[*Translating and the Computer 19*: Papers from the Aslib conference… 13 & 14 November 1997]

LEXICAL BOTTLENECK IN MACHINE TRANSLATION AND NATURAL LANGUAGE PROCESSING: A CASE STUDY

M. Victoria Arranz*
Department of Language Engineering, UMIST,
PO Box 88, Manchester M60 1QD

Abstract: This paper emphasises the need to develop efficient lexical knowledge acquisition techniques in order to tackle problems related to the so-called *lexical bottleneck.* Bearing this in mind, a semi-automatic technique for semantic clustering and word sense disambiguation is proposed. The main principles behind this method are the extraction of knowledge on a sublanguage basis and from actual corpora. Clustering and disambiguation are carried out by means of the similarity measure *Dynamic Matching.* Further, the development of a domain-specific semantic ontology is also reported.

INTRODUCTION

The *lexical bottleneck* has emerged as a major problem in natural language technology applications. As it is well known, the performance of the systems is closely related to the coverage of their lexicons. The information contained within the lexical items is required for a series of tasks, such as language analysis in NLP and MT systems, and it is therefore essential to ensure that this information is as accurate and complete as possible. Further, lexicons which hold incomplete or incoherent information can be detrimental to the functioning of their systems. Some of the lexical gaps that can be encountered within the lexicon are, for example, *missing words, compound words, word senses, collocations, idioms and phrases, metaphors, lexical orphans, lexical constraints, individual grammar properties* and *synonyms* (Zernik (1)). In terms of performance, they might affect different applications in a variety of ways, but in general they all have a negative effect on them. For instance, *missing words* represent a problem for analysis processes such as *parsing*, which will fail if reaching an unknown word. Zernik (1) suggests that "for a program to analyze a body of text 'automatically', a lexicographer must first pave the way by skilfully identifying problematic cases and by crafting them into the lexicon". As a consequence, lexical acquisition has become a necessary step to try and boost any system's performance.

Bearing these considerations in mind, the Consortium for Lexical Research was founded (CRL (2)), as an attempt to share lexical data, tools and also the results obtained in individual researches. However, although work on the acquisition of lexical knowledge began about 15 years ago, there is still a considerable amount of information which is required by our systems and which is unfortunately unavailable in the existing computational lexicons. The problem of lexical knowledge acquisition has thus become the concern of a large number of research studies currently taking place (Velardi and Pazienza (3), Church and Hanks (4), Vanderwende (5)). Such researches fall into two general categories (Boguraev and Pustejovsky (6)): those that encode the lexical knowledge bases by hand and those that extract the knowledge

* Researcher sponsored by the *Departamento de Educación, Universidades e Investigación* of the Basque Government, Spain.

automatically from on-line resources. With regard to manual acquisition, albeit still in use due to its low initial costs, it is the most labour intensive method, in particular when aiming at the construction of broad-coverage lexicons for MT. In what refers to automatic acquisition, this has become increasingly popular over the past few years, with the application of new analysis approaches, specially those based upon statistical techniques.

LEXICAL KNOWLEDGE ACQUISITION FROM CORPORA

The types of on-line text resources employed during the automatic acquisition of lexical information are either machine-readable dictionaries (MRDs) or corpora. Both types are important sources of information and at the same time both present their drawbacks.

One of the arguments against MRDs is that in spite of being the largest repositories of organised knowledge about words that are available, they are far from complete, consistent and coherent. In fact, they could be described as biased, since they reflect the lexicographer's interests and inclinations at the time of compilation (Zernik (1)). As a consequence, it is no easy task to evaluate the extent to which they are going to be of help and not in detriment of the newly derived lexicons (Boguraev and Pustejovsky (6)). Furthermore, language is a dynamic object undergoing constant evolution, while dictionaries are static objects and the information they contain could easily become obsolete if it is not duly updated. Therefore, critical assessment of the information comprised in an MRD should be carried out before proceeding to its use. Finally, MRDs also lack customisation, i.e., they might include many items which are of no relevance to some particular application and lack the required domain-specific terms and definitions. Unless the dictionary has been written *a priori* with the purpose of covering a particular semantic domain, it will not be a reliable source for the extraction of that sublanguage-specific knowledge.

