Multilinguality and Reversibility in Computational Semantic Lexicons

Evelyne Viegas and Stephen Beale
viegas@crl.nmsu.edu and sb@crl.nmsu.edu
Computing Research Laboratory
Las Cruces, NM 88003 USA

Abstract

In this paper, we address the issue of generating multilingual computational semantic lexicons from analysis lexicons, showing the necessity of relying on a conceptual lexicon. We first discuss the type of information which should be found in NLP lexicons, whatever their use (analysis, generation, speech, robotics). We claim that we should take advantage of the existing large-scale analysis lexicons and use them as the starting point in the process of building large-scale generation lexicons, by first reversing them and then enhancing them. This implies having access to a conceptual lexicon, which will serve as a pivot point between the analysis and the generation lexicons. We implemented the work reported here for Spanish and English MT projects, within the knowledge-based paradigm. From a theoretical point of view, regenerating the source text with the reversed analysis lexicon enabled us to enhance several issues as diverse as: evaluating analysis lexicons, testing the semantic analyser, evaluating which information should be added to the generation lexicon; and testing the grain-size of the pivot point between analysis and generation.

Introduction

There is no consensus on the type of lexicons which should be used for generators. It seems to depend on the type of generator. It also seems to depend on the kind of application involved: monolingual generation, multilingual generation, machine translation; generation of sentences vs. texts vs. speech; or also generation from raw data vs. from conceptual representations built with generation in mind.

Once one has an application in mind, then there are three main approaches one can adopt to build the lexicon:

lexicographic: very attractive for NLP applications at first sight, as they provide a useful description of the vocabulary; entries are distinguished on the basis of multiple senses and subcategorisations. But in practice, this approach complicates the process of lexical disambiguation for parsing and lexical choice in generation by an unjustified proliferation of entries.

statistical: very attractive for NLP applications as it seems to replace knowledge-based approaches and therefore supplant the needs for human acquisition of large-scale semantic lexicons. which is a very time consuming task. However, the limits of statistical approaches have been pointed out by [Smadja, 1993]. Moreover, some phenomenon, such as *event ellipsis* (EE) cannot be handled by a pure statistical approach nor by a lexicographic approach, as its recovery necessitates a semantic treatment, ([Viegas and Nirenburg, 1995a]), which is, in fact, handled by a computational linguistic approach making use of semantics.

computational linguistic: the main advantage of this approach is that it is usually theoretically grounded, and is domain- and application-independent. Moreover, it can handle phenomena which are out of the reach of other approaches and yet are necessary to enhance lexical choice in generation. For instance, the EE triggered by *enjoy* as in i) *I enjoyed the salmon very much*, must be modeled with a semantic representation so that its recovery can be taken care of as in ii) *I enjoyed eating....* This is part of lexical choice as one can choose to realise the **synthetic** version or the **analytic** version of the EE, as exemplified in i) and ii) respectively (cf. [Viegas and Nirenburg, 1995a]).

However necessary, adopting a linguistic approach is a difficult task, as one of the main drawbacks of this approach is that it is time consuming as far as the building of the lexicon is concerned. We address in next section how to bypass this drawback.

We would like to thank in the mikrokosmos team, Tom Herndon, Jeff Longwel, Oscar Cossio and Javier Ochoa.

A Multi-purpose Knowledge Base

Since building computational semantic lexicons is a very time-consuming task, we should aim at lexicons which conform to the three following conditions:

a - **multi-lingual:** French, English, Japanese, Russain, Spanish, etc..., (format of the lexicon)

b - **multi-media:** such as having linguistic information for natural language processing, phonological information, essentially for speech recognition and production, (structure of the lexicons)

c - **multi-use:** so that they can be used for analysis, generation (mono/multi-lingual), MT, or speech processing. (reversibility of the lexicons)

The way we organised and structured our lexicons directly follows these conditions. Large-scale computational generation lexicons carrying semantic information are indeed not that common, the obvious reason being that acquiring semantic information is a difficult and time consuming task. However, it is by no means an unattainable task, if we structure and organise our analysis lexicons in such a way so that the information they contain can be used at best for building generation lexicons.

Acquiring a large-scale lexicon is very expensive, which is why building lexicons that are reusable for other domains or applications is recommended. It is well known in computational lexical semantics that a sense enumeration approach only based on subcategorisation differences is computationally expensive and unrealistic from a theoretical viewpoint.

Our lexicons are composed of superentries, where each entry consists of a list of words, stored there independently of their part of speech (the verb and noun form of *walk* are under the same superentry), as described in length in [Onyshkevych and Nirenburg, 1994].

Reversing an Analysis Lexicon

Before addressing the issue of reversing the analysis lexicon, we want first to show how we could acquire a large-scale analysis lexicon.

Acquisition of the Analysis Lexicon

We acquired a Spanish semantic lexicon of about 40,000 word meanings, for an MT Project, described in [Beale et al., 1995]. We automated as much as possible the task of acquisition by providing the lexicographers with access to on-line dictionaries, on-line corpora, and also software allowing lexicographers to access all this on line information in an easy way (see [Viegas and Nirenburg, 1995b] for the task of acquisition). Our interfaces have been designed with respect to users needs, and continue to evolve on a needed basis.

