combinations do not belong to the lexicon and thus cannot compose synsets in WordNet.

However, as the boundaries between idioms, restricted collocations, and free combinations are not clear-cut, it is sometimes very difficult to properly distinguish a restricted collocation from a free combination of words. Moreover, applying this distinction in a rigorous manner leads to the consequence that a considerable number of expressions which are recurrently used to express a concept are excluded from wordnets as they are not lexical units.

For example, the English verb “to bike” is always translated in Italian with “andare in bicicletta” but the Italian translation equivalent seems to be a free combination of the word “andare” in one of its regular senses (dictionary definition: to move by walking or using a means of locomotion) with the restricted collocation “in bicicletta” (by bike). Expressions like “andare in bicicletta” contain relevant information about the

co-occurrence of word meanings such as “andare” and “bicicletta, which should be independently coded in any Italian wordnet. We call these expressions Recurrent Free Phrases (RFPs). The main characteristics of RFPs are the following (some of them refer to the native speaker intuition, others are more corpus oriented):

i. RFPs are free combinations of words, which means that they fail the linguistic and semantic tests usually carried out to identify multiword expressions.
ii. RFPs are phrases, i.e. syntactic constituents whose head is either a noun or a verb or an adjective or a preposition. For instance, "eats the" is not an RFP.
iii. High frequency. E.g. "legge elettorale" (electoral law) is found at position 38 on a total of 2,108,000 bigrams extracted from an Italian reference corpus.
iv. High degree of association between the component words. For example, calculating association measures on the reference corpus, we found expressions like "paese europeo" (European country) which score very high.
v. Salience. It refers to the intuition of the native speaker lexicographer that a certain expression picks up a peculiar concept. The concept of salience is not necessarily related to frequency and word association. For example, our lexicographers think that "coscia destra" (right thigh) is less salient than "vertice internazionale" (international summit) whereas it has both a higher frequency and association score.

We are aware that, whereas characteristics from (i) to (iv) are all relatively well defined, the notion of salience is a little vague and needs more investigation. We make the hypothesis that salience is related to the amount of world knowledge that is attached to a certain expression and that cannot be simply derived from the composition of the meanings of the words making up the expression. To see this point consider the difference between “right thigh” and “right hand”. Both are fully compositional, but we feel that “right hand” is more salient than “right thigh”. The “right hand” is not only the hand that is attached to the right arm. This is also the hand we use to write, to swear etc. Note also that high frequency, high degree of association, and salience are all typical but not defining characteristics of RFPs.

RFPs can provide useful information for various kinds of NLP tasks, both in a mono- and multi-lingual environment. For instance, RFPs can be useful for knowledge-based word alignment of parallel corpora, to find correspondences when one language has a lexical unit for a concept whereas the other language uses a free combination of words. Another task which could take advantage of RFPs is word sense disambiguation. RFPs are free combinations of possibly ambiguous words, which are used in one of the regular senses recorded in WordNet. Take for instance the Italian expression “campo di grano” (cornfield). Its component words are highly ambiguous: “campo” has 12 different senses and “grano” 9, but in this expression they are used in just one of their usual senses. Now, suppose that when encoding RFPs, we annotate the component words with the WordNet sense they have in the expression; then, when performing word sense disambiguation, we only need to recognize the occurrence of the expression in a text to automatically disambiguate its component words.

Some RFPs are particularly relevant to the purposes of NLP tasks and we think they should be given priority for inclusion in any wordnet:

- RFPs expressing a concept which is not lexicalized in one language but is lexicalized in another language (i.e. in correspondence with a lexical gap).
- RFPs which are synonymous with a lexical unit in the same language.
- RFPs whose components are highly polysemous. This is meant to facilitate Word Sense Disambiguation algorithms.
- RFPs that are frequent, cohese and salient within a particular corpus considered as a reference corpus.

In the following two sections we will propose two ways of encoding in WordNet the co-occurrence information contained in RFPs, depending on their characteristics.

3 Extending WordNet with Phrasets

The first way to encode collocability information in wordnets is through the introduction of a new data structure called *phraset*, as proposed by [1]. A phraset is a set of RFPs (as opposed to lexical units) which have the same meaning. Phrasets can be added in correspondence with empty or non-empty synsets. We are currently studying the integration of phrasets in the framework of MultiWordNet [8], a multilingual lexical database in which an Italian wordnet has been created in strict alignment with the Princeton WordNet [6].

In a multilingual perspective, phrasets are very useful to manage *lexical gaps*, i.e. cases in which a language expresses a concept with a lexical unit whereas the other language does not. In MultiWordNet lexical gaps are represented by adding an empty synset aligned with a non-empty synset of the other language. Previously, the free combination of words expressing the non lexicalized concept was added to the gloss of the empty synset, where it was not distinguished from definitions and examples. With the introduction of phrasets, the translation equivalents expressing the lexical gaps have a different status, as it is shown in Example 1 below.

