Beyond Lexical Units: Enriching Wordnets with Phrasets

Luisa Bentivogli, Emanuele Pianta
ITC-irst, Trento, Italy
{bentivo,pianta}@itc.it

Abstract

In this paper we present a proposal to extend WordNet-like lexical databases by adding *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. Two experiments are presented to check the possibility of acquiring recurrent free phrases from dictionaries and corpora.

1 Introduction

WordNet (Fellbaum, 1998) is a popular lexical database for English in which content words are organized into sets of synonyms (synsets), each representing one underlying lexical concept. Words and concepts are further connected through various lexical and semantic relations. WordNet has been widely adopted in the NLP community for a variety of practical tasks such as word sense disambiguation, question answering, information retrieval, summarization, etc. The English WordNet database is being used as a basis for the development of different multilingual databases such as EuroWordNet, MultiWordNet, and the recent BalkaNet project. To make it more useful in NLP applications, WordNet is constantly updated and extended with different kinds of information such as domain information, syntactic information, topic signatures, syntactic parsing and PoS tagging of the glosses, etc.

In this paper we propose to extend the WordNet model by adding a new data structure called *phraset*. A phraset is a set of free combinations of words (as opposed to lexical units) which are recurrently used to express a concept.

Phrasets can provide useful information for different kind of NLP tasks, both in a monolingual and multilingual environment. For instance, phrasets 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 phrasets is word sense disambiguation. The expressions contained in phrasets 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 adding an expression to a phraset, 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.

We are currently studying the integration of phrasets in the framework of MultiWordNet (Pianta et al., 2002), a multilingual lexical database in which an Italian wordnet has been created in strict alignment with the Princeton WordNet.

To enrich the Italian lexical database with phrasets, we explored techniques exploiting both machine-readable bilingual dictionaries and corpora. The results of two preliminary experiments will be presented in Section 4.

2 Lexical units in WordNet

Following the Princeton WordNet model adopted in MultiWordNet, synsets can include both single

words and multiwords which are idioms or restricted collocations. See Sag et al. (2002) for a recent discussion on the linguistic status of multiword expressions.

An *idiom* is a relatively frozen expression whose meaning cannot be built compositionally from the meanings of its component words. Also, the component words cannot be substituted with synonyms. The following examples are taken from MultiWordNet: E- stands for the English wordnet and I- for the Italian one.

E-synset {rollercoaster, big dipper, ...}
I-synset {montagne_russe}

A *restricted collocation* is a sequence of words which habitually co-occur and whose meaning can be derived compositionally. Restricted collocations have a kind of semantic cohesion mainly due to use and, therefore, they considerably limit the substitution of their component words. Usually, restricted collocations do not have a literal translation in other languages.

E-synset {criminal_record, record}
I-synset {precedenti_penali}

Idioms and restricted collocations must be distinguished from free combinations of words. A *free combination* is a combination of words following only the general rules of syntax: the elements are not bound specifically to each other and so they occur with other lexical items freely (Benson et al., 1986).

While idioms and restricted collocations are lexical units, free combinations do not belong to the lexicon and thus cannot compose synsets in MultiWordNet.

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 MultiWordNet 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). The same holds for the Italian phrases "punta di freccia" and "punta della freccia" which can hardly be considered restricted collocations but are recurrently used to translate the English word "arrowhead".

3 Introducing Phrasets

To be able to include in our lexical database expressions such as "andare in bicicletta" or "punta di freccia", we propose to extend the (Multi) WordNet model by adding *phrasets*. A phraset is a set of free combinations of words which are recurrently used to express a concept. Let's call the members of a phraset *recurrent free phrases*.

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 the current version of MultiWordNet we represent lexical gaps by adding an empty synset aligned with a non-empty synset of the other language. The free combination of words expressing the non lexicalized concept is added to the gloss of the empty synset, where it is not distinguished from definitions and examples.

With the introduction of phrasets, the translation equivalents expressing the lexical gaps would have a different status, as it is shown in the examples below.

E-synset {cornfield}
I-synset {GAP}
I-phraset {campo_di_grano}

E-synset {toilet_roll}
I-synset {GAP}
I-phraset {rotolo_di_carta_igienica}

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

E-synset {dishcloth}
I-synset {canovaccio}
I-phraset {strofinaccio_dei_piatti, strofinaccio_da_cucina}

3.1 Recurrent Free Phrases versus Definitions

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.

E-synset	{tree}
I-synset	{albero -- ogni pianta perenne con fusto legnoso ramificato}
I-phraset	{ -- }
E-synset	{paperboy}
I-synset	{GAP -- ragazzo che recapita i giornali}
I-phraset	{ragazzo_dei_giornali}
E-synset	{straphanger}
I-synset	{GAP -- chi viaggia in piedi su mezzi pubblici reggendosi ad un sostegno}
I-phraset	{ -- }

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.

4 Recurrent Free Phrases in Dictionaries and Corpora

We did some experiments to verify the possibility of acquiring recurrent free phrases both from dictionaries and from corpora.

