Analogy-based semantic clustering for terminological information retrieval

Gerardo Sierra
CCL, UMIST
PO Box 88
Manchester M60 1QD

gerardo@ccl.umist.ac.uk

Abstract

The majority of lexicographers recognise the need for the existence of onomasiological dictionaries that, contrary to the semasiological or alphabetical ordering of entries, help users to look for a word that has escaped their memory when remembering the concept. In the context of computational lexicography, it has been shown that machine readable dictionaries (MRDs), which are conventional semasiological dictionaries, can be used as an onomasiological dictionary 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 full-text database, the retrieved words may be far from the concept desired. As a result, it has been found advantageous to expand the original query with closely related keywords. 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. Finally, clusters can be provided in the search session in order to allow the user to select the best ones, or alternatively they can be automatically used by the system. The success of the dictionary relies on the accurate identification of the semantic clusters.

Semantic clustering methods are broadly divided into relation-based methods which use relations in an ontology, such as WordNet and Roget's thesaurus, and distribution-based methods which use statistical analysis. Although each of these methods looks promising for clustering, there are several reasons why they produce strange results. For example, any ontology lacks some semantic relations, and even the sense distinctions are not always satisfactory. It has been suggested therefore that relation-based analysis should combine several information sources. Distribution-based techniques require very large corpora, and each technique seems to apply to different ranges of frequency of words from a corpus.

An analogy-based clustering method is proposed, through the alignment of definitions from two different 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 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. In order to judge the degree of similarity between the candidates for semantic clusters, a variation of the edit distance is applied. This identifies the minimum cost for each operation required to convert one definition into another. Clusters are finally defined after extracting enough matches given a threshold.

1. Background

According to Baldinger (1970: 141) there are two methods for the study of meaning: conventional lexicography, which uses the semasiological approach, starts from the name and looks for the sense attached to it; the terminological method, which uses the onomasiological approach, starts from the sense and seeks to identify the

name connected with it. Many authors agree with Balanger on the difference between both types of dictionaries as well as on the purpose of the onomasiological approach, since during the text production process, it is frequently the case that a specific term to represent a given concept escapes the memory of the writer. As traditional linguistic prompts to find a forgotten term, writers have used concept-oriented dictionaries, such as thesauri, synonym dictionaries and pictorial dictionaries, which allow them to search from a concept, image or word to find the desired term. However, these dictionaries only allow writers to look for associated terms in general, not for specific terms, i.e., those which a writer may be said to have "on the tip of the tongue". In order to solve the need of writers who need to go from a meaning or a concept to a corresponding word, lexicographers have nowadays provided "reverse dictionaries" [Bernstein 1975], with specific characteristics that allow a user to search directly from a clue word rather than from an index or a conceptual tree. However, a limitation of these works is the subjectivity with which the lexicon is ordered, which means one of the first difficulties a user meets is the choice of the accurate word or phrase to begin the search.

Since the onomasiological search can be understood as the search for a name that designates a given concept, it is important to consider that the onomasiological approach is the opposite process from a search in a dictionary where the user, instead of introducing isolated words, is allowed to present any idea or concept coming to his mind, using natural language. Thus, any "writer" looking for a specific term must refer primarily to the concept, idea, thought or image expressed by words. Even more, due to the fact that today text production is mostly carried out on the computer, it is convenient to count on an easily accessible tool that allow the user to express the concept in natural language and to obtain as a result the appropriate term. A tool of such magnitude constitutes, undoubtedly, a contribution to the domain of terminology, since it facilitates the communication between different speakers who, though referring to the same concepts, use different terms.

