[*MT Summit III*, Washington, DC, USA, July 1991]

Automation of Bilingual Lexicon Compilation

I. Arturo Trujillo*
Computer Laboratory
University of Cambridge
Cambridge CB2 3QG, England

David A. Plowman†
Computer Laboratory
University of Cambridge
Cambridge CB2 3QG, England

Abstract

This paper shows that, there are a number of common concepts which are used to define a class of nouns in standard, monolingual English and Spanish dictionaries. An experiment is described to show how a small sot of such concepts was derived semi-automatically by automatically analysing the definitions in each language and then matching equivalent definitions manually. Also, some of the benefits of constructing such sets are described, together with the problems encountered while carrying out the experiment.

1 Introduction

The bilingual dictionary is a crucial component in all Machine Translation (MT) systems which do not adopt the interlingua strategy. Examples include TAUM-METEO, SYSTRAN, METAL, GETA-AR1ANE, the systems at PAHO, EUROTRA and many others (see [Hutchins, 1986], [Nirenberg, 1987], [Slocum, 1988]). Its construction, however, is time consuming and inefficient, especially when translation between more than two languages is involved. One way to automate bilingual lexicon compilation is to describe each monolingual entry in terms of a set of common meaning primitives and then use these primitives as a search key to match entries from different languages.

Pure meaning primitives are an elusive idea, and there is no agreed methodology on how to search for them. For the purposes of bilingual dictionary construction, it might suffice to have a set of common concepts for which there are equivalent words in most languages, and in terms of which other word definitions might be made. Since these concepts would be common to all languages they could be thought of as an interlingua vocabulary in which monolingual lexicons could be written. For this reason they will be referred to as the Lexicon's Interlingua (henceforth LI).

The Longman Dictionary of Contemporary English (LDOCE) [Procter, 1978] shows that approximately 2200 basic words might suffice to define all other words in English. Now, comparing LDOCE definitions with those found in the Spanish monolingual dictionary 'VOX Diccionario Manual Ilustrado de la Lengua Española' (VOX) [Gili Gaya, 1988], we find that they share a set, of common concepts. It would therefore be interesting to base our search for these concepts on existing monolingual dictionaries.

2 Evidence

Before comparing word sense definitions from each dictionary two points should be borne in mind. Firstly, LDOCE has been written with the foreign language student in mind and hence definitions have been made as simple as possible. In contrast, VOX contains word explanations aimed at the native user of the language. Secondly, VOX does not claim to use a restricted defining vocabulary, thus making its applicability to natural language processing less straightforward.

The following two definitions show typical entries for a common noun:

VOX
aeroplano: *vehículo* aéreo más pesado que el aire.
lit. aeroplane: air *vehicle* more heavy than the air

LDOCE
airplane: a flying *vehicle* that is heavier than air, that has wings, and has at least one engine

As we can see, both definitions share the same generic term, namely *vehicle.* This generic term, or genus for short, stands for a very general concept, which has a very direct translation in Spanish. It also happens that many other words in English and Spanish are defined in terms of this concept. For instance in LDOCE we have 'bicycle', 'bus', 'car' and 'tank', all defined as vehicles of some sort. Similarly in VOX we have *bicicleta* †, *ómnibus*†, *automóvil* † and *tanque* (the † means that in the definition of that word, *vehículo* is not used as the genus term but that whatever the genus term is, it is defined somewhere else in the dictionary using the genus *vehículo).*

Given the above exampie, we might consider including the concept corresponding to *vehicle* in our experimental English-Spanish LI. Similarities like the above are

* This work was funded by an SERC, UK, Advanced Course grant.

† Now at Harlequin Ltd., Harrington Hall, Barrington, Cambridge CB2 5RG, England.

very common throughout the noun definitions in both dictionaries. An attempt was therefore made at semi-automatically pairing the genus term of equivalent definitions from each dictionary.

