Extraction of Semantic Clusters for Terminological Information Retrieval from MRDs

Gerardo Sierra*, John McNaught[†]

*Instituto de Ingeniería, UNAM Apdo. Postal 70-472 México 04510, D.F. gsm@pumas.iingen.unam.mx

†Centre for Computational Linguistics, UMIST P.O.Box 88 Manchester, U.K., M60 1QD jock@ccl.umist.ac.uk

Abstract

This paper describes a semantic clustering method for data extracted from machine readable dictionaries (MRDs) in order to build a terminological information retrieval system that finds terms from descriptions of concepts. We first examine approaches based on ontologies and statistics, before introducing our analogy-based approach that lets us extract semantic clusters by aligning definitions from two dictionaries. Evaluation of the final set of clusters for a small set of definitions demonstrates the utility of our approach.

1. Background

The majority of lexicographers recognise the need for dictionaries that, contrary to the alphabetical ordering of entries, help users to look for a word that has escaped their memory even though they remember the concept. From a semantic point of view, Baldinger (1980) identifies two kinds of dictionaries. The semasiological one corresponds to the viewpoint of the person interpreting the speaker, thus one starts with the form of the expression to look for the meaning. It is alphabetically arranged. The onomasiological type, from the perspective of the speaker, allows one to start from the mental object and look for its designations. It is arranged by concepts and splits up the fields of meanings. E.g., the name "green" may be found under the concepts "colour", "vegetable" or "inexperience".

In the context of computational lexicography, it has been shown that machine readable dictionaries (MRDs), which are conventional semasiological dictionaries, can be used for onomasiological searches. This is based on the assumption that semasiological dictionaries have the necessary information in the first place. Kipfer (1986) stated that a dictionary can be considered as a matrix that maps between word senses, and that an on-line dictionary can be entered via words or senses. She added that in an on-line dictionary it is possible to find a word by following semantic links or by genus. For example, if a user needs to locate the word expressing a group of ducks (flock), they can check the entry for duck. Calzolari (1988) stated that an on-line dictionary can be used to seek a word through examining definitions that contain words supplied by the user (clue words).

The success of an onomasiological search relies upon the accuracy of all clue words in the concept definition that might represent the target word the user is looking for. Since the user often does not employ precisely the same terminology as the indexed keywords or stored fulltext database, the retrieved words may be far from the concept desired. When the result is not satisfactory, the user can expand the query with closely related keywords which enhance the meaning, such as alternative forms, synonyms or cross-references. As a result, it has been found advantageous to expand the original query with closely related keywords (Fox, 1988). In addition to the user's own knowledge of expressing the same concept in alternative ways, a relational thesaurus brings related words together and thereby helps to stimulate their memory. Some systems provide an on-line thesaurus as a facility for the user in this regard. However, formalising a concept with the exact clue words is sometimes a heavy task for the user, but searching can become harder if the user has also to identify clusters of related keywords, particularly when the query is expressed in natural language. The systematisation of this task has been hence placed on the system, which can provide clusters during the search session in order to allow the user to select the best ones, or alternatively they can be automatically used by the system. In order to help the user focus on the search, it is convenient that the system produces and manages the semantic clusters transparently, without any intervention by the user. In fact, this is the goal of a userfriendly onomasiological search system, and the success of such a system relies on the accurate identification of the semantic clusters.

2. Survey of clustering

Clustering has been applied to almost every discipline. The process of identifying clusters has variously been called cluster analysis, classification, categorisation, taxonomy, typology or clumping, according to the discipline. The purpose of clustering varies from classification and sorting to the development of inductive generalisations (Anderberg, 1973). The primary goal of clustering is to collect together into clusters a set of elements associated by some common characteristic. Each element or member within a cluster A is strongly associated with each other because they share the same property, while members of other clusters show distinct characteristics from those of A. Clustering may alternatively be oriented either to discover the strongest association among members or to seek members which are isolated from each other. Clustering is often based on measurements of the similarity or dissimilarity between a pair of objects, these objects being either single members or other clusters. A cluster is defined by its members and often by the "central concept" with which all the cluster's members are associated (McRoy, 1992). This central concept could be the common characteristic, the particular conceptual parent or even any member when there is no need to specify the exact nature of the association among the members. The identification of the central concept relies on the variables that are used to characterise the elements of the problem, either the characteristics, attributes, class memberships or other such properties. Here, our focus is on semantic variables.