4.1 Bilingual Dictionaries

For each word sense, bilingual dictionaries provide one or more translation equivalents (TEs), which can be a single word or a complex expression. Some of the complex expressions are lexical units (idioms or restricted collocations), other are free combinations of words. When none of the TEs of the word sense in the source language is a lexical unit, a lexical gap occurs in the target language. Bentivogli and Pianta (2000) analyzed the English to Italian section of the Collins bilingual dictionary and found that 92.2% of the English word senses correspond to at least an Italian lexical unit, whereas 7.8% correspond to an Italian lexical gap (all the TEs are free combinations of words).

Starting from the results of this study, we carried out an experiment to verify in how many cases the free combinations of words provided by the Collins as TEs to express an Italian lexical gap include at least a recurrent free phrase. By manually checking 300 Italian lexical gaps, a lexicographer found out that in 67% of the cases the TEs include a recurrent free phrase. In the remaining cases the TEs are definitions. We can use the result of this experiment to infer that more than half of the synsets which are gaps in the Italian section of MultiWordNet 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, together 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 recurrent free phrase. In the remaining cases the complex expressions provided as TEs are either lexical units or definitions.

During the manual control, in order to distinguish between recurrent free phrases and definitions, the lexicographer used the web to check if the expression provided by the dictionary is really used in general language.

4.2 Corpora

A second experiment has been carried out on an Italian corpus to compare complex lexical units and recurrent free phrases from a frequency point of view, and thus to assess the possibility of extracting recurrent free phrases from corpora with techniques similar to those used for collocation extraction. More specifically, we considered contiguous bigrams and trigrams. A standard package for the analysis of n-grams has been used (Banerjee and Pedersen, 2003).

First we extracted from a 2 year newspaper corpus of 32 million words all the *bigrams* with frequency higher than 3. A list of stop words has been used to exclude from the final list all bigrams containing at least one function word. This yielded a list of 118,464 bigrams, ordered ac-

cording to the number of occurrences (rank). The highest rank turned out to be 5,914 (the bigram "New York" occurs 5,914 times in the corpus), the lowest rank (4) included 31,453 bigrams (26,5% of the total). The 497 distinct ranks occurring in the frequency list have been divided into 9 groups with the following ranges (in parenthesis the number of bigrams included in the group): A: 5,914-509 (100); B: 505-257 (257); C: 256-129 (731); D: 128-65 (1,956); E: 64-33 (4,525); F: 32-17 (10,477); G: 16-9 (22,167); H: 8-5 (46,798); I: 4 (31,453). A lexicographer manually checked the first 100 bigrams of each group, classifying them in three groups: lexical units, recurrent free phrases, other. The following table summarizes the results of the manual check:

	A	B	C	D	E	F	G	H	I
Lex. Unit	82	79	74	65	58	55	42	35	28
R. F. P.	14	4	9	14	17	4	15	3	15
Other	4	17	17	21	25	41	43	58	57

The table shows 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 recurrent free phrases seems not to be correlated with the rank of the bigrams, fluctuating irregularly between a mininum of 3 and a maximum of 15. A similar experiment carried out on trigrams gave very similar result.

5 Open Issues

Introducing phrasets will not solve all the problems related to the inclusion of multiword expressions in MultiWordNet. In some cases it will still be difficult to decide which expressions are to be included in synsets, which ones in phrasets and which ones are just definitions. For example, the English word "backyard" can be translated in Italian with "giardino posteriore", "giardino sul retro", "giardino sul retro della casa". The first two expressions are on the borderline between synset and phraset, while the third is on the borderline between phraset and definition.

However in most cases phrasets provide a flexible tool to aid lexicographers in the process of choosing the lexical status of multiword expressions. Moreover, phrasets store information which otherwise would be lost and which is useful for NLP applications.

6 Conclusion

We presented a proposal to extend the (Multi) WordNet model with phrasets, which requires the inclusion in the lexical database of expression that are not lexical units. Such expressions are useful to handle lexical gaps in multilingual databases, but can also be added to regular synsets to provide alternative ways to express/translate a concept. 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. Evidence has been provided that recurrent free expressions can be extracted from both bilingual dictionaries and corpora with techniques similar to those used for collocation extraction.

References

Morton Benson, Evelyn Benson, and Robert Ilson, 1986. *The BBI combinatory dictionary of English: a guide to word combinations.* John Benjamins Publishing Company, Philadelphia.

Luisa Bentivogli, and Emanuele Pianta, 2000. "Looking for lexical gaps". In *Proceedings of the ninth EURALEX International Congress*, Stuttgart, Germany.

Christiane Fellbaum, editor, 1998. *WordNet: An electronic lexical database.* The MIT Press, Cambridge, Mass.

Emanuele Pianta, Luisa Bentivogli, and Christian Girardi, 2002. "MultiWordNet: developing an aligned multilingual database". In *Proceedings of the First International Conference on Global WordNet*, Mysore, India.

Ivan Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger, 2002. "Multiword Expressions: A Pain in the Neck for NLP". In *Proceedings of CICLING 2002*, Mexico City, Mexico.

Satanjeev Banerjee, and Ted Pedersen, 2003. "The Design, Implementation and Use of the Ngram Statistics Package". In *Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics*, Mexico City, Mexico.