Is there content in empty heads?*

*Louise Guthrie, Brian M. Slator,** Yorick Wilks, and Rebecca Bruce*

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

ABSTRACT

We describe a technique for automatically constructing a taxonomy of word senses from a machine readable dictionary. Previous taxonomies developed from dictionaries have two properties in common. First, they are based on a somewhat loosely defined notion of the IS-A relation. Second, they require human intervention to identify the sense of the genus term being used. We believe that for taxonomies of this type to serve a useful role in subsequent natural language processing tasks, the taxonomy must be based on a consistent use of the IS-A relation which allows inheritance and transitivity. We show that hierarchies of this type can be automatically constructed, by using the semantic category codes and the subject codes of the Longman Dictionary of Contemporary English (LDOCE) to disambiguate the genus terms in noun definitions. In addition, we discuss how certain genus terms give rise to other semantic relations between definitions.

1. Introduction

In order to extract meaning from text, we must at least disambiguate the words used in the text. The normal scenario is to locate the disambiguated words in some knowledge base, which gives additional information about the words and their properties, so that some of the higher-minded tasks of natural language processing (NLP) can be accomplished: tasks like determining speech acts, identifying topic and focus shifts for discourse analysis, and drawing inferences. The work described here is a step toward developing an initial knowledge base for NLP by automatically transforming information found in machine readable dictionaries into a data base suitable for a variety of NLP applications.

The overall scheme of the work at CRL on machine readable dictionaries is described in Wilks et al. (1988, 1989, 1990). As part of that work, Slator (1988a, 1988b; Slator and Wilks, 1987, 1990) developed a program called Lexicon Provider which creates frames from the dictionary definitions of word senses provided by the Longman's Dictionary of Contemporary English (LDOCE). Each frame is connected via an IS-A link to some other word (spelling form) in the dictionary. Our goal in this paper is to refine that work so as to connect each frame to another word *sense* in the dictionary. This insures that properties can be consistently inherited in this graph structure (since A IS-A B allows A to inherit properties of B). We can think of this task as building a taxonomy of the word senses in LDOCE.

This paper presents our techniques for automating this task for noun definitions, using the special coded information found in the machine readable version of LDOCE. We also present ways to extract other semantic relations automatically as part of the process.

2. Background

Dictionary definitions of nouns are normally written in such a way that one can identify a "genus term" for the headword (the word being defined) via an IS-A relation. The information following the genus term, the differentia, serves to differentiate the

** Present address: Department of Computer Science, North Dakota State University, Fargo, ND 58105

* This research was supported by the New Mexico State University Computing Research Laboratory through NSF Grant No. IRI-8811108 — Grateful acknowledgement is accorded to all the members of the CRL Natural Language Group for their comments and suggestions.

headword from other headwords with the same genus. For example, (from LDOCE):

knife - a blade fixed in a handle, used for cutting as a tool or weapon.

Here "blade" is the genus term of the headword "knife" and "fixed in a handle, used for cutting as a tool or weapon" yields differentia. In other words, a "knife" IS-A "blade" (genus) distinguished from other blades by the features of its differentia. In order to create a taxonomy of word senses, this genus term must be identified and also sense-tagged (in this case, by ruling out blade of grass, propeller blade, and an amusing fellow).

Previous research on constructing taxonomies from machine readable dictionaries, i.e. Amsler & White (1979) and, to some extent, Chodorow et. al. (1985), has relied on a good deal of human intervention whenever the taxonomy is composed of word senses rather than spelling forms. Nakamura & Nagao (1988) automatically constructed a taxonomy, but did not distinguish the senses of nouns and hence cannot allow inheritance of properties along the links of the implied network created by the taxonomy. Because of the semantic category markings in LDOCE, we have been able to develop heuristic procedures (described in section 4), that, to a great extent, automate the task of developing a hierarchy of word senses.

Constructing taxonomies from the genus terms of definitions forces one to take a stand on how to treat a large class of noun definitions which are not as "standard" as the definition given above for **knife.** The characteristic property of these definitions is that the head of the first noun phrase (the usual place to find a genus term) seems vacuous, and another easily identifiable noun in the definition gives information about the headword. Nakamura & Nagao (1988), identify these non-standard definitions syntactically as:

{det.} {adj.}* <Function Noun> of <Key Noun> {adj. phrase}*

For example, the following definitions have the property that the head of the noun phrase following the "of" is more semantically relevant to the headword than the head of the first noun phrase.

arum (LDOCE) - a tall, white type of Lily

cyclamate (LDOCE) - any of various man-made sweeteners ...

deuterium (Merriam-Webster Pocket Dictionary) - a form of hydrogen that is twice the mass of ordinary hydrogen

academic (LDOCE) - a member of a college or university

The form of this type of definition is predictable whenever certain words are used as the head of the first noun phrase. Amsler and White (1979) kept a list of these words, referring to them as partives and collectives. Nakamura & Nagao (1988) call them Function Nouns. Chodorow et al., (1985) refer to a subset of these as "empty heads". Since we dIS-Agree with certain elements of these characterizations, we will use the terminology "disturbed heads". The question at issue is: what to do with these cases?