Clustering methods to identify semantically similar words are broadly divided into relation-based and distribution-based approaches (Hirakawa, Xu & Haase, 1996). The former analyse relations in an ontology, while the latter use statistical analysis. According to the terminology of Grefenstette (1996), these methods can be called knowledge-rich, based on a conceptual dependency representation. and knowledge-poor, based on distributional analysis. From a methodological point of view, there is, in addition to the above two approaches, a little known approach called the analogy-based approach. This employs an inferential process and is used in computational linguistics and artificial intelligence as an alternative to current rule-based linguistic models.

2.1. Relation-based clustering

Relation-based clustering methods rely on the relations in a semantic network or ontology to judge the similarity between two concepts. Since an ontology connects concepts, each located in a node, it is then possible to analyse either the taxonomic relations or just the conceptual distance between the nodes. A taxonomy lets us extract semantic relations, such as is-a or a-kind-of, to judge the similarity between two concepts by comparing their parent. A semantic network lets us derive similarity by determining the path-length or number of links between the nodes.

The most important lexical knowledge resources that provide a basic ontology for clustering are the semantic taxonomy WordNet (Fellbaum, 1998) and Roget's Thesaurus (1987).

2.1.1. WordNet

WordNet is an organised lexical resource of English nouns, verbs, and adjectives, widely used for relationbased applications. It is a hand-constructed system, designed on psycholinguistic principles at Princeton University. The lexicon relies, for each concept, on a set of words that can be used to express that concept, namely synonym sets or synsets for short. The semantic relations between concepts and other relations are represented by cross-references between synsets, such as synonymy and antonymy, hyponymy and hypernymy, meronymy (subordinate/ superordinate) and holonymy (part–whole relation), and troponymy (manner-of, in verbs). These relations allow the lexicon to be structured into hierarchies, so that we are able to request, e.g., a list of all superordinates in the hierarchy for a given concept.

In order to determine closeness in meaning among words, Agirre and Rigau (1996) introduce the measure of conceptual distance, which is the shortest length that connects two concepts in the hierarchical net. Their method enables us to check the closeness of candidate hypernyms for a given hyponym. This measure may be applied to cluster similar concepts, so that candidate concepts with the shortest path in the hierarchy should be clustered. On the other hand, Resnik (1995) suggests WordNet for semantic clustering on the basis of the information content shared by the synsets in comparison.

However, we may add that the technique of conceptual distance or number of links is highly dependent on the degree of density of coverage of the conceptual space in an area that the WordNet lexicographers have been able to achieve. Besides, it is also appropriate to point out some drawbacks observed by researchers applying WordNet to specific purposes (Arranz, 1998; Agirre & Rigau, 1996; Basili, Pazienza & Velardi, 1996): restricted types of semantic relationships; lack of cross-categorical semantic relations among nouns, verbs and adjectives; the sense distinctions are not always satisfactory; there are similar words that are not recognised in WordNet; tags in WordNet create overambiguity.

2.1.2. Roget's Thesaurus

Roget's Thesaurus has become rather popular for many applications. One reason is that it is a general thesaurus with broad vocabulary coverage, although it is likely to be missing many domain-specific words. Another reason is its well-organised structure in three hierarchical levels above the basic level of words, namely category. Grefenstette (1996) has used Roget's Thesaurus as a gold standard to evaluate distribution-based clustering methods on the premise that there is a very low chance, 0.4%, of finding two words together under the same category. Therefore, he evaluates the results of these methods over Roget categories, in such a way that there is a hit when two words appear under the same category. Morris and Hirst (1991) used Roget's Thesaurus as a knowledge base to identify lexical chains, not only on the basis of two words sharing a common category, but of other relationships. E.g. two words with different categories that both point to another common one. The assumption that two words connected by a category can be clustered together is however not always reliable. Exploration of Roget's reveals how members of different semantic clusters may belong to the same category.