In response to this need, at the Engineering Institute of the Universidad Nacional Autónoma de México (UNAM), a prototype version of the Onomasiological Search System (OSS I) was implemented. Its purpose is to help users find a word when they only have the concept. The prototype was elaborated for searching 33 terms in the domain of destructive phenomena which are taken from a Mexican conceptual framework on the topic of disasters [Sierra 1995]. Among its advantages over other dictionaries, it allows introducing the request through natural language, expressing the concept either by comprehension or extension. Therefore, the database foresees the possible syntagmas referring to a concept and, for each keyword of the syntagma, a paradigm of keywords. A syntagma is a chain of words that together represent a concept. The paradigm constitutes the set of keywords that present common features and that can be used with the same sense in the context of the final term to which they correspond, that is, any member of the paradigm can be substituted in the corresponding elements of the syntagma without changing the meaning. The OSS I works as an inverted file constituted by an indexed vocabulary of keywords, with each keyword having links to the items that carry the corresponding clue words given in the query. The identification of the 835 clue words and the determination of such paradigms was hand-elaborated, based mainly on definitions and the conceptual framework of disasters, through a semantic model that permits the analysis of the internal and external structure of each one of the concepts, supported by experts and the context given by the literature [Sierra 1997].

A parallel development of recent creation by Los Alamos National Laboratory, is the "Casey's Snow Day Reversed Dictionary", an onomasiological dictionary available at the Internet to seek terms in English [Faber 1996]. Similar to OSS I, it claims to solve the problem of the user who does not remember a word and who is looking for the word describing it. The user submits the query in a window using natural language, either a definition, a question or a set of words. The system matches the inputted text and the database definitions through a n-gram analysis [Frakes 1992]. The output is a list of 48 single terms, apparently sorted according to a similarity measure of occurrence of n-grams. Its principal advantage over OSS I is that it is able to obtain terms in any thematic area. However, its disadvantage is the lack of paradigms, in such a way that the change of a word in the query can alter totally the result.

In order to use leading technology in language engineering, information retrieval and computational linguistics, the UNAM has sponsored a doctoral program through a scholarship, at UMIST, with the goal to redesign the OSS I and to apply it more effectively and efficiently to any terminology.

2. Outline of the search system

The success of an onomasiological search relies upon the accuracy of all clue words in the concept 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 full-text database, the retrieved words may be far from the concept desired. As a result, it has been found advantageous to expand the original query with closely related keywords. For inverted file systems, the use of paradigms allows compression of the database file and expansion of the initial query keywords, since they are mapped to the file of index paradigms, and the system just retrieve the items corresponding to the paradigm [Porter 1980]. Conversely, for full-text searching, the main goal of using clusters is to expand the search. The query clue word is substituted by all the members of the paradigm and every one is used to search in the full-text database [Calzolari, Picchi & Zampolli 1987].

As a capability of some systems, paradigms can be provided in the search session through an on-line relational thesaurus, which brings related words together, such as alternative forms, synonyms or cross-references, and thereby helps to stimulate the user's memory [Swanson 1960]. It allows the user to select the best words and express the same concept in alternative ways.

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 expand the query with closely related keywords which enhance the meaning, particularly when the query is expressed in natural language. In order to help the user focus on the search, the systematisation of this task has been hence placed on the system, in such a way that the system produces and manages the semantic paradigm transparently, without any intervention by the user [Calzolari, Picchi & Zampolli 1987]. In fact, this should be a goal of a user-friendly onomasiological search system.

In the context of computational lexicography, it has been shown that machine readable dictionaries (MRDs), which are conventional semasiological dictionaries, can be used as an onomasiological dictionary for practical terminological data searching, seeking a word whose definition contains the clue word [Calzolari 1988; Wilks et al 1996]. An example of onomasiological search in a dictionary where the user enters a description of the concept and the system looks up every word in the full-text database of definitions is Casey's Snow Day Reverse Dictionary. As stated above, it still needs to expand the search with a database of paradigms in order to increase its efficiency. Therefore, the success of natural language retrieval relies on the automatic manipulation of the keywords, and on the online expansion of queries, together with ranking and matching processes. The central point of this paper is oriented to the accurate identification of the semantic clusters.

3. Clustering

The primary goal of clustering is to collect together into paradigms or clusters a set of elements associated by some common characteristic. Each member within a cluster is strongly associated with each other because they share the same property, while members of other clusters show distinct characteristics from one another. According to Gordon [1981], clustering may alternatively be oriented either to discover the strongest association among elements or to seek clusters which are isolated from each other. Clustering is based on measurements of the similarity between a pair of objects, these objects being either single elements or other clusters.

