Genus Disambiguation: A Study in Weighted Preference*

Rebecca Bruce and Louise Guthrie

Computing Research Laboratory
Box 30001
New Mexico State University
Las Cruces, NM 88003-0001

ABSTRACT

The automatic construction of an IS_A taxonomy of noun senses from a machine readable dictionary (MRD) has long been sought, but achieved with only limited success. The task requires the solution to two problems: 1) To define an algorithm to automatically identify the genus or hypernym of a noun definition, and 2) to define an algorithm for lexical disambiguation of the genus term. In the last few years, effective methods for solving the first problem have been developed, but the problem of creating an algorithm for lexical disambiguation of the genus terms is one that has proven to be very difficult. In COLING 90 we described our initial work on the automatic creation of a taxonomy of noun senses from Longman's Dictionary of Contemporary English (LDOCE). The algorithm for lexical disambiguation of the genus term was accurate about 80% of the time and made use of the semantic categories, the subject area markings and the frequency of use information in LDOCE. In this paper we report a series of experiments which weight the three factors in various ways, and describe our improvements to the algorithm (to about 90% accuracy).

1. Introduction

Much of the previous research on the construction of networks of genus terms from MRD's (Amsler and White 1979; Chodorow et al. 1985; Nakamura and Nagao 1988; Vossen 1990) required human intervention to distinguish the senses. Recently, several researchers (Veronis and Ide 1990; Klavans et. al 1990; Copestake 1990; Vossen 1991) have suggested techniques for automatic disambiguation of these taxonomies based on neural net techniques, word overlap, or bilingual dictionaries. The techniques we have used to construct a network of noun senses automatically from the Longman Dictionary of Contemporary English (LDOCE) differ substantially from any of those methods.

In (Guthrie et al. 1990), we suggested and algorithm for disambiguating the genus terms of noun definitions in LDOCE. The procedure we used was based on the assumption that the semantic relationship between the headword and its genus should be reflected in their LDOCE semantic categories. In other words, the semantic category of the genus word should be identical to, or an ancestor of, the semantic category of the headword (an ancestor is a superordinate term in the hierarchy of semantic codes). Using a random sample of 520 noun word sense from LDOCE, we tested this assumption.

The semantic categories used (there are thirty-four in all) were defined by the LDOCE lexicographers, who placed sixteen of the basic categories in a hierarchy. The notion of a "more general semantic category" was somewhat subjective, as is illustrated in the next section.

The disambiguation algorithm presented in (Guthrie et al. 1990) utilized three factors in determining the correct genus sense. The algorithm is stated as follows:

- Choose the genus sense with the same semantic category as the headword (or closest more general category if this is not possible).
- In the case of a tie, choose a sense with has the same pragmatic code
- In case there is still a tie, or no genus sense meeting the above criteria, choose the most frequently used sense‡ of the genus word.

* This research was supported by NSF Grant No. IRI-8811108.

‡ In the 2nd edition of LDOCE, the publishers state that the order in which word senses are listed corresponds to the frequency with which each sense is used (ie. the first sense listed is the most commonly used, etc.). We have observed

The algorithm was successful about 80% of the time.

In an effort to improve the disambiguation algorithm, we conducted a series of experiments designed to identify more completely the contribution of each factor consider in the algorithm. Since we considered three factors in determining the correct genus sense (the semantic code relationship, the pragmatic code relationship, and the frequency information), we designed experiments to first test each factor separately, and then again in combination, weighting each input according to its individual predictive value. Below we describe those experiments, beginning with the formulation of each factor, and ending with the assignment of weights to the contribution of each input in the final disambiguation algorithm.

2. Sense Selection Based on LDOCE Semantic Codes

This section describes our investigation of the use of semantic category information for disambiguation, and outlines the problems in using that type of information. The basic hierarchical structure of the semantic codes provided by LDOCE is depicted in Figure 1. In addition to the codes positioned in that tree structure, seventeen other codes, which we refer to as "composite" are defined as follows:

E = solid or liquid
U = collective and animal or human
O = animal or human (sex unspecified)
K = male (animal or human)
R = female (animal or human)
V = plant or animal (not human)
W = abstract and inanimate
Y = abstract or animate
X = not concrete or animal (abstract or human)
Z = unmarked (no semantic restriction)
1 = human and solid
2 = abstract and solid
3 = "it" as subject or object
4 = physical quantities
5 = organic materials
6 = liquid and abstract
7 = gas and liquid