Word 1	Word 2	Score
measure	Meter	10
instrument	Measure	9
measure	Scale	8
instrument	Meter	7
instrument	Record	7
apparatus	Device	1
apparatus	Instrument	1
device	Instrument	1
graduate	Measure	1
machine	Tool	1
gauge	Weigh	1

Table 1: Top words sharing Roget's categories

Some words, such as "apparatus" and "device", are connected by one category, while other words, such as "instrument" and "tool", share up to four categories. Assuming that two words are more semantically related if they share a larger number of categories, top scores for both paradigms are presented in table 1. Half of the high scores relate words of two different paradigms, while low scores poorly relate members of the same intuitive paradigm.

2.1.3. Remarks

There are two possible uses of ontologies for clustering. The first one is to use them as a lexical knowledge base to extract information and build clusters. The second one is to use them as a "gold standard" to check the candidate clusters previously determined by any other method and other resources. In relation with the first possibility of use, it is convenient to know the amount of information each one provides. WordNet and Roget's Thesaurus describe a huge number of members for a paradigm, that is, they seem sufficient. Nevertheless, few words of a category may be interchangeable in the same context and then used as members of the same paradigm. This means that not all words in a category are necessary. Better possibilities of use exist for these lexical resources as a "gold standard" for clustering. WordNet and Roget's information seem quite sufficient to corroborate the similarity of a candidate pair of words, but only in the case such a pair refers to two words that already are similar. As observed above, members of two different paradigms may belong to the same category.

2.2. Distribution-based clustering

Conversely to the semantic relations extracted from an ontology, distribution-based clustering methods depend on pure statistical analysis of the lexical occurrences in running texts. The basis for the statistical approach is that similarity of words can be judged by analysing the similarity of the surrounding context in which they occur, since it has been observed that two synonym words share similar context when they occur separately.

There has been a great deal of research to find the similarity between words on the basis of the similarity of the context in which they occur. The methods derived from the distribution of words in corpora vary in both the type and size of the corpus, the span of the context for the analysis, and the measure of significance applied. Brown et al (1992) use mutual information in a window of 1,001 words, excluding the two words before and after the keyword, on large corpora. Mutual information has been used to analyse on-line dictionary definitions and textual corpora (Fukumoto and Tsujii, 1994; Fukumoto and Suzuki, 1996). The latter used a window size of 100 words. Edmonds (1997), besides mutual information, applies t-scores within a span of four words in a corpus of almost 3 million words. Despite their generally good results and the attempts to produce clusters with semantic coherence, it is appropriate to note some drawbacks that arise with distribution-based methods.

The use of statistical techniques to find similar words faces difficulties when it is fully automated, and new methods attempt insofar as possible to solve these difficulties. Earlier studies encountered drawbacks with the treatment of independent variant forms, such as spelling variation and inflectional endings of words (Adamson and Boreham, 1974). Although most corpus analysis software allows us to analyse variations of a word in the same utterance, it requires additional effort that reduces the efficiency of the method. The foremost reason is that distribution-based methods require us to process a large amount of data in order to get more reliable results (Habert et al, 1996; Arranz, 1997). However, the use of large corpora is not always practical, due to economic, time or capabilities factors. The consequences for lacking large corpora include results based on low-frequency words, which are quite unrepresentative for clustering.

Grefenstette (1996) suggests that a mixture of different methods, rather than any single statistical technique, may be adapted to be usefully applied to all ranges of frequency of words encountered in a corpus: for more finer grained context frequent words, discrimination; for less frequent words, using windows of N words; for rare words, examining large windows, even to entire document level. Regardless of the method used and of its reliability, there is always the task of checking the accuracy of final clusters, due to some strange results that occur for reasons that are not immediately apparent. For example, Charniak (1993) shows that many clusters present typical antonyms as similar adjectives. As he states "there are some possibly intrinsic limits to how finely we can group semantically simply by looking at surface phenomena".