3 Experiment

An experiment was carried out to probe the feasibility of finding, or, more modestly, verifying a Spanish-English LI with the aid of a computer and machine readable (MR) versions of the definitions,

The set of definitions studied comprised what is called a substance hierarchy. A substance hierarchy can be thought of as a tree, with the root node representing the concept 'substance'. Below the root there are other nodes winch are seen as representing substances of different types. These might include 'powder', 'liquid', 'grease', 'food', etc. Below each of these nodes there are further nodes representing further subtypes and so on. For example, below 'liquid' we could have 'drink' and below 'drink', 'tea'.

The actual selection of the word set was done as follows: a pattern matching program, written by [Alshawi, 1989], was used to retrieve the words of the substance hierarchy from the MR version of LDOCE. The algorithm used is described in [Copestake, 1990]. Basically, it takes the word 'substance' as a key and collects all those sense definitions which have this key as a genus term. Then the word corresponding to each of these senses is taken as key and the algorithm is applied recursively, until the keys add no further words. To limit the number of definitions analysed, only those words in the LDOCE core vocabulary were included. The final set comprised 85 words. These head words were translated manually into Spanish using a bilingual dictionary, Their sense definitions were then collected to form the two sets of definitions studied.

From these definitions, a grammar was written for each language based on the Generalised Phrase Structure Grammar formalism of [Gazdar *et al.*, 1985] implemented in the GDE system described in (Carroll *et al.,* 1988]. Each grammar contained approximately 100 PS rules and could parse 70% of all definitions. For a more detailed description of the English and Spanish grammars and the experiment in general see [Plowman, 1990] and [Trujillo, 1990].

Before parsing the definitions, those word senses which did not constitute a 'substance' hierarchy were deleted from the study. For example the 'soul' sense of 'spirit' would be deleted to leave the 'drink' sense. The result of this elimination process were two files containing approximately 100 definitions in each language. These two files were then parsed in batch mode to obtain parse trees in the form of labelled, nested Lisp lists. For example, the definitions for *chocolate* were:

VOX

chocolate2: *bebida* hecha de esta *pasta.
(*chocolatel)

lit. chocolate2: *drink* made from this *paste.
(*chocolatel)

Labelled List

```
(N1   (N1   (N0 bebida))
    (AP  (Al   (A0 hecha)
            (PP  (P1   (P de)
                    (NP   (Det esta)
                        (N1   (N0 pasta)..)
```

LDOCE

chocolate4: a *drink* made from hot. milk mixed with this *powder. (*chocolate3)

Labelled List

```
(NP (Det a)
    (N1  (N0 drink)
        (VP  (V0 made)
            (PP  (P0  from)
                (NP  (AP hot)
                    (N1  (N0 milk)
                        (VP  (V0 mixed)
                            (PP  (P0 with)
                                (NP  (Det this)
                                   (N0 powder)..)
```

The asterisks have been added to note the anaphoric reference made in the definitions to other senses of the word: sense 1 in VOX and sense 3 in LDOCE.

Using the pattern matching program XS, written by Plowman [Plowman, 1990], head-attribute (H-A) structures were constructed from these labelled lists. H-A structures consisted of a genus term and its differentia. The differentia comprised a predicate such as CONSTITUTION or PURPOSE and its value. The function of the differentia was to further restrict the meaning of the genus.

Each predicate was added to the representation when a certain pattern appeared in the parse tree. For example, the H-A structures corresponding to the above two definitions were:

```
(BEBIDA (CONSTITUTION
           (PASTA)))

(DRINK  (CONSTITUTION
           (MILK  (CONSTITUTION
                       (POWDER)))))
```

The CONSTITUTION predicate was added by XS when the patterns *hecha de, made of* and *mixed with* were found. Its argument was built from the head of the noun phrase which followed the pattern.