Phrasets are also useful in connection with non-empty synsets to give further information about alternative ways to express/translate a concept (Example 2).

Finally, it is important to stress that phrasets contain only free combinations which are recurrently used, and not definitions of concepts, which must be included in the gloss of the synset (Example 3). When the synset in the target language is empty and no expression is found in the phraset, this means that the target language lacks a synonym translation equivalent. The definition allows to understand the concept, but it is unlikely to be used to translate it.

Example 1

Eng-synset	{ toilet_roll }
Ita-synset	{ GAP }
Ita-phraset	{ rotolo_di_carta_igienica}

Example 2

Eng-synset	{dishcloth}
Ita-synset	{canovaccio}
Ita-phraset	{strofinaccio_dei_piatti, strofinaccio_da_cucina}

Example 3

Eng-synset	{straphanger}
Ita-synset	{GAP -- chi viaggia in piedi su mezzi pubblici reggendosi ad un sostegno}
Ita-phraset	{ -- }

Up to now 1,216 phrasets have been created in MultiWordNet, containing a total of 1,233 RFPs.

4 Extending WordNet with Syntagmatic Relations

In some cases word meaning co-occurrence information can be coded through semantic or lexical relations. Some steps in this direction have already been done in the framework of the MEANING project [9], an EU funded project aiming at enriching wordnets with semantic information useful for disambiguation purposes. One of the relations which is being added is the "envolve" semantic relation which encodes deep selectional restriction information, by relating verbal concepts with other concepts that typically occur as arguments (or participants) of the verb.

On the contrary, in our approach we deliberately focus on the kind of co-occurrence information that is not explained by selectional restriction phenomena. Consider for instance the RFP "campagna antifumo" (campaign against smoking). This expression is quite frequent in Italian newspapers, and shows a good degree of log-likelyhood association. Also the noun "campagna" in Italian is ambiguous between the meanings "campaing" and "country-side", but is monosemous in the above RFP, so it is worth including it in WordNet. If we choose to encode the co-occurrence of "campagna" and "antifumo" through a phraset, we need to create a new empty synset which is hyponym of "campagna" in the "campaign" sense, to add a phraset containing "campagna antifumo" in correspondence with such empty synset, and to annotate "campagna" and "antifumo" with their meanings in WordNet.

In principle we could follow a simpler strategy. If WordNet had a "has_constraint" relation relating nominal concepts with adjectival concepts that typically constrain the former, then all we would need to do is add an instance of such relation between the correct synsets for the noun "campagna" and the adjective "antifumo". The use of relations looks like as a concise and smart way of encoding meaning co-occurrence information. This has however a number of limitations:

- It is more suitable for representing bigrams than higher order n-grams. For instance we could somehow represent the fact that "campo" and "grano" co-occur in the RFP "campo di grano", but in this way we would lack the possibility of representing the fact that the two words are connected through the "di" (of) preposi-

tion. Also, using relations to represent RFP with more than two content words is completely impossible.

- It is not possible to represent the fact that two RFPs are synonyms.
- It is not possible to represent the fact that a certain RFP is the translation equivalent of a lexical unit in another language.
- It is not possible to represent restrictions on the order or the morphological features of the words of the RFP.

The solution currently adopted in MultiWordNet to represent syntagmatic relations tries to get the best of both phrasets and explicit relations. RFPs are indeed explicitly represented in phrasets, but a new lexical relation (composes/composed-of) between phrasets and synsets is used to annotate the senses of the words in the RFPs. Figure 1 shows how the RFP "campo di grano" is represented in MultiWordNet.