To evaluate our assumption that the semantic category of the genus word is the same or more general than the semantic category of the headword, it was necessary to define what we meant by "more general" for the composite categories. We did this by incorporating the composite codes into the hierarchical structure display in Figure 1, and defining a semantic distance between word senses based on the placement of their respective codes in the hierarchy. It was obvious from the start that the addition of these codes to the tree depicted in Figure 1 would create a tangled hierarchy. The problem was to decide where these codes should be placed in the tree structure in order to preserve inheritance. For example, should "E" (the code for "solid or liquid") be placed above or below "solid" and "liquid", and would a similar placement hold for code 7, which reads "gas AND liquid" (as opposed to "liquid OR solid")?

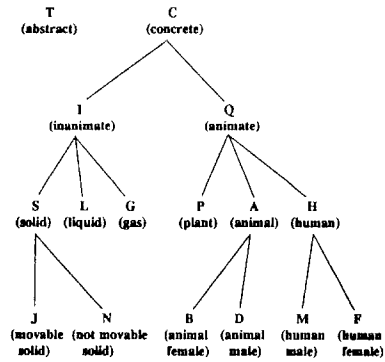

Figure 1:
Basic Hierarchy of LDOCE Semantic Codes

To answer such questions, two types of studies were conducted. The first was an in-depth look at the words marked with composite codes (nouns marked to identify a semantic category and adjectives and verbs marked as to their selection restrictions). The second was a survey of the genus senses for headwords with composite semantic codes. As might be expected, there were inconsistencies in the assignment of nouns categories. For example, within the "liquid" categories, we observed that nouns which represent both liquids and solids can be found in both categories L and E, and abstractions of liquids can be found in categories L, 6, and 7. This is not surprising, as it is difficult to create distinct categories for overlapping concepts.

that the listing order of senses in the 1st edition of LDOCE is similar to that of the 2nd, and have found empirical evidence in the work of Guo (1989) and this study to show that a similar connection between the order in which word senses are listed and the frequency with which they are used (in LDOCE) holds for the 1st edition as well.

Our proposed placement of composite codes within the hierarchy structure provided by LDOCE is presented in Figure 2. In constructing Figure 2, we attempted to create a hierarchy which would reflect not only the data gathered on the properties of words assigned to each category, but also the most frequently occurring superset for each composite code, based on the results of the second study.

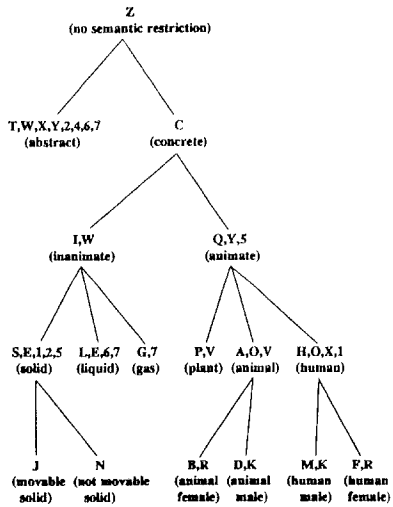

Figure 2:
Revised Hierarchy of LDOCE Semantic Codes

Based on this study of the semantic codes used in LDOCE, three implementations of a partial genus sense selection algorithm (partial because at this time we are only considering the contribution made by the semantic code comparison to sense selection) were found to be possible. They are as follows:

1. Selection of the genus sense with a minimum semantic distance from the headword sense, where semantic distance is measured by the placement of the respective codes in the hierarchy presented in Figure 2. (This formulation of a genus sense selection criteria is the basis of the algorithm reported in Guthrie et al. 1990.)
2. Choose the genus sense with a semantic code belonging to the same code set as the code of the headword, where the code sets are the nodes of the tree structure presented in Figure 2.
3. Select the genus sense with a semantic code identical to the headword.

3. Sense Selection Based on LDOCE Pragmatic Codes

The pragmatic codes in LDOCE are another set of terms organized into a hierarchy, although the hierarchy provided by LDOCE is quite flat. As stated earlier, these terms are used to classify words by subject area. The LDOCE pragmatic coding system divides all possible subjects into 124 major categories, ranging from *aeronautics, aerospace,* and *agriculture,* to *winter-sports,* and *zoology.* The hierarchy is only two layers deep, and the 124 major categories have equal and unrelated status.