The purpose of clustering varies from classification and sorting to the development of inductive generalisations [Anderberg 1973]. A cluster is defined by its elements and by the "central concept" to which all 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 to be used to characterise the elements of the problem, either the characteristics, attributes, class memberships and other such properties. Here the focus is on semantic variables.

Clustering methods to identify semantically similar words are usually divided in relation-based and distribution-based approaches [Hirawaka, Xu and 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 also a little known approach called analogy-based. This

analysis. From a methodological point of view, there is also a little known approach called analogy-based that employs an inferential process and is used in computational linguistics and artificial intelligence as an alternative to current rule-based linguistic models.

3.1 Relation-based clustering

Relation-based clustering methods rely on the relations in an 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. Semantic relations may be extracted from a taxonomy, such as ISA or a-kind-of [Chakravarthy 1994], to judge the similarity between two concepts by comparing their parent. Similarity measures derived may be from a semantic network by determining the path-length or number of links between the nodes. The most important lexical knowledge resources that supply a basic ontology for clustering are machine readable dictionaries (MRDs), the semantic taxonomy WordNet [Miller et al 1990] and Roget's thesaurus.

An ontology can be used in two ways for clustering. The first one is to use it as a lexical knowledge base to extract information and get the clusters. The second one is to use it as a "gold standard" to check the candidate clusters previously determined by some 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. For our purpose, dictionaries provide the necessary information, i.e., all words in a definition, except functional words, can be considered as keywords of the concept. Because definitions are considered as an authority in the subject, they may constitute the base for the paradigm construction. However, definitions do not provide sufficient information, and this means they are not a reliable ontology for clustering. In the other extreme, WordNet and Roget's thesaurus describe a huge number of members for a paradigm, that is, they seems 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 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 just in 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. On the other hand, inconsistency in lexicographic definitions means dictionaries become less reliable to verify candidate clusters, besides their structure as an ontology is not immediately obvious.

3.2 Distribution-based clustering

In contrast 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 synonymous words share a similar context when they occur separately.

The methods derived from the distribution of words in text corpora or dictionaries start extracting the co-occurrences of words in a window of set size. Afterwards, the methods must to assign a numerical value to each property one cares about. As an example, mutual information is a statistical significance measure widely used to calculate the words most strongly collocated. Finally, the algorithms compare the vectors associated to each word with the numerical values and cluster the words with the most similar vectors.

The use of statistical techniques to find similar words faces difficulties when 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 the analysis of 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 to process 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, either due to economic, time or capabilities factors. The consequences for lacking large corpora result in low-frequency words, which are quite unrepresentative for clustering. Grefenstette [1996: 216] 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 frequent words, finer grained context 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 to check the accuracy of final clusters, due to some strange results occurring for not immediately apparent reasons. 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".

3.3 Analogy-based clustering

As an alternative to the two traditional approaches described above, analogy-based methods has 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 the analogy to a set of unselected examples, analogy-based approaches can generalise and find all the rules to apply correctly, in contrast to rule-based approaches, which apply a single rule to a given context [Jones 1996].

4. Alignment

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. Alignment is concerned with bilingual parallel texts, in order to provide an aid to translators, bilingual lexicographers and machine translation systems, since it is possible to find how a sentence in a source language is translated in the other language [Oakes 1998; Wu 1998]. Sentence alignment methods are nearly all statistical in nature, and are based on the facts described by Gale and Church [1991], such as the length of a sentence in one language tends to be similar in the other language.

Although the methods have not been applied to the alignment of two sentences in the same language, the purpose of corpus alignment concerns both monolingual and bilingual texts. For example, given the following two definitions for alkalimeter, from two on-line dictionaries, Collins English Dictionary (CED) and Oxford English Dictionary (OED2):

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

Alignment may identify which words in those definitions are equivalent. A quick observation of the sentences lets us identify three likely pairs of words: (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 indiscriminately in similar contexts. The appeal of using definitions as corpora for alignment is twofold. 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.

5. Edit distance

In order to align two sentences in the same language, Waterman [1996] proposes one technique for measuring the similarity between lexical strings, named edit distance. This matches the words of two sentences in linear order and determines their correspondence. Sellers [1972] defines the edit distance as the smallest number of steps required to make two sequences identical. Wagner and Fisher [1974] associate each step for changing one string into another as the cost of three edit operations, so that the minimum cost is desired. The operations are: substitution of a word to another, insertion of a word into the sentence and deletion of a word from the sentence.