2.3. Analogy-based clustering

As an alternative to the two traditional approaches described above, analogy-based methods have been proposed in computational linguistics for language processing. Federici and Pirrelli (1997) describe generalisation by analogy as the inferential process by which one can acquire knowledge of an unfamiliar linguistic object by drawing an analogy to more familiar objects, i.e., by extracting the right amount of linguistic knowledge from the examples of similar objects. Through analogy to a set of unselected examples, analogy-based approaches can generalise and find all the rules to apply correctly, instead of rule-based approaches, which apply a single rule to a given context.

Jones (1996) suggests corpus alignment as a feasible analogy-based approach. The automatic alignment of parallel texts aims to discover which words of the target sentence are most likely to correspond with which words of the source sentence. For example, given the following two definitions for alkalimeter:

CED	(1994):	An	apparatus	for	determining	the		
concentration of alkalis in solution								
OED2	(1994):	An	instrument	for	ascertaining	the		
amount of alkali in a solution								

alignment may identify which words in these definitions are equivalents of each other. Observation of the sentences lets us identify three pairs of words, namely: (apparatus, instrument), (determining, ascertaining) and (concentration, amount). Therefore, if the alignment of definitions of different dictionaries is used instead of pure statistical distribution-based methods or relation-based methods, it is possible to identify pairs of words used indistinguishably in similar contexts.

The appeal of using definitions as corpora for alignment is founded on two reasons. Firstly, dictionaries contain all necessary information as a knowledge base for extracting keywords (Boguraev and Pustejovsky, 1996). Secondly, it is much easier to find the sentences for aligning, since definitions are distinguished by the entry.

Through the alignment of definitions from two or more different sources, it is possible to retrieve pairs of words that can be used indistinguishably in the same sentence without changing the meaning of the concept. As lexicographic work relies on the same basis, such as genus and differentia, a concept is similarly defined by different dictionaries. The difference in words used between two lexicographic sources lets us extend the knowledge lexical base, so that clustering is possible through merging two or more dictionaries into a single database and then using an appropriate alignment technique. Since alignment starts from the same entry of a dictionary, clustering is faster than any other technique.

3. Our clustering algorithm

The analogy-based method here proposed, to identify automatically semantic clusters, aligns definitions from two different dictionary sources. Definitions are used as a lexical knowledge base because of the application of the clustering analysis to terminological information retrieval. The method relies on the assumption that two authors use different words to express a definition. The use of several dictionaries and terminologies is an advantage, because of the variety of senses and words that can be used to define or describe a concept. A dictionary designed for foreign learners of a language provides much information with a minimum of words, such as LDOCE (Procter, 1978) which uses a core controlled vocabulary of about 2000 words. A dictionary for children has less coverage, but it defines concepts easily and with examples (Barrière and Popowich, 1996). A dictionary for monolingual speakers can cover a variety of words. A terminology uses technical words to define a concept.

The alignment matches the words of two definitions and shows the correspondence between words that can replace each other in the definition without producing any major change of the meaning. The difference in words used between two or more lexicographic definitions enabled us to infer paradigms by merging the dictionary definitions into a single database and then using our own alignment technique.

3.1. Alignment of definitions

In order to align pairs of definitions for a same entry, drawn from two different dictionaries, our algorithm uses the Levenshtein distance (Levenshtein, 1966), a variation of a technique for measuring the similarity between lexical strings, named edit distance (Sellers, 1974; Waterman, 1996). Both match the words of two sentences in linear order and determine their correspondence. As they are defined as the smallest number of steps required to make two sentences identical, they identify the minimum cost for each operation required to change one definition into another. The operations are: substitution of a word for another, insertion of a word into a sentence and deletion of a word from a sentence.

Experimental results have shown that the application of the Levenshtein distance using stem forms gives better matches than using full forms. We will use the stemming algorithm of Porter (1980), which removes endings from words. This algorithm, widely used in IR, has been chosen because it performs slightly better than other similar algorithms. It should be noted that the Porter algorithm causes some over- and understemming. The risk of overstemming is low, since there is not too great a chance of having two different full forms in the same definition with a same stem form. Understemming is more probable and can cause some words not to match however the proposed clustering procedure will eventually match them.