Slator (1988) implemented a scheme which imposed deeper structure onto the LDOCE pragmatic code hierarchy. He restructured the LDOCE pragmatic code hierarchy by making *Communication, Economics, Entertainment, Household, Politics, Science,* and *Transportation* fundamental categories, and grouping all other pragmatic codes under those headings. His restructuring of the code hierarchy revealed that words classified under *Botany* have pragmatic connections to words classified as *Plant-Names,* as well as connections with other words classified under *Science.*

We investigated four implementations of a genus sense selection algorithm based on pragmatic codes. The first implementation utilized the hierarchy developed by Slator. In that scheme, the pragmatic codes were arranged in a tree structure in which each node of the tree is a single pragmatic code.

In addition, pragmatic code sets were defined directly from Slator's hierarchy by creating seven large groups corresponding to the seven subtrees of the top level of the hierarchy. Each of the seven code sets contained all codes descendant from the corresponding top level node. Within this construction, lack of common set membership is a strong indication of disjoint subject areas.

In summary, we proposed four approaches to genus sense selection based on pragmatic codes:

1. Choose the genus sense with minimum pragmatic distance from the headword sense, where pragmatic distance is measured by the placement of the respective codes in the hierarchy implemented by Slator.
2. Select the genus sense with a pragmatic code belonging to the same code set as the code of the headword. Seven code sets were constructed corresponding to the seven major divisions of Slator's hierarchy.
3. Rule out all headword/genus sense combinations with pragmatic codes that are not in the same code set.

4. Select the genus sense with a pragmatic code identical to the headword.

4. Results of the Experimentation

All tests of the proposed sense selection criteria were run on the same random sample of 520 definitions. Table 1 provides a summary of the relevant test results. Although each selection mechanism was evaluated separately, because of the large number of word senses having either redundant code markings, or no markings at all (particularly with pragmatic codes), it was necessary to introduce a default or "tie breaking" mechanism for all selection criteria other than usage frequency. Usage frequency was established as the default selection mechanism for all tests. When no sense selection (or no unique sense selection) could be made based on the criteria being tested, the sense selection was based on usage frequency (ie., of the competing senses, the sense occurring first in the listing order was selected).

The variation in performance between all approaches developed for genus sense selection was relatively small - no more than 8%. Both the best and the worst performance of a single sense selection parameter was achieved using pragmatic code relationships. The best performance (80% success rate) resulted from requiring identical code markings for headword and genus senses. The worst disambiguation performance was the result of sense selection based on common pragmatic code set membership. The variation in disambiguation performance was small in the experiments which used only the semantic code information. The maximum success rate of 77% resulted from stipulating common code set membership, while the minimum success rate was 75% for identical code designation.

Some of the test results were unexpected: for instance, we did not expect selection of the first sense listed to yield a 76% success rate. Nor did we expect sense selection based on a subset/superset relationship between codes to be as unsuccessful as it was, yielding no more than a 78% success rate for both pragmatic and semantic codes.

Although the experiments showed that a direct match of pragmatic codes was the most successful single selection mechanism, the result is somewhat misleading. Because many words have no pragmatic code, the default rule was applied often, resulting in the selection of the most frequently used sense. Having said that, it remains true that the tests show pragmatic code information to be the best predictor of the correct genus sense, when it is present.

SUMMARY OF DISAMBIGUATION EXPERIMENTS	
GENUS SENSE SELECTION MECHANISM	TEST RESULTS
Selection based on semantic codes:	
subset/superset relationship implemented with code hierarchy	75% correct
common code set membership	77% correct
identical code designation	75% correct
Selection based on pragmatic codes:	
subset/superset relationship implemented with code hierarchy	78% correct
common code set membership (preferred)	72% correct
common code set membership (exclusive)	72% correct
identical code designation	80% correct
Selection based on usage frequency:	
select first sense listed	76% correct
Weighted, 3 Parameter Selection Algorithm	
common semantic code set - weight 1 identical pragmatic code - weight 1 usage frequency - tie breaker	80% correct
common semantic code set - weight 1 identical pragmatic code - weight 2 usage frequency - tie breaker	80% correct
semantic code hierarchy - weight 1 | 2 pragmatic code hierarchy - weight 2 usage frequency - tie breaker	79% correct
common semantic code set - weight 1 identical pragmatic code - weight 2 usage frequency - tie breaker hand-coded exceptions included	90% correct

Table 1: Summary of Disambiguation Experiments

Table 1 also displays the results of tests performed using all three factors in combination. These experiments were conducted to determine the optimum weight to assign each of the three factors when considering their cumulative predictive capability. The selection of weights was based on the performance of each factor individually. Again, the variation in performance across all tests of different weighings was small (less than 1%). The highest success rate was achieved when pragmatic code information received the greatest weight.