Given two sentences $A = (a_0, a_1,..., a_m)$ and $B = (b_0, b_1,..., b_n)$ for aligning, where a_i corresponds to the i word in

the sentence A, Waterman defines the cost $D(a_i,b_j)$ of changing a_i to b_j, for all $i \geq 1$, as:

$$D(a_i, b_j) = \min \begin{cases} D(a_i, b_{j-1}) + D_{ins}(b_j) \\ D(a_{i-1}, b_j) + D_{ins}(a_i) \\ D(a_{i-1}, b_{j-1}) + D_{sub}(a_i, b_j) \end{cases}$$

Where the cost of insertion, $D_{ins}()$, is one, and the cost of substitution, $D_{sub}()$, is 0 or 2, according as a_i and b_j differ or not. By definition, the starting cost $D(a_0,b_0)$, corresponding to the null word, is equal to zero.

Through the alignment of a definition from two or more different sources it is possible to retrieve likely pairs of words that can be used indiscriminately in the same sentence without changing the meaning of the concept. As the lexicographic work relies on the same basis, such as genus and differentia, a concept is similarly defined by different dictionaries. The difference of words used between two lexicographic sources let extend the knowledge lexical base, so that clustering is available 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.

However, some drawbacks arise using different dictionaries. Dolan [1994] considers the arbitrary sense divisions for the same entry among dictionaries, in such a way that a sense in a dictionary can correspond to two or more senses in another one. The alignment technique itself can determine which senses of two dictionaries are closely related to one another, though it increases the performance cost. After solving the sense divisions, the type of definitions is another factor that can affect the performance of alignment. Some dictionaries tend to give concise definitions while some others provide broad explanation of a term. By splitting definitions in smaller sub-strings, it is possible to identify more closely sentences for aligning.

As an analogy-based method, clustering can be seen as a sequential process from the most elemental to the most complex. Starting with the minimal edit distance, one can get the first pair of words, such that it will be corroborated with other alignments. After a certain number of clusters, sentences with higher distance are realigned considering clusters instead of single words.

6. Conclusions

Jones [1996] suggests corpus alignment as a feasible analogy-based approach. The use of alignment of definitions has been outlined as a promising application to semantic clustering, as well as an algorithm to match similar pair of words from aligned sentences. Although some drawbacks arise, some ideas to solve them are presented, which enable this method to be considered for implementation in a potential application.

Acknowledgement

To John McNaught for his invaluable supervision and suggestions to improve this paper.

References

1. ***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*

identify semantically related pairs of words and document titles". Information Storage and Retrieval 10, 253-260.