It is essential then that researchers have access to other reliable on-line resources from which to carry out the knowledge acquisition (KA) process. One such reliable resource is the corpus, which can provide us with a new and more realistic understanding of language, as well as help us to achieve tasks which would have been almost unthinkable in what regards human introspection (Aijmer and Altenberg (7)). Computerised corpora represent a valuable basis for comparing different varieties of language as well as exploring the quantitative and probabilistic aspects of the language.

Given that the present section has briefly introduced the concept of *sublanguage,* the next one defines the term and explains the advantages of approaching NLP and MT on a sublanguage basis.

SUBLANGUAGE-BASED KNOWLEDGE ACQUISITION

In general terms, *sublanguages* represent those subsystems within a certain natural language that we normally refer to as "the language of medicine" or "the language of law", etc., because they behave like well defined languages which happen to be used by specialists in various fields. The reason to focus on the study of sublanguages rather than on that of unrestricted language is a practical one: this approach allows us to achieve better results as well as eases the discovery task since we consider a constrained domain which presents restricted lexis, syntax and semantics. As Grishman and Kittredge (8) emphasise, the diversity of language encountered in a sublanguage is considerably smaller and more systematic in structure and meaning than that of the whole language. This proves to be so, for example, in the case of the MT system TAUM-METEO, developed at the *Université de Montréal* to translate weather reports from English to French, which is a clear example of an outstanding success for sublanguage-based approaches to MT, and text processing in general.

The difference in linguistic behaviour that characterises sublanguages is also studied by Lehrberger (9), who, in an application of the concept of sublanguage to MT, proposes several factors as an aid to characterise a sublanguage. These factors are the following:

- limited subject matter,
- lexical, syntactic and semantic restrictions,
- "deviant" rules of grammar,
- high frequency of certain constructions,
- specific text structure,
- use of special symbols.

Bearing this in mind, the use of corpora as on-line resources can be considered the best possible choice. In particular when working with a very restricted sublanguage, the information provided in MRDs is bound to be very general and therefore incapable of providing us with the necessary knowledge.

SEMANTIC CLUSTERING AND DISAMBIGUATION

Our Approach

For the reasons above explained, the current work focuses on the extraction of domain-specific lexical information from corpora. In particular, we are interested in the issues of semantic clustering and word sense disambiguation in a Unix-specific corpus and by means of a semi-automatic KA process. Regarding the paradigm approach to be adopted, despite the present popularity of probabilistic methods, we are well aware of the fact that the acquisition of knowledge from sublanguage-restricted corpora by such methods can prove problematic. As it is well known, purely statistical techniques require large amounts of data in order to

obtain reasonable results, which is not always practical for current NLP systems. This is particularly the case when working with domain-specific texts, since these are not usually available in such large amounts. For instance, the present research takes place on a Unix-specific corpus of about 100,000 words. To this problem one should add the fact that probabilistic processes are completely opaque to the human specialist, making it difficult to judge intuitionally uninterpretable results. Therefore, following on the idea of KA as an evolutionary process detailed by Tsujii and Ananiadou (10), we attempt to develop a quasi-statistical approach to sublanguage-based semantic KA and disambiguation from small corpora.

Dynamic Matching Similarity Measure

This technique is based on the technique *Dynamic Alignment* employed in *EBMT* by Somers *et al.* (11), and it represents the main strength of our semantic clustering and disambiguation processes. It allows us to compare the degree of similarity between two words by looking at their contextual surroundings. Firstly, it discovers all potential matches between two sets of individual words which form the contexts for the pair of words, and attaches a value to each match according to its level of importance. Then, it calculates the strongest match for the pair of contexts, establishing thus a value on their similarity relation (see Arranz *et al.* (12)).

Given that the keywords under comparison are bound to occur in more than two contexts, while *Dynamic Matching* only considers two contexts at a time, this process is repeated for all possible combinations of context matchings. Thus, the best match value is calculated for each pair of contexts, which results in a *correlation matrix* containing the match values for all the contexts. An example of a *correlation matrix* for the pair of words *discussed/VBN* and *listed/VBN* (see Arranz *et al.* (13)) is the one presented below, where the cluster for contexts (2,3) presents the highest value and thus symbolises the strongest relation for the terms under study:

```
% dynamic discussed/VBN listed/VBN +5 -5 < corpus
Post context length set to 5
Pre context length set to 5
CIWK data read. 9 records found.
0  1  2  3  4  5  6  7  8
-------------------------
0 : 5 10  7  6  8 10  7 10
1 :    8  7  8  3  5  5  4
2 :      27  7 11  9  5  4
3 :         6  8  6  4  4
4 :            14  9  6  4
5 :               14  1  3
6 :                  4  5
7 :                     5
8 :
```