A dynamic programming method (Wagner & Fisher, 1974) lets us align the elements of two strings from the arrangement given by the calculation of the Levenshtein distance and obtain the ordered pairs of the alignment. Every pair of words has associated with it the cost of the Levenshtein distance.

Word 1	Word 2	Cost
An	An	0
Instrument	instrument	0
For	for	0
Measuring	ascertaining	1
The	the	1
Amount	quality	2
Of	or	3
Nitrogen	value	4
In	of	5
A	niter	6
Substance		7

Table 2. Alignment for nitrometer

Table 2 gives the pairs of words aligned for the definition of nitrometer. The pairs of different words (measuring ascertaining) and (amount quality) are related semantically, considering the stem forms, but the other pairs of matched words are far from being related. Therefore, it is convenient to use a factor to measure the degree of similarity among the pairs of matched words.

3.2. Degree of similarity

The purpose of semantic clustering is to cluster those words that can be used indistinguishably in the same definition. For example, the pair (measuring ascertaining) in table 2 means that, if we replace "measuring" with "ascertaining" or vice versa, the definitions by which that pair was obtained do not alter their meaning. As a measure to judge the degree of similarity between the candidates for semantic clusters, we introduce the concept of *longest collocation couple* (lcc). This quantifies the surrounding pairs of equal words (stem forms) above and below a potential pair of similar words. Given the alignment of two strings consisting of ordered pairs of words, the lcc of a pair of matched words, whose stem forms are different, is the maximal length of the string of couples such that there is just one matched couple surrounded above and below by those couples whose stem forms are equal. Otherwise, the value of the *lcc* is equal to zero.

By this definition, only the pair of matched words (measuring ascertaining) of table 2 presents a valuable degree of similarity. Its *lcc* is equal to 5. The pair (amount quality) is not considered as the couple below it is not an equal couple of stem forms.

Experimental results on 314 terms for measuring instruments, extracted with their definitions from CED and OED2, show us that the greater the value of *lcc*, the greater the similarity between the pair of words. By inspection of table 3, we observe that a length of *lcc* equal to 5 is a reliable threshold.

ff_I	$f\!f_j$	lcc _{ij}					
Any	an	9					
Determining	measuring	9					
Celestial	heavenly	8					
Intensity	amount	8					
One	that	8					
Swinging	turning	8					
That	which	8					
Determining	which	7					
Inclination	direction	7					
Instrument	telescope	7					
That	for	7					
•••							
radiofrequency	radio	3					
resonator	combination	3					
spherical	closed	3					
such	at	3					
to	in	3					
to	on	3					
used	serving	3					
variation	form	3					
variation	motion	3					
variously	various	3					
wavelength	frequency	3					

Table 3. Some lcc triplets for our corpus

3.3. Stoplist discrimination

For the information retrieval domain, there are stopword lists containing commonly occurring words that are unlikely to be used for retrieval purposes [Salton 1968]. Removing stop words or non-relevant words from definitions must be done carefully in the case of onomasiological search, as a query can contain several important words that are usually removed for IR purposes. For example, "where", "when" and "who" are usually stop words, but for the user they can be equivalent to "place", "time" and "person", respectively. Function words can also be clustered. As an example, from table 3 we can observe that one of the highest *lcc* scores corresponds to "any-an", and other pairs are "any-a" and "an-the". However, in general, stop words interfere in the identification of clusters and can give more wrong than good results

Therefore, we use a stoplist to automatically identify any pair of words where a non-relevant word appears and exclude it, on the grounds that these are not very useful words for clustering. Thus, when the program comes across a matched pair of different words in a context and if that matched pair contains a word from the stoplist, then the pair is rejected. Essentially, this is the same thing as using a tagger and looking at the tags as well as the words, since one would not want to choose a noun pairing with a determiner or a relative.