2. ***Anderberg, M.R.*** *1973. Cluster Analysis for Applications. New York: Academic Press.*
3. ***Arranz, M.V.*** *1997. "Lexical Bottleneck in Machine Translation and Natural Language Processing: A Case Study". Aslib 97. London.*
4. ***Bernstein, T.M.*** *1975. Bernstein's Reverse Dictionary. London: Routledge & Kegan Paul.*
5. ***Boguraev, B., and Pustejovsky, J.*** *1996. "Issues in Text-based Lexicon Acquisition". Corpus Processing for Lexical Acquisition. B. Boguraev and J. Pustejovsky (eds.). Cambridge: The MIT Press.*
6. ***Calzolari, N.*** *1988. "The dictionary and the thesaurus can be combined". In Relational Models of the Lexicon: Representing knowledge in semantic networks. M.W. Evens (ed.). Cambridge: Cambridge University Press.*
7. ***Calzolari, N., Picchi, E., and Zampolli, A.*** *1987. "The use of computers in lexicography and lexicology". In Lexicographica: the dictionary and the language learner. A. Cowie (ed.).*
8. ***Chakravarthy, A.S.*** *1994. "Representing Information Need with Semantic Relations". Proc. COLING-94. The 15th International Conference on Computational Linguistics, Kyoto, 737-741.*
9. ***Charniak, E.*** *1993. Statistical Language Learning. Cambridge: The MIT Press.*
10. ***.Church, K., et al.*** *1991. "Using Statistics in Lexical Analysis". Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon. U. Zernik (Ed.). New Jersey: Lawrence Erlbaum Associates.*
11. ***.Dolan, W.B.*** *1994. "Word Sense Ambiguation: Clustering Related Senses". In Proc. COLING-94. The 15th International Conference on Computational Linguistics, Kyoto, 712-716.*
12. ***.Faber, V.*** *1996. "Casey's Snow Day Reverse Dictionary (and Guru)". Los Alamos National Laboratory. Web page http://www.c3.lanl.gov:8064/.*
13. ***.Federici, S., and Pirrelli, V.*** *1997. "Analogy, computation and linguistic theory". New Methods in Language Processing. D.B. Jones & H.L. Somers (eds.). London: UCL Press, 16-34.*
14. ***.Frakes, W.B.*** *1992. "Stemming algorithms". In Information Retrieval: Data Structures & Algorithms. W.B. Frakes and R. Baeza-Yates (eds.). New Jersey: Prentice Hall.*
15. ***.Gordon, A.D.*** *1981. Classification. Cambridge: University Press.*
16. ***.Grefenstette, G.*** *1996. "Evaluation Techniques for Automatic Semantic Extraction: Comparing Syntactic and Window Based Approaches". Corpus Processing for Lexical Acquisition. B. Boguarev and J. Pustejovsky (eds.). Cambridge: The MIT Press.*
17. ***.Habert, B., Naulleau, E., and Nazarenko, A.*** *1996. "Symbolic word clustering for medium-size corpora". Proc. COLING-96. The 16th International Conference on Computational Linguistics, Copenhagen, 490-495.*
18. ***.Hirakawa, H., Xu, Z., and Haase, K.*** *1996. "Inherited Feature-based Similarity Measure Based on Large Semantic Hierarchy and Large Text Corpus". Proc. COLING-96. The 16th International Conference on Computational Linguistics. Copenhagen: Center for Sprogteknologi.*
19. ***.Jones, D.*** *1996. Analogical Natural Language Processing. London: UCL Press.*
20. ***.McRoy, S.W.*** *1992. "Using Multiple Knowledge Sources for Word Sense Discrimination". Computational Linguistics 18(1), 1-30.*
21. ***.Miller, G.A., Fellbaum, C., Kegl, J., and Miller, K.*** *1990. "The Priceton Lexicon Project: A Report on WordNet". In BudaLEX'88 Proceedings. T. Magay and J. Zigány (eds.). 543-558.*
22. ***.Oakes, M.P.*** *1998. Statistics for Corpus Linguistics. Cambridge: Edinburgh University Press.*
23. ***.Porter, M.F.*** *1980. "An algorithm for suffix stripping". Program 14(3), 130-137.*
24. ***.Sellers, P.H.*** *1992. "An Algorithm for the Distance Between Two Finite Sequences". Journal of Combinatorial Theory (A) 16, 253-258.*
25. ***.Sierra, G.*** *1995. "Outline of an Onomasiological Dictionary Software in the Disaster Area". Journal of the International Institute for Terminology Research IITF 6/2.*
26. ***.Sierra, G.*** *1997. "Estructura semántica del léxico en un diccionario onomasiológico práctico". Estudios de Lingüística Aplicada 23/24.*
27. ***.Swanson, D.R.*** *1960. "Searching natural language text by computer". Science 132, 1099-1104*
28. ***.Wagner, R.A., and Fisher, M.J.*** *1974. "The String-to-String Correction Problem". Journal of the Association for Computing Machinery 21(1), 168-173.*
29. ***.Waterman, S.A.*** *1996. "Distinguished Usage". In Corpus Processing for Lexical Acquisition. B. Boguraev and J. Pustejovsky (eds.) Cambridge: The MIT Press.*
30. ***.Wilks, Y.A., Slator, B.M., Guthrie, L.M.*** *1996. Electric Words. Cambridge: ACL-MIT Press.*
31. ***.Wu, D.*** *1998. "Alignment". A Handbook of Natural Language Processing. R. Dale, H. Moisl and H. Somers (eds.). New York: Marcel Dekker.*