Semantic Clustering and Disambiguation

As mentioned in the previous section, the semantic clustering and disambiguation processes rely on *Dynamic Matching* in order to discover the degree of similarity between the candidates to semantic clusters. Once their corresponding contexts have been matched and compared and the *correlation matrix* created, the strongest cluster is determined by means of a simple clustering algorithm which operates as follows:

1. The pair of contexts presenting the highest value in the matrix is selected as the core of the cluster.
2. Then, all remaining contexts are considered in turn and added to the cluster if their correlation value is above a certain pre-established threshold with respect to more than half the contexts already in the cluster.
3. Step 2 is repeated until no more contexts are appropriate for that cluster.

Basically, the clusters representing the different meanings (or different usages) of a pair of words are created based upon the values obtained in the matrix. The idea behind this is that the values included in the matrix offer a quantitative estimate of the similarity between the contexts involved and thus, between the words they represent. When all the possibilities for clustering are exhausted for a particular semantic group, the process is repeated for the next strongest cluster, as an attempt to extract any further possible meanings or usages.

Applying Structural Constraints

One of the main principles behind our similarity measure *Dynamic Matching* is its *linear ordering constraint.* Despite the fact that context matching techniques have been frequently used in NLP, the position of the words in the contexts is not normally considered of importance. In contrast, our context matching technique applies structural constraints to the context similarity measurement process, making word order relevant (Radford *et al.* (14)). That is, during the context matching process carried out by *Dynamic Matching*, maximal sets of individual matches are discovered on the grounds that no match can cross any other match in the set. This ensures that corresponding matching terms occur in the same order in both contexts. Figure 1 shows an example of the functioning of this *linear ordering constraint,* where the crossing constraint avoids having simultaneously the two matches on the left involving the first instances of *the/DT* and those of *of/IN.*

The reason for applying this *linear ordering constraint* lies in the importance of word order in natural language. Although no parsing is used during our lexical KA processes, it is observed that semantic relations such as Subject-Verb can hold essential information for disambiguation. Therefore, it is considered important that word order be maintained for our context matching. Nevertheless, in spite of the importance of this linear word order, it is often

discerned that the same semantic meaning can be phrased in different manners, thus complicating our disambiguation task and being the cause for ill-formed clusters. Two such cases which occur frequently in our Unix corpus are that of active-passive structural changes within the sentence, and that of the changes within compounding collocational nominal constructions. In order to solve these problems, the following two modules are developed, which interact with *Dynamic Matching:*

1. a **grammar mapping module:** which is in charge of restoring the canonical order within the contexts before the matching takes place. This helps to relax the *linear ordering constraint* imposed by our matching tool (cf. Arranz (15) for more details).
2. a **compound-collocation mapping module:** which recovers clusters lost due to the change of order within compound structures. This module recognises potential compounding constructions and allows for any permitted form of such constructions to be matched with them.

Albeit rather simple, both modules become an essential part of our clustering and disambiguation processes, due to the considerable number of cases they handle in our particular corpus. This Unix corpus presents a very high number of both passive constructions and compound-collocational elements with the order of their components altered.

SUBLANGUAGE ONTOLOGY DESIGN

As previously established in this paper, an important idea behind this work is that semantic clustering, and the measuring of the semantic information in general, does not require the structural analysis and annotation of the corpus. Although parsing the input data *a priori* is the usual approach in other semantic-related studies (for example, in Hirschman *et al.* (16) and Habert *et al.* (17)), *Dynamic Matching* does not require such pre-processing. This tool identifies the keywords' usage constraints which are reflected in their contexts, and uses them as cues for disambiguation. It can thus be inferred that the concept of *meaning* is viewed in terms of *usage,* where *usage* reflects every particularity of a word and thus helps to establish meaning according to the multiple semantic aspects that can be encountered for a particular word.

Moreover, this multi-aspect characteristic of a word's meaning shows its dynamic nature, i.e., meaning is not a static object but a dynamic one, which represents the evolutionary character of a word's semantics. Therefore, meaning cannot be defined in terms of fixed primitives, but rather in terms of pre-established classes and their actual inter-relations. In agreement with this, and closely related to Pustejovsky's notion of *qualia structure* (in Pustejovsky (18)), our research supports the existence of a semantic relation network within Unix which covers all the possible usages for the words included in that sublanguage. That is, when defining a word in that domain, this is to be done bearing in mind the relations into which such word may enter and thus judging purely on empirical usage-related terms.

