Current Techniques in Lexical Information Retrieval and Manipulation

Antonio J. Moreno Ortiz

Abstract

The aim of this paper is, first, to present a brief account of current techniques used in NLP concerning lexical information needs, acquisition, retrieval, representation, and handling, offering a contrastive view of the most significant approaches; secondly, to offer an overview of what should the ideal NL system look like in order to account for today's needs. In doing so, I will briefly describe "Linguist's Workbench" a program that offers several facilities for the linguist.

1 Introduction

Karen Sparck Jones (1994) refers to four well differentiated phases in the history of Computational Linguistics, the MT phase (late 1940's to late 1960's), the AI phase (late 1960's to late 1970's), the grammatico-logical phase (late 1970's to late 1980's) and finally, from the 1980's onwards, the "massive data-bashing period." By this, she means the ever growing trend towards using computers to analyse and handle massive amounts of lexical data. Further, we could distinguish two trends in this phase (Bindi et al. 1991): lexicalism (the major trend of the 80's) and stochasticism (the trend of the 90's).

In sum, we have witnessed the move towards "data-driven" linguistics. Computers are used to extract lexical information from raw data in several ways, but mainly two: machine-readable dictionaries (MRDs) and corpora, directing the two main trends mentioned above. Lexicalism was favoured by the prosperity of lexicalist theories of language. The central role of the lexicon in the present day picture of linguistics (computational or not) is all but taken for granted, and any NL system nowadays has in the lexicon its central component and/or its main objective.

However, the growing complexity of techniques is an inevitable consequence of this evolution; in the following pages I will try to shed some light on present day system architectures, capabilities, and limitations. My goal is not to enumerate each NLP system proposed so far, but rather, to describe the most outstanding and generally employed approaches in state-of-the-art computational lexicography praxis.

2 System features

2.1 Lexicon Based

The common feature that all present day NLP systems share is that they are invariably lexicon-centred. Nicoletta Calzolari (1994: 267) asserts that "It is almost a tautology to affirm that a good computational lexicon is an essential component of any linguistic application (...)"

It was long ago realised that the so-called "lexical idiosyncrasies of language" where not to be regarded as such, but as a substantial, central component of language itself. The "lexicon bottleneck" (Briscoe 1991; Levin 1991), as this situation used to be referred to, was an important issue of debate for some time. The creation of a well-structured lexicon is thus the first step in any system that aims to deal with lexical information appropriately.

2.2 Theory Independence

The co-operation between different research teams from different countries with different methodologies and linguistic backgrounds, fostered chiefly by EU funding and industrial interests, has brought about a predictable consequence. Linguists had to face the shortcomings of their linguistic theories and were forced to reach a consensus if they wanted to achieve greater efficiency. Linguistic eclecticism is now commonplace in any NLP project, linguistic theories are evaluated against performance and cost-effectiveness, in terms of money, time, and effort; and ideas from different theories are put together and employed in order to achieve a positive result.

2.3 Reusability

Innumerable references to this feature can be found in the literature. Early systems were, almost invariably, ad hoc implementations of a given linguistic theory, or single applications that carried out a given task, commonly fruit of the efforts of an isolated researcher or group of researchers. It was soon realised that new applications had the same needs of previous ones, whose resources could not be exploited because they were application-specific. This often meant having to build those resources again, with the subsequent waste of time and effort. In fact, this problem has been, and still is, the plague not only of computational linguistics, but of information science in general.

We support the view that a NLP system is a dynamic, ever changing framework. All its components should work together synergically in order to take full advantage of its capabilities and to enrich itself. The concept of reusability acquires a new dimension in this perspective. The point is not just to be able to re-use components, such as a morphological tagger or a syntactic parser so that they can be used later, say to retrieve information from corpora; rather, our efforts should be aimed from the very beginning at achieving total integration of system components and close interaction between them. The metaphor of a toolbox with several small tools that

work together in a like manner reflects this idea fairly well. Our system, *Linguist's Workbench,* follows these integration guidelines.

2.4 Multifunctionality

That a system have multifunctional capabilities is a direct consequence of the above feature. A properly integrated system should be able to carry out several tasks and serve a number of purposes. We highlight the production of the following end-user applications: intelligent spell checkers, on-line thesauri, grammar checkers, CALL applications, translation assistants, natural language interfaces, etc. For a comprehensive account of present and future applications, see for example Marcos Marín (1994).