Once structures were found for each pair of equivalent sense definitions, a program was run which not only paired the genus but also the value assigned to corresponding predicates. The two definitions of *chocolate* above would result in:

```
((DRINK BEBIDA)     (MILK PASTA))
```

4 Results

The following list of genus and predicate value pairs was obtained:

```
?(DRINK INFUSION) ?(DRINK COSA) *(DRINK SEMILLA)
(DRINK BEBIDA)  (DRINK BEBIDA)  ?(EMBER SUBSTANCIA)
(FAT MANTECA)   (FAT GRASA)   (FLESH CARNE)
(FOOD ALIMENTO)  *(FOOD MASA)   *(FOOD HOJA)
(GRAIN GRANOS)  ?(LIQUID SUBSTANCIA)  *(LSD NOMBRE)
(MATTER MATERIA)   (MEAT CARNES)
?(MEDICINE SUBSTANCIAS)  *(MEDICINE SUAVIZAR)
```

```
*(METAL LAMINA) *(MILK PASTA)  *(MIXTURE MASA)
*(MIXTURE FLUIDO) *(PART CARNE)
(PREPARATION CONFECCION) ?(SOMETHING CUERPO)
?(SUBSTANCE MEDICAMENTO) ?(SUBSTANCE GRASILLA)
(SUBSTANCE SUBSTANCIA)  (SUBSTANCE SUBSTANCIA)
?(SUBSTANCE COSA)  ?(SUBSTANCE NATA)
?(SUBSTANCE PORCELANA)  ?(SUBSTANCE MANTECA)
?(SUBSTANCE COMPUESTO) (WATER AGUA) (WATER AGUA))
```

In the above list, pairs of words which have no relation to each other have been marked with an asterisk. Those which share some sort of relation have been marked with a question mark. The following English words have been paired with Spanish words of similar meaning: drink, fat, flesh, food, grain, matter, meat, preparation, substance and water (it should be pointed out that for this pairing we are ignoring issues of polysemy between paired words). These are all common words which have direct translations at least into most Indo-European languages. Their concepts could comprise a very simple LI for substance nouns. For instance, we could define *beer* in both languages in terms of the concepts DRINK, WATER and GRAIN. Obviously this is an extremely broad description, but for all substances which have a name in a language, it could be used by a program to narrow the choices to a few words and then allow a human to select between them.

5 Problems

The development of both grammars posed two problems. Firstly, we faced the problem of structural ambiguity common to most natural language processing enterprises. Secondly, it proved very difficult to write a grammar which was both general and complete. Initially we expected the sublanguage of definitions to be restricted, but found that it was hard to predict where parenthetical material occurred within definitions, or how definitions would deviate from the general 'dictionary' style format.

Constructing the representations required inspecting their content in order to write pattern rules which retrieved semantically loaded genus such as *vehicle* above and not what are sometimes called Linkers (see [Vossen *et al.*, 1989]), such as *kind, sort*, *type,* etc. which are relatively empty as regards meaning (although see [Vossen, 1990] for a more detailed discussion on this issue). It was also found that very general predicates such as RELATED_TO and PROPERTY, which are often used as the meaning of prepositions and adjectives, were too vague for our purposes.

Another difficulty was incorporating information contained in relative clauses into this framework, since the way in which they relate to the genus is difficult to predict using pattern matching on the VP structure. Consider the following definitions for 'blood':

VOX
sangre: *líquido* que lleva en suspensión células
de distintas formas y funciones ...
lit. blood: *liquid* which carries in suspension
cells of different forms and functions ...
LDOCE
blood: red *liquid* which flows round the body

As we can see, the genus terms have the same meaning, but if we were also to pair the verbs of which they are the subject we would get *lleva(carries)=flows,* which is wrong.

The main reason for the discrepancies in the list of pairs given in the previous section is not that words must inherently be defined in terms of different concepts in different languages but that each team of lexicographers chose to establish a different level of simplicity in their sense definitions. To mitigate the effect of this problem we should pair each English genus with the Spanish genus AND the genus of the genus. This would result in many more correct pairings, as seen from the *vehicle* example above.

6 Conclusion

One thing which is worth re-emphasizing is the use of a restricted vocabulary in LDOCE. Without it, the above experiment would have yielded little result. Conversely VOX's unrestricted vocabulary reduced the number of possible correct pairings and consequently the success of the experiment. One effect VOX definitions have is to make the substance hierarchy deeper since their lexicographers tend to use genus which are fairly specific in meaning thus creating more nodes between the root 'substance' and the leaves.