The semantic ontology designed for Unix covers all the main content words (nouns, adjectives and verbs) encountered in our corpus and it classifies such elements by means of a rather simple hierarchy of function categories. Such categories reflect the various meaning aspects for each word as well as the relationships that can be established

a. between the pair of words to form a cluster, and
b. between the aspects contained in these words.

Further, these template categories are defined as *objects, actions, action arguments* and *classifying elements,* and they subdivide into several other subcategories so as to describe a term in as specific a way as possible. Details and examples on the adaptation of this theory and our domain-specific ontology for particular cases can be seen in the following section. It will just be added here, though, that the sublanguage ontology built is currently being used as reference for evaluation, so as to check the results from the clustering and disambiguation processes.

CASE DESCRIPTION

Prior to the implementation of the symbolic modules described above, this technique had already proved to be rather flexible (Arranz *et al.* (13)) by successfully discovering the different meanings of some of our cases and partially locating some of the senses of many others. Still, further work has been done to solve the remaining ambiguity and despite the fact that some of it still remains, some words which in earlier stages could only be partially disambiguated are fully disambiguated at this point. This section aims to describe two such cases in detail, so as to illustrate the actual functioning and performance of both the clustering and disambiguation processes. Furthermore, the particular semantic ontology for those semantic clusters will be shown, in order to exemplify the type of functional hierarchy built for every semantic group in Unix. At present, such ontologies are used as a reference for performance evaluation, but it is intended to develop a conceptual-graph-based knowledge base which will allow user-friendly access to the information for various purposes.

The first case to be described is that of the pair *CTRL/NN - SHIFT/NN.* Once all the submodules are applied with *Dynamic Matching,* the initially ambiguous cluster succeeds in having those individual non-related meanings filtered out. The set of contexts in which the elements of this cluster take place in the corpus and the contextual clusters obtained by the clustering process are detailed below:

Contexts:
0 :)/) then/RB the/DT value/NN used/VBN is/VBZ the/DT corresponding/JJ -*CTRL/NN*- character/NN (/(for/IN instance/NN ,/, `/`` ^D/NN '/"
1 : ,/, the/DT move/NN is/VBZ constrained/NNP ./. hold/VB the/DT -*CTRL/NN*- key/NN and/CC press/VB and/CC hold/VB the/DT MIDDLE/NN mouse/NN

2 : ,/, the/DT resize/NN is/VBZ constrained/NNP ./. hold/VB the/DT -*CTRL/NN*- key/NN and/CC click/VB the/DT LEFT/NN mouse/NN button/NN while/IN
3 : frame/NN header/NN or/CC outer/JJ border/NN ./. hold/VB the/DT -*SHIFT/NN*- key/NN and/CC click/VB the/DT LEFT/NN mouse/NN button/NN while/IN
4 : overall/JJ size/NN of/IN the/DT application/NN :/: hold/VB the/DT -*CTRL/NN*- key/NN ,/, press/VB the/DT MIDDLE/NN mouse/NN button/NN over/IN
5 : their/PRP$ output/NN appearing/VBG in/IN separate/JJ windows/NNS ./. typing/VBG -*CTRL/NN*- -L/NN redraws/VBZ the/DT screen/NN ,/, while/IN your/PRP$ erase/NN
6 : A/DT selection/NN is/VBZ made/VBN pending-delete/NN by/IN holding/VBG the/DT -*CTRL/NN*- key/NN while/IN clicking/VBG the/DT LEFT/NN or/CC MIDDLE/NN mouse/NN
7 : ,/, towards/IN the/DT end/NN ./. holding/VBG down/RB the/DT -*SHIFT/NN*- key/NN while/IN invoking/VBG find/NN searches/VBZ backward/RB through/IN the/DT

Clusters :
((1,((2,6),(3,7))),4)

As it can be observed, the resulting cluster shows that one common meaning has been found for both *CTRL/NN* and *SHIFT/NN,* where both function as **keyboard keys.** However, there is more to their meaning than this mentioned aspect. If we have a look at figure 2, which presents the corresponding semantic ontology for the pair of words, we will be able to observe that a wider semantic richness has been captured in the ontology by means of establishing the different meaning usages regarding the two words under study. For example, the shared boxes (those receiving arrows from both keywords) indicate the common meaning aspects for both words. Thus, *CTRL/NN* and *SHIFT/NN* share the aspects of being a *physical object* **key** (as already mentioned before) as well as of *undergoing a certain type of action:* **hold (down).** That is, these two terms can be considered semantically related in the Unix sublanguage when they share these usages.