3 System components

In Figure 1, we show in a simplified way what the main components of a typical NLP system are and how they interact with one another. This diagram does not mean to be exhaustive, and may not be in accordance with some existing systems. Its aim is to exhibit our idea of the exemplary system, while reflecting most features found in *Linguist's Workbench*.

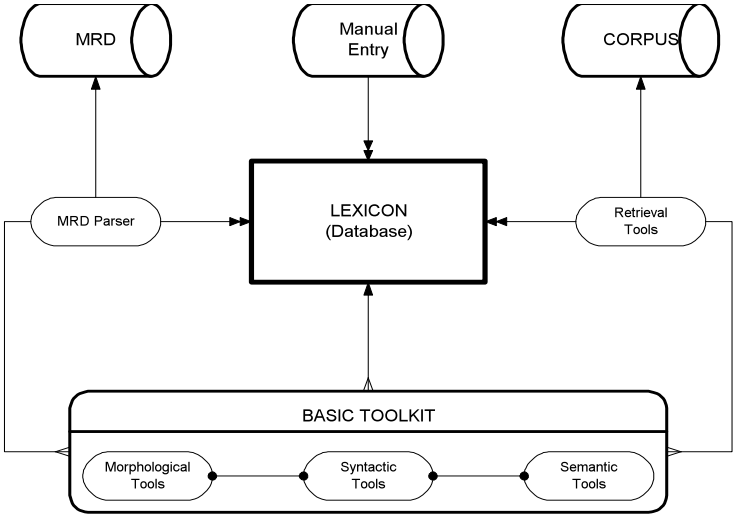

Figure 1: Basic Components of a Typical NLP System

3.1 Lexical Information Sources

Two are the basic sources of information for the creation of a NLP system: machine-readable dictionaries and corpora. These two sources should not be seen as opposed, but rather as complementary. MRDs are used to populate the Computational Lexicon (CL) initially, and is the basis of it. The use of computer corpora is a widespread technique in computational linguistics and lexicography (see, for instance, Pérez Hernández, 1994); its main function in the creation of the CL is to refine the information extracted from the MRD as well as to contribute with further information.

The role of direct manual entry of information derives from the following situation. Printed dictionaries are a rich source of lexical information. The obvious drawback of using MRDs to extract the basic information for the CL is that they were designed for humans, not machines; this means that their entries have to be *parsed* so that their information can be homogeneously extracted and entered in the database. We might distinguish three kinds of information to be found in a common

dictionary (Boguraev, 1994), information related to *form* (spelling, syllabification, phonetic description, and variants)*, function* (distributional behaviour, i.e. grammatical subcategorization)*, and meaning*. The first two types of information are usually well formatted and can relatively easily be retrieved; retrieving information about meaning is not so straightforward. Most work done in this direction has only been able to extract relationships among words or word senses, e.g., WORDNET (Miller, 1985). Proper extraction and normalisation of meaning would call for advanced language understanding capabilities, and although much progress has been made, the catch-22 situation (Boguraev and Briscoe, 1989) has not yet been resolved.

To summarise, only the integration of these three lexical sources will fulfil the necessities of a full-fledged CL.

3.2 The Computational Lexicon

Following this integrational approach, the
compilation of an exceptionally detailed lexicon, has been the focus of the project conceived and conducted by Martín Mingorance (1984, 1990) for several years. The valuable information so compiled has been modelled and structured in the shape of a computational lexicon that is at the core of *Linguist's Workbench*, so that the information contained in it can be used by the rest of the components. A thorough description of this lexical database (LDB) falls outside the scope of this paper; we refer the reader for an in-depth discussion of its features and design in terms of conceptual modelling to Moreno Ortiz (1994).

3.3 The Toolkit

The toolkit is the set of programs designed to manipulate the other elements in the system. A satisfactory level of integration is necessary in order to achieve good results. The idea of modularity is certainly relevant in this respect. We have designed the toolkit piece by piece, in such a way that each of the programs is able to perform a given task and whose output can be employed by other programs in the toolkit. For a practical application of one of the features in *Linguist's Workbench's* toolkit, concerning bilingual corpus work, see Pérez Hernández (this volume).

Although not indicated in the diagram, the set of tools should be handled by a Graphical User Interface (GUI) in order to take full advantage of its possibilities in a user-friendly manner (see Moreno Ortiz and Pérez Hernández, in progress).