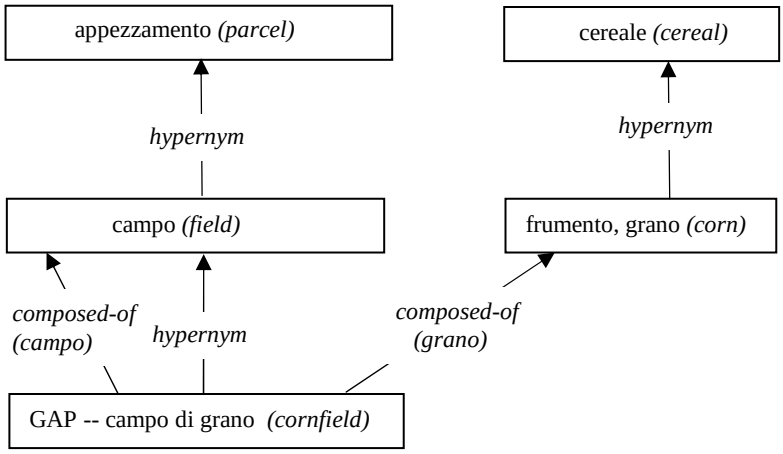

Fig. 1. Representing syntagmatic relations in MultiWordNet

5 Recurrent Free Phrases in Dictionaries and Corpora

In [1] a study is presented to verify the possibility of acquiring RFPs from both dictionaries and corpora. First, we studied all the Translation Equivalents (TEs) of the Collins English/Italian dictionary corresponding to English to Italian gaps (7.8% of the total). By manually checking 300 Italian lexical gaps, a lexicographer found out that in 67% of the cases the TEs include a RFP. In the remaining cases the TEs are definitions. We used the result of this experiment to infer that more than half of the synsets which are gaps in any Italian wordnet potentially have an associated phraset.

In Section 3 we saw that phrasets can be associated also to regular (non empty) synsets. To assess the extension of this phenomenon, we first looked for cases in which the Collins dictionary presents an Italian TE composed of a single word, to-

gether with at least a TE composed of a complex expression. This happens in 2,004 cases (12% of the total). A lexicographer manually checked 300 of these complex expressions and determined that in 52% of the cases at least one complex expression is a RFP. In the remaining cases the complex expressions provided as TEs are either lexical units or definitions.

A second experiment has been carried out on an Italian corpus to compare multiword expressions and RPFs from a frequency point of view, and thus to assess the possibility of extracting RFPs from corpora with techniques similar to those used for collocation extraction. More specifically, we considered contiguous bigrams and trigrams with frequency higher then 3, and excluding stopwords. The results of the experiment show that, as expected, the number of bigrams that are lexical units decreases regularly along with the rank of the frequency, whereas non lexical units increase complementary. However, within non-lexical units the number of RFPs seems not to be correlated with the rank of the bigrams, fluctuating irregularly between a minimum of 3% and a maximum of 15%.

6 Conclusions and future work

We presented a proposal to extend the WordNet model with syntagmatic information about the co-occurrence of word meanings. This information is contained in RFP, that is free combinations of words characterized by high frequency, cohesion, and salience. Such expressions can be listed in phrasets (sets of synonymous RFPs), which are useful to handle lexical gaps in multilingual databases, and to provide alternative ways to express a concept in correspondence with regular synsets. The information contained in phrasets can be used to enhance word sense disambiguation algorithms, provided that each expression of the phraset is annotated with the specific meaning that its component words assume in the expression. The annotation of RFPs is implemented through a new lexical relation (composes/composed-of) relating phrasets and synsets. Evidence has been provided that RFPs can be extracted from both bilingual dictionaries and corpora with techniques similar to those used for collocation extraction. A lot of work need still to be done to better understand the lexicographic status of RFPs, and the practical implications of their inclusion in wordnets.

References

1. Bentivogli, L. and Pianta, E.: Beyond Lexical Units: Enriching WordNets with Phrasets. In: *Proceedings of EACL-03*, Budapest, Hungary (2003)
2. Benson, M., Benson, E., Ilson, R.: *The BBI combinatory dictionary of English: a guide to word combinations*. John Benjamins Publishing Company, Philadelphia (1986)
3. Calzolari, N., Fillmore, C., Grishman, R., Ide, N., Lenci, A., MacLeod, C., Zampolli, A.: Towards Best Practice for Multiword Expressions in Computational Lexicons. In: *Proceedings of LREC 2002*, Las Palmas, Canary Islands (2002)
4. Cruse, D.A.: *Lexical semantics*. Cambridge University Press, Cambridge (1986)
5. de Saussure, F.: *Cours de linguistique générale*. Payot, Paris (1916)

6. Fellbaum, C. (editor): *WordNet: An electronic lexical database*. The MIT Press, Cambridge, Mass. (1998)
7. Heid, U.: On ways words work together: research topics in lexical combinatorics. In: *Proceedings of Euralex-94*, Amsterdam, Holland (1994)
8. Pianta, E., Bentivogli, L., Girardi, C.: MultiWordNet: Developing an Aligned Multilingual Database. In *Proceedings of the First International Conference on Global WordNet*, Mysore, India (2002)
9. Rigau, G., Magnini, B., Agirre, E., Vossen, P., Carroll, J.: A Roadmap to Knowledge Technologies. In: *Proceedings of COLING Workshop "A Roadmap for Computational Linguistics"*. Taipei, Taiwan (2002)
10. Sag, I., Baldwin, T., Bond, F., Copestake, A., Flickinger, D.: Multiword Expressions: a Pain in the Neck for NLP. In: *Proceedings of CICLING 2002*, Mexico City, Mexico (2002)