Moreover, the category template also includes other information regarding certain meaning aspects that have been found in the corpus and which are exclusive to the word *CTRL/NN.* For example, those of being a *non-physical object* **character** or **-L** (the latter combines with the term *CTRL/NN* in order to refer to a particular Unix command), as well as that of *undergoing the action* **type.** To conclude with this case, it will just be added that when the meaning aspect *non-physical object* **character** occurs in the corpus with *CTRL/NN,* it takes place together with the *classifying element* **corresponding.** It is due to the aspects and relationships established between these aspects that the elements in the semantic cluster are defined and classified, given that semantic clustering and disambiguation take place according to the actual usage of the terms within the corpus.

The second case to be considered is that of *BEGIN/NNP - END/NNP.* Like in the previous case, the process succeeds in prioritising those contextual cues which are more relevant for clustering and disambiguation, and recognises the two candidate meanings while it rejects the only non-related sense. For a full description on this cluster, refer to figure 3, which presents the semantic ontology for both *BEGIN/NNP* and *END/NNP.* According to the information contained in the ontology, both keywords can be characterised by being a *non-physical object*

pattern, which can be modified by the following *classifying elements*: **special, first** and **last.** In addition, *BEGIN/NNP* and *END/NNP* are also defined by the type of *action they can undergo,* which is that of **use.** Furthermore, *action* **use** requires the use of the elements **special pattern $'** in the corpus. Similar to the previous cluster case considered, both words in this case also present their own meaning usages. For instance, *BEGIN/NNP* can also refer to an element which *performs the action* of **starting a program,** and *END/NNP* can represent the *non-physical object* **feature,** which *undergoes the action* of **demonstrating,** instead of performing it.

CONCLUSIONS

This paper is an attempt to provide some insight into an issue of current concern in NLP and MT: the *lexical bottleneck.* We begin by emphasising the advantages of sublanguage-based and corpus-based approaches over other existing methods. Then, focusing on the phenomena of semantic clustering and word sense disambiguation we present a tool for such purpose. This tool relies on *Dynamic Matching,* a word similarity measure which calculates the relation between two words by looking at the similarity between their respective contexts. Due to the domain-specific nature of the corpus as well as its small size, the basic statistical nature of the clustering process is combined with some symbolic submodules, which help to increase the robustness of the whole process.

As shown above, the results are very promising and *Dynamic Matching* succeeds in capturing the similarity between words based on the similarity between the information in their contexts. This information is expressed by means of semantic relations within the sentences, i.e., two words will be semantically related if they share the same content elements as subjects, objects, etc. Despite the fact that no structural analysis is required *a priori, Dynamic Matching* looks at the surrounding content words, which are indirect members of such syntactic relations. Further work is currently being done in order to eliminate the noise encountered in some of the clusters extracted by the system.

In addition, a Unix-specific ontology has been developed and this paper shows how it applies to two particular cases. This ontology classifies the information regarding some Unix-specific word in terms of the different usages that this word might present. In order to do so, it considers all surrounding content elements and the relations they establish both between themselves and with the words under study.

REFERENCES

1. Zernik, U. ,1991, "Introduction". In Zernik, U. (ed.). Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon. Lawrence Erlbaum Associates, Hillsdale, New Jersey.

2. CLR, 1991, "The Consortium for Lexical Research". Proc. of DARPA Speech and Natural Language Workshop. Pacific Grove, California.

[1] The dollar sign stands for the position of the keyword within the context.

3. Velardi, P. and Pazienza, M. T., 1989, "Computer Aided Interpretation of Lexical Cooccurrences". Proc. of ACL'89: 27th Annual Meeting of the Association for Computational Linguistics.

4. Church , K. W., Hanks, P. , 1990, "Word Association Norms, Mutual Information and Lexicography", Computational Linguistics. 16(1).

5. Vanderwende, L., 1994, "Algorithm for Automatic Interpretation of Noun Sequences". Proc. of COLING'92: 15th International Conference on Computational Linguistics.