As a result of these tests, our disambiguation algorithm was formulated as follows:

- Choose the most frequently used genus sense unless an alternate sense choice is indicated by a strong relationship between headword and genus codes, either semantic or pragmatic.
- If the sense selection based on semantic codes differs from that inferred by the pragmatic

codes, base the sense selection on the pragmatic codes.

- Select among competing genus senses with identical code markings by choosing the most frequently used sense.

By a "strong relationship" in the case of semantic codes, we mean membership in the same code set. This is not surprising due to the limited scope of the code sets, and the inherent overlap of the composite codes. Strong relationship for pragmatic codes means an exact match.

5. The Final Disambiguation Algorithm

Review of the output data from each disambiguation trial using the three parameter algorithm revealed that the majority of the failures were on a very small number of frequently occurring genus words. Often, the pragmatic and semantic classifications of these word senses were either deficient (lacking in code information), or redundant (more than one word sense having the same markings). Such situations frequently arise with very abstract words (e.g. **part**, **quality**, **piece**, and **number**) where there are numerous word senses, and most (if not all) senses have identical semantic codes and no pragmatic codes.

The final modification to our genus sense selection algorithm was introduced to solve this problem: the correct sense selections for words with errors in their code information, as well as certain very general words are pre-selected, and assumed to be constant. Fewer than ten words required hand coding of the correct sense and almost all were abstract words such as **part** or **quality**. While it is true that the majority of these words are "disturbed heads" (Guthrie et al. 1990), and will, in the future, not serve as genus terms but rather as identifiers of alternate link types, we still require that they be sense disambiguated to serve as relation descriptors. This final modification to the sense selection algorithm increased performance by 10%, resulting in success rate of 90%.

6. References

Amsler, Robert A., and John S. White (1979). Development of a Computational Methodology for Deriving Natural Language Semantic Structures via Analysis of Machine-readable Dictionaries. *Technical Report MCS77-01315*, NSF.

Copestake A. (1990). An approach to building the hierarchical element of a lexical knowledge base from a machine readable dictionary, *Proceedings of the First International Workshop on Inheritance inNatural Language Processing*, Tilburg, The Netherlands, pp. 19-29.

Chodorow, Martin S., Roy J. Byrd, and George E. Heidorn (1985). Extracting Semantic Hierarchies from a Large On-Line Dictionary. *Proceedings of the 23rd Annual Meeting of the ACL*, Chicago, IL, USA, pp.299-304.

Guo, Cheng-Ming (1989). Constructing a Machine Tractable Dictionary From Longman Dictionary of Contemporary English, *Memoranda in Computer and Cognitive Science*, MCCS-89-156. Computing Research Laboratory, New Mexico State University.

Guthrie, Louise, Brian Slator, Yorick Wilks, and Rebecca Bruce (1990). Is there content in Empty Heads? *Proceedings of the 13th International Conference on Computational Linguistics (COLING-90)*, Helsinki, Finland, 3, pp.138-143.

Klavans, J., Chodorow, M., Wacholder, N. (1990). From Dictionary to Knowledge Base Via Taxonomy. *Proc. of the 6th Conference UW Center for the New OED*, Waterloo, pp. 110-132.

Nakamura, Jun-ichi, and Makoto Nagao (1988). Extraction of Semantic Information from an Ordinary English Dictionary and its Evaluation. *Proceedings of the 12th International Conference on Computational Linguistics (COLING-88)*, Budapest, Hungary, pp.459-464.

Slator, Brian M. (1988). Constructing Contextually Organized Lexical Semantic Knowledge-bases. *Proceedings of the Third Annual Rocky Mountain Conference on Artificial Intelligence (RMCAI-88)*, Denver, CO, pp.142-148.

Ide, N.N. and J. Veronis (1990). Very Large Neural Networks for Word Sense Disambiguation. *European Conference on Artificial Intelligence, ECAI '90*, Stockholm.

Vossen, P. (1991). Polysemy and Vagueness of Meaning Descriptions in the Longman Dictionary of Contemporary English. In J. Svartvik and H. Wekker (eds.), *Topics in English Linguistics*. Mouton de Gruyter.