3.4. Clustering

We introduce the term **binding** to represent a candidate cluster, i.e. a pair of words drawn from two definitions in such a way that the words are equivalent, in a determined context, after stoplist discrimination according to a determined threshold ($lcc \ge 5$).

As bindings represent pairs of words that are used with the same meaning in particular contexts, by replacing the full forms according to the bindings in the definition, we preserve the same meaning. As an example, we follow the alignment of "alkalimeter". By replacing the three bindings together (apparatus instrument), (determining ascertain) and (concentration amount), the final Levenshtein distance will be only 1, which indicates a strong similarity of both strings:

CED:	an	appa	ratus	in	strun	nent	for	ė	letermini	ng
ascerta	ining	g the	conce	ntra	ution	amou	int	of	alkalis	in
solution	n.									
OED2:	an	instru	ment	for	ascer	rtainin	g	the	amount	of
alkali i	n a s	olutior	1.				-			

In a consecutive sequence of bindings, it may happen that a stem form occurs in two or more different bindings. In this case, one can cluster all bindings with a common stem form according to the transitive property. The transitive property is more oriented, in the context of substitution of words in the strings, to preserve the meaning, rather than the synonymy of words. Even so, it is important to note that the transitive property applies in particular contexts, in the same way as a binding means that two words are used indistinguishably in the same definition. For example, the binding (instrument telescope) means that if we replace "instrument" with "telescope" or vice versa, the strings by which that cluster was obtained do not alter their meaning. However, replacing "instrument" with "telescope" in other strings can change the meaning of the string. This particular binding is a case of a hyponym relationship (Cruse, 1986), where a word A can be substituted with word B in all contexts without producing any difference of meaning, but word B cannot be substituted with word A in all contexts. As the clustering algorithm is based on aligning definitions and a strategy of lexicographers is to use the superordinate with the same meaning as the hyponym, it is common to find this kind of relationship between the words of a binding.

The algorithm to cluster bindings consists of three loops. We firstly assign a cluster number to each binding, so those bindings with a common word have the same cluster number. Then we cluster bindings with the same cluster number, but remove duplicate stem forms in the same cluster. Finally, we check if it is possible to merge new clusters with those of previous cycles

After substituting the bindings, it may happen that several pairs of words already present a high *lcc* score, even those pairs of words which initially did not yield matches with any word. It is then advantageous to replace the bindings in the definitions and repeat the entire process until no new clusters are found. The first cycle runs from the reading of definitions up to merging of clusters. All subsequent cycles will start by replacing the bindings in the definitions. The replacement of bindings in the definitions is a necessary step before starting a new cycle. It implies likely changes in the Levenshtein distance calculated for these modified definitions, so that words that were not aligned previously can now be aligned, and matched couples with a previously low *lcc* could increase it above the threshold value required for these couples to participate in clustering. The process iterates, replacing similar pairs of words in the definitions until no new clusters are found.

3.5. Modifying the strings

In order not to manipulate the strings to retrieve biased clusters, definitions were not modified to "tidy up" the data, before being submitted to the main process. No words in definitions were replaced or moved. In fact, entry words were chosen randomly, but always in the domain of "measuring instruments". Although good precision is observable in the final clusters, there are still some relevant words in the strings that are semantically similar to some of those of the clusters. For example, the word "device" is frequently used instead of "instrument", but because of the definition of *lcc*, the matched couple (device instrument) rarely can be a binding for clustering, as the preceding determiner of each word is different. The former use "a", while the latter use "an". As we observed, the matched couples (any an) and (any a) present a $lcc \ge 5$, so that by our clustering algorithm they should belong to the same cluster and then we can replace one with the other in the strings. By running the program without stoplist discrimination, we observe two clusters related to function words:

Cluster 1: a an any the	
Cluster 2: for that which	

Therefore, table 4 demonstrates clusters by first replacing all the strings according to these clusters of function words.