6. Boguraev, B., Pustejovsky, J., 1996, "Issues in text-based lexicon acquisition". In Boguraev, B., Pustejovsky, J. (eds.). Corpus Processing for Lexical Acquisition. MIT Press, Cambridge.

7. Aijmer, K. and Altenberg, B., 1991, "Introduction" to Aijmer, K. and Altenberg, B. (eds.). English Corpus Linguistics. Longman, London.

8. Grishman, R., Kittredge, R,, 1986, "Preface". In Grishman, R., Kittredge, R. (eds.). Analyzing Language in Restricted Domains: Sublanguage Description and Processing. Lawrence Erlbaum Associates, Hillsdale, New Jersey.

9. Lehrberger, J., 1986, "Sublanguage Analysis". In Grishman, R. and Kittredge, R. (eds.), Analyzing Language in Restricted Domains: Sublanguage Description and Processing. Lawrence Erlbaum Associates, Hillsdale, New Jersey.

10. Tsujii, J., Ananiadou, S., 1993, "Epsilon: Tool Kit for Knowledge Acquisition Based on a Hierarchy of Pseudo-texts". Proc. of NLPRS'93: Natural Language Pacific Rim Symposium. Fukuoka, Japan.

11. Somers, H., McLean, I. and Jones, D., 1994, "Experiments in Multilingual Example-Based Generation". Proc. of the 3rd Conference on the Cognitive Science of Natural Language Processing. Dublin, Ireland.

12. Arranz, M. V., Radford, I., Ananiadou, S., Tsujii, J., 1995, "Tools for Sublanguage-Based Semantic Knowledge Acquisition from Corpora". Proc. of SEPLN'95: *Congreso de la Sociedad Espanola para el Procesamiento del Lenguaie Natural.* Bilbao, Spain.

13. Arranz, M. V., Radford, I., Ananiadou, S., Tsujii, J., 1995, "Towards a Sublanguage-Based Semantic Clustering Algorithm". Proc. of RANLP'95: International Conference on Recent Advances in Natural Language Processing. Tzigov Chark, Bulgaria.

14. Radford, I., Arranz, M. V., Ananiadou, S., Tsujii, J., 1995, "Dynamic Context Matching for Knowledge Acquisition from Small Corpora". In Bolasco, S., Lebart, L., Salem, A. (eds.). Analisi Statistica dei Dati Testuali. CISU, Rome, Italy.

15. Arranz, M. V., 1997 (forthcoming). Sublanguage-Based Semantic Clustering and Disambiguation from Corpora. PhD Thesis. CCL, UMIST, Manchester.

16. Hirschman, L., Grishman, R., Sager, N., 1975, "Grammatically-Based Automatic Word Class Formation". Information Processing and Management. 11.

17. Habert, B., Naulleau, E., Nazarenko, A., 1996, "Symbolic Word Clustering for Medium-Size Corpora". Proc. of COLING'96: 16th International Conference on Computational Linguistics. Copenhagen.