In the original work of Amsler and White (1979) with the Merriam Webster Pocket Dictionary (MPD, 1964), the disturbed head cases were handled by asking paid human "disambiguators" to sense-tag the head of the first noun phrase in the definition and also to sense-tag any other noun in the definition which "made a significant semantic contribution to an IS-A link" (Amsler and White, 1979: p. 55) with the headword being defined (i.e. for the **deuterium** definition above, "hydrogen" was sense tagged as well as "form"). The taxonomy actually contained both a link from **deuterium** to "form" and a link from **deuterium** to "hydrogen", although the hydrogen sense was marked in a special way to indicate it is not the syntactic head of the definition. In cases like the "hydrogen" example just given, the marked "semantic contributors" were never given ancestors, since the link often represented a more loosely defined relation than the strictly transitive "is a subset of" definition of IS-A, which ideally relates the headword and its genus sense. This degenerate form of IS-A precludes inheritance in the network. It is included in the taxonomy in order to form links to words which may not be related in a strict IS-A sense, but which convey useful information about the word being defined.

There have been various proposals over the years suggesting different specialized link types to be added to the taxonomy (besides the degenerate IS-A). Markowitz et al., (1986) suggest HAS_MEMBER links be created in definitions which use the phrase "member of" (i.e. "college" HAS_MEMBER "academic" in the definition of **academic** above). Nakamura & Nagao (1988) identify 41 different function nouns and replace the IS-A link in their taxonomy with various other links in these cases (except in the "kind of", "type of", etc., definitions). Amsler (1980) suggests the incorporation of an IS_PART_OF link in addition to the IS-A links in the earlier taxonomy of Amsler & White (1979).

Chodorow et al., (1985) automate the genus finding process for nouns and verbs in Webster's Seventh (W7, 1967). However, in their work, only the spelling form of the genus is identified automatically; the sense selections are made by humans. The disambiguation here is not to attach a sense number, but rather to perform a function termed "sprouting"

which interactively selects among all words which have a given word-sense as a genus. Their taxonomy contains only IS-A links, but they partially attack the "disturbed head" problem by identifying a small class of what they call "empty heads". The effect of their method is to skip over seemingly vacuous terms (located where a genus is usually expected), and treat the more semantically relevant term as the actual genus.

3. Description of LDOCE and its limitations

The Longman Dictionary of Contemporary English (LDOCE; Procter et al. 1978), is a full-sized dictionary designed for learners of English as a second language that contains 41,122 headword entries, defined in terms of 72,177 word senses, in machine-readable form (a type-setting tape). The book and tape versions of LDOCE both use a system of grammatical codes of about 110 syntactic categories which vary in generality from, for example, *noun* to *noun/count* to *noun/count/followed-by-infinitive-with-TO*. The machine readable version of LDOCE also contains "box" and "subject" codes that are not found in the book. The box codes use a set of primitives such as *abstract*, *concrete*, and *animate*, organized into a type hierarchy. This hierarchy of primitive types conforms to the classical notion of the IS-A relation as describing proper subsets. These primitives are used to assign type restrictions on nouns and adjectives, and type restrictions on the arguments of verbs. The subject codes are another set of terms organized into a hierarchy. This hierarchy consists of main headings such as *engineering* with subheadings like *electrical*. These terms are used to classify words by subject. For example, one sense of **current** is classified as *geology-and-geography* while another sense is marked *engineering/electrical*.

This paper's overall goal is to make implicit semantic information in the dictionary explicit. However, we are not doing "psychology of lexicography": the test of our derived structures is not whether they match any conscious or unconscious inferences of lexicographers, but whether they improve subsequent natural language processing (e.g. machine translation). Nor are we in any way concerned here with low-level issues of the syntax of dictionary entries, its expression on tapes or pages, or by what device the information enters the computer. It is of course a strong assumption that a fallible dictionary designed for human learners of a second language also implicitly contains the information needed for successful natural language processing. We make this assumption consciously as an empirical hypothesis. Even though LDOCE has beneficial features, such as its restricted vocabulary for sense definition, we see no reason to believe at this stage that the taxonomic relations we derive are in any way non-standard.