- 1. apparatus device instrument meter telescope
- 2. analyse ascertaining astronomical counting detecting determining estimating indicates location making measuring provides recording taking testing
- 3. amount concentration intensity percentage proportion rate salinity strength
- 4. hyperbolic radio radiofrequency
- 5. frequency wavelength
- 6. mass weight
- 7. conditions variations
- 8. swinging turning
- 9. direction inclination
- 10. accurate precise
- 11. distances heights
- 12. set specific
- 13. method system
- 14. field limits
- 15. observing tracing

16. day sunlight
17. celestial heavenly
18. photometric relative
19. displaying producing
20. angle slope
21. photographic visual
22. reticle time

Table 4. Final clusters for measuring instruments	Table 4.	Final	clusters	for	measuring	instruments
---	----------	-------	----------	-----	-----------	-------------

4. Evaluation

A noteworthy question arises of what does a "good" cluster mean. An objective answer is always preferred to a subjective one. However, there is no clear boundary between one and the other. By comparing clusters with other lexical resources, such as WordNet, Roget's thesaurus or a synonym dictionary, for example, then the question is to evaluate them and define which of them is the "best". In an onomasiological search, as proposed in this paper, the user and only the user has the answer. Our clustering program is only one approximation to the truth of one kind of user: the lexicographer.

In this perspective, we assess our semantic clustering algorithm and demonstrate the utility of our approach by identifying the recall and precision for a small set of texts.

4.1. Clusters for barometer

General language dictionaries present the advantage of using well-established lexicographic criteria to normalise definitions. These criteria, as for example the use of analytical definitions by genus and differentia, have been nowadays implemented by terminological or specialised dictionaries, with the addition of a richer vocabulary and the identification of properties that are not always considered relevant in other resources. Unfortunately, these are more oriented to a specific domain, so that it is sometimes necessary to search in two or more resources to compile the data. We used many online lexical resources, some of them available on the Internet. This allowed us to easily use different databases to extract semantic clusters. As an example, for the term "barometer" we selected from the Internet 17 sources from general language dictionaries, terminologies and specialised dictionaries, in addition to OED2 and CED.

The use of several terminologies lets us identify the clusters corresponding to a specific term or a more concentrated group of data. For example, it is interesting to analyse our 19 definitions of "barometer" and identify the words that are semantically similar. Table 5 demonstrates the use of our clustering program for the 19 definitions of "barometer", using $lcc \ge 5$, stoplist discrimination, and a previous modification of the original strings with the clusters of function words.

Cluster 1: air atmospheric	
Cluster 2: device instrument	
Cluster 3: determining measures shows	

Table 5. Clusters for barometer

From this table, we see there are only 3 clusters, but comparing these with the strings we observe that these clusters are complete with low recall and high precision. No more clusters can be extracted from the strings, there are no more relevant words in the strings that still can be clustered, and there are no unnecessary words in any of these clusters.

5. Acknowledgments

The research reported here was part of a Ph.D. project in the Department of Language Engineering, UMIST, 1996–1999. Support was provided by a scholarship from the Universidad Nacional Autónoma de México (UNAM), and the Secretaría de Educación Pública, Mexico.

6. References

Adamson, G.W. and Boreham, J. (1974). The use of an association measure based on character structure to identify semantically related pairs of words and document titles. *Information Storage and Retrieval*, 10, 253–260.

Agirre, E. & Rigau, G. (1996). Word sense disambiguation using conceptual density. In Proc. Coling-96, 16–22.

Anderberg, M.R. (1973). Cluster analysis for applications. New York: Academic Press.Arranz, M.V. (1998). Sublanguage-based semantic

clustering and disambiguation from corpora. PhD Thesis, UMIST.

Baldinger, K. (1980). Semantic theory: towards a modern semantics. Oxford: Basil Blackwell.

Barrière, C. and Popowich, F. (1996). Concept clustering and knowledge integration from a children's dictionary. Proc. COLING-96, 65–70.

Basili, R., Pazienza, M.T. & Velardi, P. (1996). Context driven conceptual clustering method for verb classification. In B. Boguraev & J. Pustejovsky (eds) (1996), 117–142.

Boguraev, B. and Pustejovsky, J. (1996). Issues in textbased lexicon acquisition. In B. Boguraev and J. Pustejovsky (eds) (1996).