We give below the example of the partial entry *compañía* in Spanish, with two different mappings, represented by the following concepts in our world model (or ontology): CORPORATION, INTERACT-SOCIALLY. One important point here to notice is our transcategorial approach. There is no one-to-one mapping between semantic categories or concepts and lexical items, and some EVENTs, (such as INTERACT-SOCIALLY here) can be lexicalised as nouns (Figure 1)[1] or verbs such as in *acompañar*:

$$\begin{bmatrix} \text{compañía-N1} & \\ \text{syn:} & \begin{bmatrix} \text{root: } \boxed{0} \begin{bmatrix} \text{cat:} & \textbf{N} \\ \text{sem:} & \boxed{00} \end{bmatrix} \end{bmatrix} \\ \text{sem:} & \begin{bmatrix} \boxed{00}\ \textbf{corporation} \end{bmatrix} \\ \text{compañía-N2} & \\ \text{syn:} & \begin{bmatrix} \text{root: } \boxed{0} \begin{bmatrix} \text{cat:} & \textbf{N} \\ \text{sem:} & \boxed{00} \end{bmatrix} \end{bmatrix} \\ \text{sem:} & \begin{bmatrix} \boxed{00}\ \textbf{interact-socially} \end{bmatrix} \\ \ldots & \end{bmatrix}$$

Figure 1: Sense Entries for the Spanish lexical item *compañía*.

Let us now consider some of the entries for the Spanish verb *adquirir* with the following corresponding semantics: ACQUIRE, LEARN, displayed in (Figure 2).

The sub-entries for *adquirir* have different selectional restrictions for the theme, OBJECT and INFORMATION for ACQUIRE and LEARN respectively.

We have acquired about 1/5 of our lexicon semi-automatically and have developed a morpho-semantic acquisition program, which has allowed us to acquire the remaining 4/5 entirely automatically to create at the end a large-scale lexicon of about 40000 word senses.[2] The main advantage of our approach is that it enabled us to economically multiply the size of the lexicon. The main drawback is that the entries produced automatically need some semi-manual checking. We bypass this drawback by

[1] We use the typed feature structures (tfs) as described in [Pollard and Sag, 1987)[].

[2] See [Viegas et al., 1996] which describes the advantages and drawbacks of using lexical rules to build lexicons.

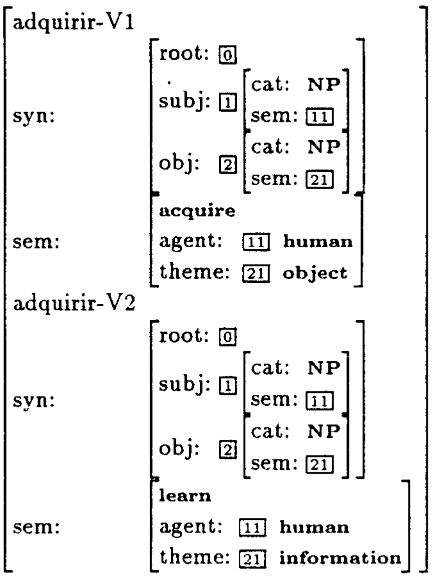

Figure 2: Sense Entries for the Spanish lexical item *adquirir*.

using the reversed lexicon to regenerate the entry as explained below.

The Reversed Lexicon

The algorithm to "reverse" the analysis lexicon (AL) to produce the generation lexicon (GL) mainly involves rearranging, modifying, deleting, and adding certain items. We focus below on the zones which have been reversed, namely: SYN (subcategorisation information) and SEM (providing the semantic information with associated selectional restrictions), as shown in (Figure 3).

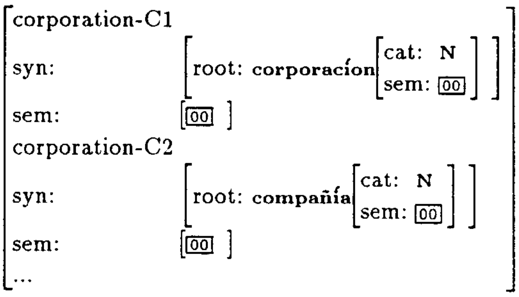

Figure 3: Partial Entry for the concept CORPORATION.

Our transcategorial approach to sense discrimination is a good basis for paraphrasing, thus the concept ACQUIRE from the ontology, can be lexicalised in our Spanish lexicon, at least in: *adquirir, obtener, conseguir* (verbs), *adquisición, obtención, enriquecimiento* (nouns), *codicioso* (adjective). We only show partial entries for superentry of the concept ACQUIRE, as shown in (Figure 4).

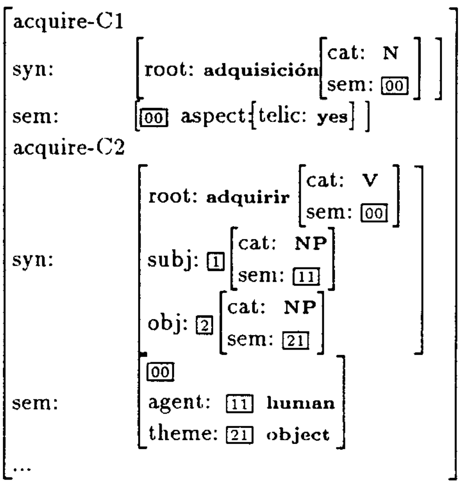

Figure 4: Partial Entry for the concept ACQUIRE.