4. Automatically finding genus senses

A heuristic procedure that automatically finds disambiguated genus terms for nouns has been developed. The initial stage of this procedure is to automatically identify the genus term in the definition. The Lexicon Provider (Slator 1988a, 1988b; Slator and Wilks, 1987, 1990) mentioned above has a parser which does this. The parser accepts LDOCE definitions as Lisp lists and produces phrase-structure trees. LDOCE sense definitions are typically one or more independent clauses composed of zero or more prepositional phrases, noun phrases, and/or relative clauses. The syntax of sense definitions is relatively uniform, and developing a grammar for the bulk of LDOCE has not proven to be an intractable problem. Chart parsing was selected for this system because of its utility as a grammar testing and development tool. The chart parser is driven by a context free grammar of 100-plus rules and has a lexicon derived from the 2,219 words in the LDOCE core vocabulary. The parser is left-corner, and bottom-up, with top-down filtering. The context-free grammar driving the chart parser is virtually unaugmented and, with certain minor exceptions, no procedure associates constituents with what they modify. Hence, there is little or no motivation for assigning elaborate or competing syntactic structures, since the choice of one over the other has no semantic consequence. Therefore, the trees are constructed to be as "flat" as possible. The parser also has a "longest string" (fewest constituents) syntactic preference. The grammar is still being tuned, but the chart parser is already quite successful and works extremely well over a fairly wide range of examples from the language of content word definitions in LDOCE. Ninety-Five percent result in a parse tree for the entire definition text. Five percent of the analyses fail at some point. In those cases where it fails the parser still returns a partial parse (of the leading constituents in the definition text), and this is the most important part of a definition anyway.

The second phase of this procedure is to find the correct sense of the genus term that has been identified by the parser. To do this, we have constructed a program called the Genus Disambiguator, which takes as input the subject codes (pragmatic codes) and box codes (semantic category codes) of the headword, taken from the machine readable version of LDOCE, and the spelling form of the genus word which has been identified by the parser described above. The output is the correct sense of the genus word.

The codes in LDOCE seem to support the thesis that the genus for a noun must be a noun, and that the semantic category of the genus word must be

the same as, or an ancestor of, the semantic category of the headword. The word ancestor refers to superordinate terms in the hierarchy of semantic codes defined by the Longman lexicographers. The strategy of the algorithm is:

1. choose the genus sense whose semantic codes identically match with the headword, if possible;
2. if not, choose the sense whose semantic category is the closest ancestor to the semantic category of the headword;
3. in the case of a tie, the subject codes are used to determine the winner;
4. if subject codes cannot be used to break the tie, the first one of the tied senses which appears in the dictionary is chosen (since more frequently used senses are listed first in LDOCE).

The following examples illustrate the algorithm. The ordered pair following the headword consists of the box code and subject code as found in dictionary (the notation following that is the English gloss for these particular codes). Many definitions are not given a subject code in LDOCE and a dash (--) is used here to indicate that. Consider the following LDOCE definition.

ambulance - (J:movable-solid, AUZV: Automotive /Vehicle-Types) - motor vehicle for carrying sick or wounded people esp. to hospital

The genus of **ambulance** is the word "vehicle", which is found by the Lexicon Provider's parser; therefore the input to the Genus Disambiguator is:

(ambulance J AUZV vehicle)

The following are the LDOCE definitions for the noun senses of vehicle.

vehicle-1 - (J:movable-solid, TNVH: Transportation /Vehicles) - something in or on which people or goods can be carried from one place to another ...

vehicle-2 - (T:abstract,--) - something by means of which something else can be passed on or spread: *Television has become an important vehicle for spreading political ideas*

vehicle-3 - (T:abstract,--) - a means for showing off a person's abilities: *The writer wrote this big part in his play simply as a vehicle for the famous actress*

In this case the Genus Disambiguator chooses the first sense of **vehicle,** because of the match between the "movable-solid" semantic codes, therefore the output is "vehicle-1". There are many cases, however, where a direct match is not found. Consider the following LDOCE definition.

dart - (J:movable-solid,GA:Games) - a small sharp-pointed object to be thrown, shot, etc. ...

The word "object" is the genus of **dart,** making the input to the Genus Disambiguator

(dart J GA object)

The following are the LDOCE noun definitions for "object"

object-1 - (S:movable-solid,--) - a thing

object-2 - (1:human-and-solid,--) - something or someone that produces interest or other effect ...

object-3 - (1:human-and-solid,--) - something or someone unusual or that causes laughter

object-4 - (T:abstract,--) - purpose; aim

object-5 - (T:abstract,LN:Linguistics-and-Grammar) - word(s) saying with whom or with what, a PREPOSITION ...