4 Lexical Acquisition in *Linguist's Workbench*

Linguist's Workbench is a multifunctional system currently under development. It has most of the features we have discussed so far, and it incorporates multilingual capabilities for English and Spanish. The focus is on the lexicon, both as a means and as an end in itself. It includes such features as generation and recognition of forms, morphological tagging, lemmatised text search, etc.

The basic philosophy that has guided its creation is self-enrichment. Thus, the corpus is constantly used to check and enrich the lexicon in a semiautomatic fashion.

It is our idea of this process of lexical information acquisition that we would like to put forward next. This process is shown in the form of a typical flow chart in Figure 2.

We will illustrate this process with an example. Let us suppose that our lexicon only contains the sense for the verb HIRE:1. "to engage the temporary use of at a set price; rent." with the corresponding selectional restriction for Argument 2 (Od) [-*Human*] and the semantic function [*Artifact*] and the parser is confronted with the set of examples shown in Figure 3. It would accept as normal examples 1 to 6, because "video camera," "car," "motorhome," "van," etc. are described in the LDB with the features mentioned above, but it would find trouble to analyse example 7, as "driver" would be described as [+*human*]; then it would prompt the user to check if this kind of pattern is common or not (e.g. an occasional metaphorical deviation); the user would reply affirmatively and a new pattern would be included in the lexicon as having a human object. The system then would inquire further information to complete the LDB entry (e.g. semantic function [*profession*]); The rest of the examples ("architect," "therapist," "mercenary," etc.), already defined as such in the LDB, would be analysed without problem.