18. Pustejovsky, J., 1991, "The Generative Lexicon". Computational Linguistics. 17(4).

FIGURES

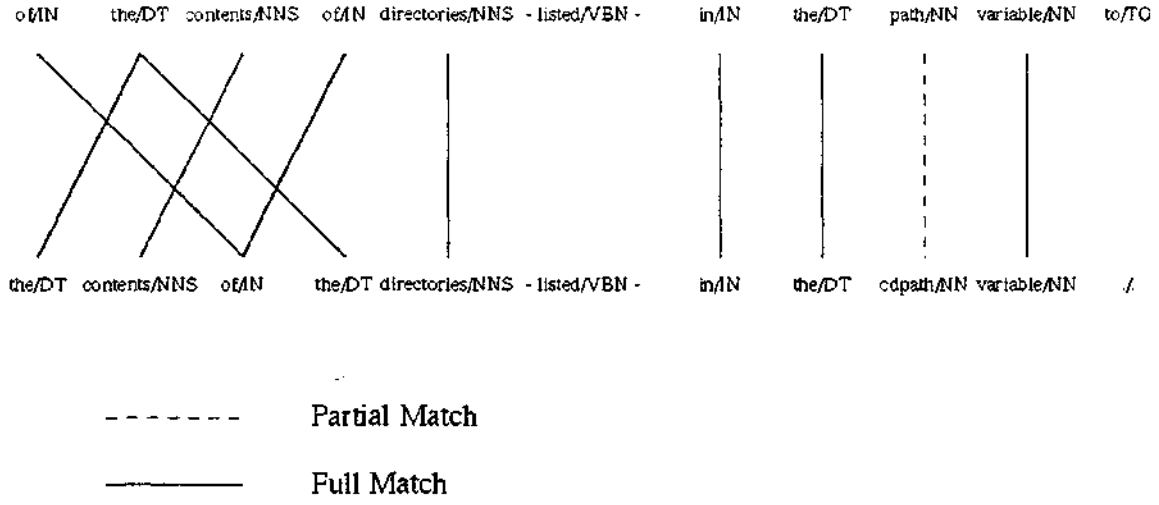

Figure 1: Example of Context Matching

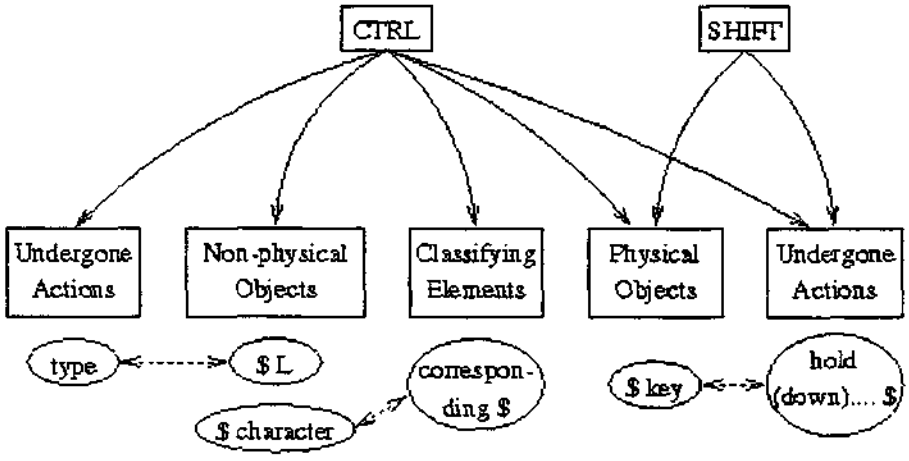

Figure 2: Semantic Ontology for the Pair *CTRL/NN - SHIFT/NN*

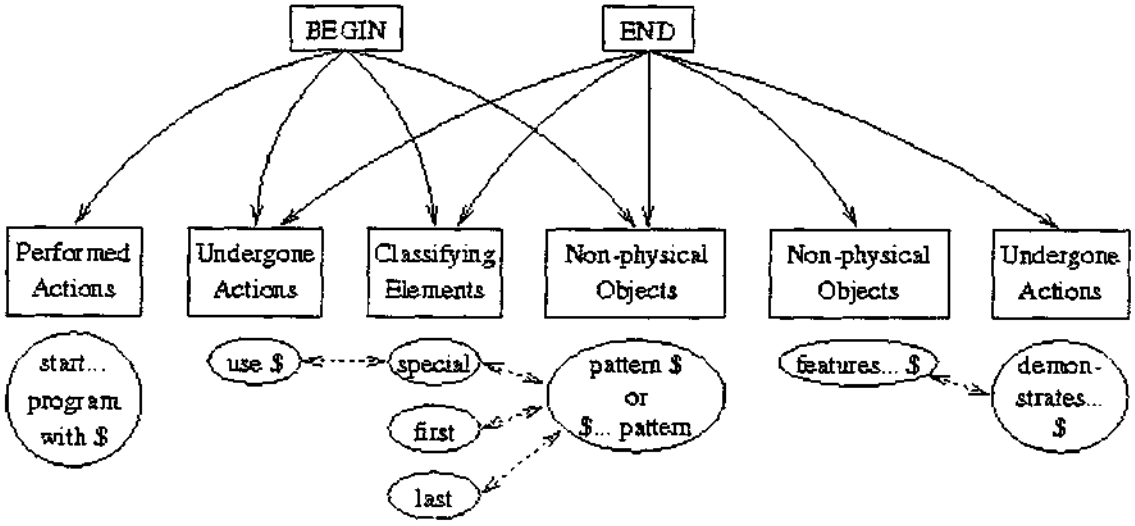

Figure 3: Semantic Ontology for the Pair *BEGIN/NNP - END/NNP*