If we had an LI for nouns available, it would be possible to discover translations of new words by defining them in terms of the LI and then searching the Target Language's lexicon to see whether there was a translation for the word or whether a paraphrase or a neologism was necessary. Also, it is worth considering the possibility of using a MR version of a bilingual dictionary to automate the pairing of senses prior to genus and argument matching.

7 Related Research

The research reported here is closely related to the Esprit-project AQUILEX. The aim of this project is the construction of large lexicons from lexical databases, such as MR dictionaries, for use in natural language processing applications. Within this project, [Vossen, 1991] presents a similar study to the one here but where Dutch and English are taken as the languages studied and where a much more extensive study is carried out using MR versions of two monolingual and one bilingual dictionary.

The work reported in [Klavans *et al.,* 1990] uses corpora and a bilingual dictionary to construct and update a bilingual database. The emphasis there is on the actual state of the two languages studied and on the way human translators actually map words and phrases onto other languages. The emphasis is therefore not on a LI but on coverage and actual use of bilingual equivalents.

Acknowledgements

Our thanks to the Longman Group for making the machine readable version of LDOCE available to us; to Trinity Hall, Cambridge, for a travel grant; to our supervisors, Ted Briscoe and Ann Copestake for their help, and to two referees for comments on an earlier version of this paper.

References

[Alshawi, 1989] Alshawi, H. Analysing the Dictionary Definitions. In: B. Boguraev and T. Briscoe (eds.), *Computational Lexicography for Natural Language Processing.* Longman Group Limited, Chapter 7, 1989.

[Carroll et al., 1988] Carroll, J., B. Boguraev, C. Grover and T. Briscoe. A Development Environment for Large Natural Language Grammars. Technical Report 127, Computer Laboratory, University of Cambridge, 1988.

[Copestake, 1990] Copestake, A. An Approach to Building the Hierarchical Element of a Lexical Knowledge Base from a Machine Readable Dictionary. *Inheritance in Natural Language Processing, Workshop Proceedings,* Tilburg, The Netherlands, 1990.

[Gili Gaya, 1988] Gili Gaya, S, *VOX Diccionario Manual Ilustrado de la Lengua Española.* Bibliograf S.A., Spain, 1988.

[Gazdar et *al.,* 1985] Gazdar, G., E. Klein, G. Pullum, and I. Sag. *Generalised Phrase Structure* Grammar, Blackwell, England, 1985.

[Hutchins, 1986] Hutchins, W. J. *Machine Translation - Past, Present and Future.* Ellis Horwood Limited, England, 1986.

[Klavans e*t al.,* 1990] Klavans, J. and E. Tzoukermann. The BICORD System - Combining Lexical Information from Bilingual Corpora and machine Readable Dictionaries. *Proceedings of COLING-90,* 1990.

[Nirenberg, 1987] Nirenberg, S. (ed). *Machine Translation - Theoretical and Methodological Issues.* Cambridge University Press, 1987,

[Plowman, 1990] Plowman, D. A. Extraction and Utilisation of Knowledge from Machine-readable Dictionary Definitions. University of Cambridge Master Thesis, 1990.

[Procter, 1978] Procter, P. (ed). *Longman Dictionary of Contemporary English.* Longman Group Limited, 1978.

[Slocum, 1988] Slocum, J. (ed). *Machine Translation Systems.* Cambridge University Press, 1988.

[Trujillo, 1990] Trujillo, I. A. Knowledge Extraction and Utilization from Spanish Dictionary Definitions. University of Cambridge Master Thesis, 1990.

[Vossen *et al.,* 1989] Vossen, P., W. Meijs, and M. den-Broeder. Meaning and Structure in Dictionary Definitions. In: B. Boguraev and T. Briscoe (eds.), *Computational Lexicography for Natural Language Processing.* Longman Group Limited, Chapter 8, 1989.

[Vossen, 1990] Vossen, P. The end of the chain: where does stepwise lexical decomposition lead us eventually? *Proceedings of ike 4th Functional Grammar Conference,* Copenhagen, Denmark, 1990.

[Vossen. 1991] Vossen, P. Comparing noun-taxonomies cross-linguistically. Working Paper No. 014, ESPRIT BRA-3030 Acquilex, English Department, University of Amsterdam, 1991.