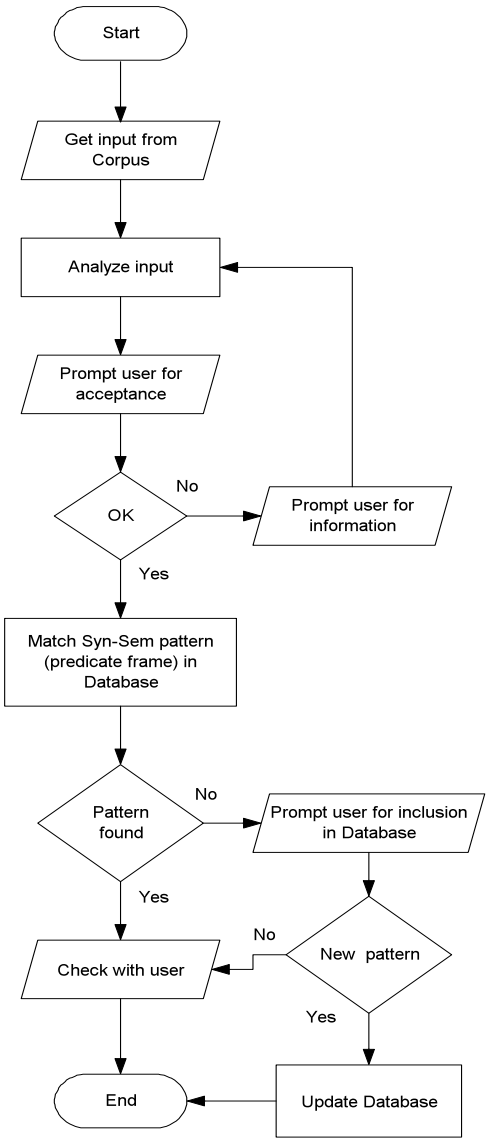

Figure 2: Lexical acquisition process

This is, of course, a simplification of the process with a very straightforward example. The actual lexicon is already rich enough to contain this pattern. Also, in most cases the user would need a higher number of occurrences to check if a new pattern should be included. A great deal of other problems are usually encountered. For instance, proper nouns, mostly ignored so far in NLP, are a rich source of problems and a large proportion of them is found in certain types of texts (e.g. newspapers) (Conlon et al., 1994). Thus, the names "Anthony Hopkins" or "Lambretta" would probably have to be entered in the lexicon and given their semantic features.

```
1. TH> LEST YOU FORGET You can now hire video cameras at Gatwick Airport as y   M07-54.TXT
2. ntertainment for all ages.<LTH> Hire a luxury Motorhome from Major Travel    M07-54.TXT
3. g Harbour (prepare for sharks). Hire a car and venture north to Palm and     M1-154.TXT
4. ll probably work out cheaper to hire a van and move your own things. Try t   M07-54.TXT
5. summer offers<LTH> canal boats. Hire a canal boat and experience life at t   M07-54.TXT
6. e park that it's a good idea to hire a car in the small town of Jasper.      M07-54.TXT
7. d car. As non-drivers we had to hire a driver with our Renault. This incre   M07-54.TXT
8. doubted that he could afford to hire an architect to design his dream hou    M55-99.TXT
9. e new NHS system, even a GP can hire a yoga therapist for his or her patie   M07-54.TXT
10.ng psychotic, giving the cue to hire Anthony Hopkins six years later.<LTH    M55-99.TXT
11.>  Have your own Roman Holiday: hire a Lambretta or Vespa, put on your cam   M07-54.TXT
12.ecting a baby? If you'd like to hire a maternity nurse to tide you over th   M07-54.TXT
13.this, Viscount Aimar decided to hire mercenaries in Gascony and denounce   LIONHEAR.TXT
```

Figure 3: Sample concordance lines

References

Bindi, R., Calzolari, N., Monachini, M. and Pirrellim V. (1991) "Lexical Knowledge Acquisition from Textual Corpora: A Multivariate Statistic Approach as an Integration to Traditional Methodologies," in L.M. Jones (ed.) *Using Corpora*. Proceedings of the 7th Annual Conference of the UW Centre for the New OED and Text Research, Oxford University Press, Oxford. pp. 170-196.

Boguraev, B. and Briscoe, T. (1989) *Computational Lexicography for Natural Language Processing*. Longman, London and New York.

Boguraev, B. (1994) "Machine-Readable Dictionaries" in A. Zampolli and N. Calzolari (eds.) (1994).

Briscoe, T. (1991) "Lexical Issues in Natural Language Processing," in E.F. Klein and Veltman (eds.) *Natural Language and Speech*, Springen-Verlag, pp. 39-68.

Calzolari, N. (1994) "Issues for Lexicon Building," in A. Zampolli and N. Calzolari (eds.) (1994). pp. 267-281.

Kiefer, F., Kiss, G. and Pajzs, J. (eds.) (1992) *Papers in Computational Lexicography. COMPLEX' 92.* Linguistic Institute Hungarian Academy of Science. Budapest.

Levin, B. (1991) "Building a Lexicon: The Contribution of Linguistics" International Journal of Lexicography, Vol. 4, No. 3.

Martín Mingorance, L. (1984) "Lexical Fields and Stepwise Lexical Decomposition in a Contrastive English-Spanish Verb Valency Dictionary," in R. Hartmann, (ed.) (1984) *LEX'eter'83: Proceedings of the International Conference on Lexicography*. Max Niemeyer, Tübingen. pp. 226-237

Martín Mingorance, L. (1990) "Functional Grammar and Lexematics in Lexicography," in J. Tomaszczyk and B. Lewandowska-Tomaszczyk (eds.) (1990) *Meaning and Lexicography*. John Benjamins. Amsterdam. pp. 227-253.

Marcos Marín, F. A. (1994) *Informática y Humanidades.* Editorial Gredos, S.A., Madrid.

Moreno Ortiz, A. J. (1994) "Semantic Modelling and Lexical Knowledge Representation" Procesamiento del Lenguaje Natural, Revista nº 16, Abril de 1995. pp. 44-60.

Moreno Ortiz, A. J. and Pérez Hernández, M. C. (in progress) "Building an Interface for Computational Lexicon Entries"

Pérez Hernández, M. C. (1994) *Corpus-Based Bilingual Lexicography: The Use of Computerized Corpora for the Identification of Translation Equivalents between English and Spanish* MA Dissertation, University of Exeter.

Pérez Hernández, M. C. (1995) "Using Corpora in Bilingual Lexicography" in this Volume.

Sinclair, J. M. (1993) "Lexicographers' Needs" *Zeitschrift für Anglistik and Amerikanistik.* XLI Jahrgang 1993. Langenscheidt. Berlin.

Sparck Jones, K. (1994) "Natural Language Processing: a Historical Review," in A. Zampolli and N. Calzolari (eds.) (1994). pp. 3-16.

Zampolli, A. and Calzolari, N. (eds.) (1994) *Current Issues in Computational Linguistics: in Honour of Don Walker*. Giardini editori e stampatori, Pisa and Kluwer Academic Publishers, Norwell, MA. pp. 267-281.