We are now in the phase of enhancing the reversed lexicon for producing the Spanish and English generation lexicons; namely, we are encoding information which is specific to the process of generation and which can be avoided in an analysis lexicon, such as word order (in Adj-noun constructions), and collocational information acquired semi-automatically from corpora (*hacer una adquisición*).

Moreover, with this technique, we can produce multilingual generation lexicons by lexicalising the concepts of the reversed lexicons in different languages, this ensures that we will have a lexical item or phrase for lexicalisation available.

Another advantage of reversing an analysis lexicon is using it to regenerate the same text that was parsed to gain some insight into the issue of the pivot point between parsing and generation, and as a result of this, what is the best input for generation.

Advantages of a Reversed Lexicon

A reversed lexicon has advantages beyond its practical use in generation. We have identified, and in some cases begun work on the following areas:

Evaluation of semantic analysis. With a reversed lexicon that is based on the original analysis lexicon, it is possible to take the output semantic representations from the analyser and submit them to a text generator. The output surface structures can then be compared to the input text. Apart from this, evaluation of semantic analyses can be difficult because it involves reading and understanding complex meaning representations.

Evaluating Text Meaning Representation language. For example, the granularity of semantic representation can be studied. Is the representation precise enough to correctly translate all meaning components, or is a specific source term mapped into a generalised one from which the original meaning cannot be recovered? This will be especially helpful in a multilingual environment where meaning components might be bundled differently.

Testing lexicon entries. We have developed a suite of tools to help in testing the analysis lexicon, to ensure the high-quality of our large-scale lexicon. These tools range in complexity from checking placement of parentheses to automatically creating sentences to test individual lexicon entries. For the latter, having a reversed lexicon available is extremely helpful. For example, a simple lexicon entry for the English word *read* might look like:

```
READ-V1
 syn-struc:
   root: read
   Subj: VAR1
   OBJ:  VAR2
 sem-struc:
   READ
    AGENT: VAR1 = HUMAN
    THEME: VAR2 = BOOK
```

We can then use the reversed lexicon to generate sentences that conform to the given syn-struc but substitute appropriate words or phrases in place of the variables. Sentences such as the following would signal problems:

The book read John.

John read into the book.

John read the cheese.

This is especially helpful for automatically generated lexicon entries such as nominalisations, which are created from verbal entries using lexical rules. Many thousands of such entries have been created; tools such as this provide a simple way to check their accuracy.

Conclusion

In this paper we argued that, although lexicographic and statistical approaches have their place in natural language processing, computational semantic lexicons are necessary for a wide range of phenomena and are applicable to a number of purposes. Unfortunately, large-scale acquisition of computational lexicons is difficult. Compounding this problem is the fact that analysis systems require different information than generation lexicons.

We have developed a method that enables us to take advantage of the large investment made in the Mikrokosmos analysis lexicon. We outlined a relatively simple process for reversing analysis lexicons for eventual use in generation. This process transfers relevant information and re-indexes it according to the needs of generation. The process is not perfect; some information required in generation, such as collocational constraints, is not typically recorded in analysis lexicons. Nevertheless, the methods described here provide a baseline to which additional information can be added.

Creating these reversed lexicons has produced a number of additional advantages beyond those originally envisioned, mostly in the area of testing and evaluating the semantic analysis system. By back-translating the results of semantic analysis, evaluation is simplified. Testing the content of individual lexicon entries can also be made easier by generating sample sentences that conform to them. And finally, theoretical issues concerning the content of analysis lexicons, generation lexicons and the text meaning representation language can be more fully investigated.

References

Beale, S., Nirenburg, S. and Mahesh, K. (1995) Semantic Analysis in the Mikrokosmos Machine Translation Project. In *Proceedings of the 2nd Symposium on NLP*, Bangkok, Thailand.

Onyshkevych, B. et S. Nirenburg (1994) *The Lexicon in the Scheme of KBMT Things.* Technical Report MCCS-94-277, CRL, NMSU.

Smadja, F. (1993) Retrieving Collocations from Texts: Xtract. In *Computational Linguistics, 19(1).*

Viegas, E. and Nirenburg, S. (1995a) The Semantic Recovery of Event Ellipsis: its Computational Treatment. In Proceedings of the *Workshop "Context in Natural Language Processing", of (IJCAI95)*, Montréal, Québec.

Viegas, E. and Nirenburg, S. (1995b) Acquisition semi-automatique du lexique. Proceedings of *LTT*, Lyon 95, France.

Viegas, E., Onyshkevych, B., Raskin, V., Nirenburg, S. (1996) From *Submit* to *Submitted* via *Submission*: on Lexical Rules in Large-scale Lexicon Acquisition. In *ACL'96*, Santa-Cruz, California.