Boguraev, B. and Pustejovsky, J. (eds) (1996). Corpus processing for lexical acquisition. Cambridge: The MIT Press.

Brown, P.F., de Souza, P.V., Mercer, R.L. Della Pietra, V.J., Lai, J.C. (1992). Class-based n-gram models of natural language. Computational Linguistics, 18(4), 467–479.

Calzolari, N. (1988). The dictionary and the thesaurus can be combined. In M.W. Evens (Ed.), Relational models of the lexicon: representing knowledge in semantic networks (pp. 75–96). Cambridge: Cambridge University Press.

CED. (1994). Collins English Dictionary. Glasgow: Harper Collins Publishers.

Charniak, E. (1993). Statistical language learning. Cambridge: The MIT Press.

Cruse, D.A. (1986). Lexical semantics. Cambridge: Cambridge University Press.

Edmonds, P. (1997). Choosing the word most typical in context using a lexical co-occurrence network. Proc. 35th Annual Meeting of the Association for Computational Linguistics, 507–509.

Federici, S. and Pirrelli, V. (1997). Analogy, computation and linguistic theory. In D.B. Jones & H.L. Somers (eds), New Methods in Language Processing. London: UCL Press, 16–34.

Fellbaum, C. (Ed.) (1998). WordNet: An Electronical lexical Database. The MIT Press.

Fox, E.A. (1988). Improved retrieval using a relational thesaurus for automatic expansion of extended Boolean logic queries. In M.W. Evens (Ed.), Relational models of the lexicon: representing knowledge in semantic networks (pp. 199–210). Cambridge: Cambridge University Press.

Fukumoto, F. and Suzuki, Y. (1996). An automatic clustering of articles using dictionary definitions. Proc. COLING-96, 406–411.

Fukumoto, F. and Tsujii, J. (1994). Automatic recognition of verbal polysemy. Proc. COLING-94, 762–768.

Grefenstette, G. (1996). Evaluation techniques for automatic semantic extraction: comparing syntactic and window based approaches. In B. Boguarev & J. Pustejovsky (eds) (1996), 205–227.

Habert, B. Naulleau, E. and Nazarenko, A. (1996). Symbolic word clustering for medium-size corpora. Proc. COLING-96, 490–495.

Hirakawa, H., Xu, Z. & Haase, K. (1996). Inherited Feature-based Similarity Measure Based on Large Semantic Hierarchy and Large Text Corpus. In Proc. Coling-96, 508–513.

Jones, D. (1996). Analogical natural language processing. London: UCL Press.

Kipfer, B.A. (1986). Investigating an onomasiological approach to dictionary material. Dictionaries: Journal of the Dictionary Society of North America, 8, 55–64.

Levenshtein, V.I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. Cybernetics and Control Theory, 10 (8), 707–710.

McRoy, S.W. (1992). Using multiple knowledge sources for word sense discrimination. Computational Linguistics, 18(1), 1–30.

Morris, J. and Hirst, G. (1991). Lexical cohesion computed by thesaural relations as an indicator of the structure of text. Computational Linguistics, 17(1), 21–48.

OED2. (1994). Oxford English Dictionary. Oxford: Oxford University Press and Rotterdam: Software B.V.

Porter, M.F. (1980). An algorithm for suffix stripping. Program, 14(3), 130–137.

Procter, P. (1978). Longman Dictionary of Contemporary English. Essex: Longman Group.

Resnik, P. (1995). Disambiguating noun groupings with respect to WordNet senses. Proc. of the 3rd Workshop on Very Large Corpora, 54–68.

Roget, P. (1987). Roget's Thesaurus of English words and phrases. Essex: Longman.

Sellers, P.H. (1974). An algorithm for the distance between two finite sequences. Journal of Combinatorial Theory (A), 16, 253–258.

Wagner, R.A. and Fisher, M.J. (1974). The string-tostring correction problem. Journal of the Association for Computing Machinery, 21(1), 168–173.

Waterman, S.A. (1996). Distinguished usage. In B. Boguraev and J. Pustejovsky (eds) (1996).