In this example there is no direct match between the semantic codes of the headword, **dart,** and any of the senses of the genus, "object"; therefore the Genus Disambiguator must traverse up the type hierarchy, described in section 3, to find the closest ancestor of boxcode "J" (movable-solid) that is present in the definitions of the genus word. In this case, boxcode "S" (solid) is found one level above "J" and the output is "object-1". There are still other cases, however, when more than one sense definition has semantic codes matching the codes of the headword. Consider the following LDOCE definition.

flute - (J:movable-solid,MU:Music) - a pipelike wooden or metal musical instrument with finger holes, played by blowing across a hole in the side ...

The genus of **flute** is the word "instrument"; therefore, the input to the Genus Disambiguator is

(flute J MU instrument)

The following are the LDOCE definitions for instrument.

instrument-1 - (J:movable-solid, HWZT: Hardware/Tools) - an object used to help in work: *medical instruments*

instrument-2 - (J:movable-solid,MU:Music) - ... an object which is played to give musical sounds (such as a piano, a horn, etc.) ...

instrument-3 - (Z:unmarked,--) - someone or something which seems to be used by an outside force to cause something to happen: *an instrument of fate*

In this case both the first and second senses of **instrument** are marked as "J", (movable-solid), which matches perfectly with the selection restriction for **flute.** However, the tie is broken by appeal to the subject code, Music, which selects the second sense of **instrument** as the genus of **flute,** and the output is "instrument-2".

There are occasional failures, many of which appear to be due to unusual markings in LDOCE. For example, the LDOCE definition for banana is:

> **banana** - (P:plant,PMZ5:Plant-Names) - any of several types of long curved tropical fruit, shaped like a thick finger, with a yellow skin and a soft, usu. sweet, inside ...

The genus of **banana** is the word "fruit", and the input to the Genus Disambiguator is

(banana P PM fruit)

The following are the LDOCE definitions for fruit.

> **fruit-1** - (J:movable-solid,FO:Food) - an object that grows on a tree or bush, contains seeds, is used for food, but is not usu. eaten with meat or with salt
>
> **fruit-2** - (S:solid,FO:Food) - these objects in general, esp. considered as food ...
>
> **fruit-3** - (J:movable-solid,FO:Food) - a type of this object
>
> **fruit-4** - (J:movable-solid,BO:Botany) - a seed-containing part of any plant
>
> **fruit-5** - (T:abstract,--) - a result, good or bad: *His failure is the fruit of laziness*
>
> **fruit-6** - (M:male/human,--) - fellow (in the phr. *old fruit*)

In this case, **banana** is marked as a "plant" but, for some reason, the likely candidates defined under fruit are all marked "solid" or "movable-solid". Since neither solid nor movable-solid are ancestor to plant in the LDOCE type hierarchy they are all equally bad, from the point of view of the Genus Disambiguator, and the default is invoked, which is to choose the lowest numbered sense from among the competitors. Therefore the first sense is selected and the output is "fruit-1". This happens to be correct, but it is an unsatisfying resolution.

In a piece of related work, Slator (1988a) has implemented a scheme in the Lexicon Provider which imposes deeper structure onto the LDOCE subject hierarchy (e.g. terms like Food, Botany, and Plant-Names in the "fruit" definitions above) relating these categories in a natural way, in order to discover important relationships between concepts within text. This manual restructuring simply observes that words classified under **Botany** have pragmatic connections to words classified as **Plant-Names,** as well as connections with other words classified under **Science** (connections *not* made by the LDOCE hierarchy as given), and that these connections are useful to exploit.

The Lexicon Provider system relates these codes through a specially restructured hierarchy created for that purpose, making **Communication, Economics, Entertainment, Household, Politics, Science,** and **Transportation** the fundamental categories. Every word sense defined with a subject code therefore has a position in the new hierarchy, attached below the node for its subject code. Once this feature is implemented in the Genus Disambiguator, the subject code hierarchy can be used to resolve the "banana-fruit" case above in a somewhat more satisfactory way, by choosing sense 4 of **fruit.**

5. Identifying other relationships automatically

The identification of a satisfactory genus term and the construction of a taxonomy is not straightforward in all cases. It is clear that the problems in this area are difficult, numerous, and can be seen to encompass a great variety of relationships. We believe that a thorough study of this shadowy area is necessary in order to make optimal use of the semantic information available in machine readable dictionaries. Although we do not have complete solutions, we have additional insights into the problem of extracting supplementary information from the "disturbed head" definitions.

Chodorow et al. (1985) examined a phenomenon that they described as follows:

> "If the word found belongs to a small class of "empty heads" (words like *one, any, kind, class, manner, family, race, group, complex,* etc.) and is followed by *of,* then the string following *of* is reprocessed in an effort to locate additional heads." (pg. 301).

Although the empty head rule seems to be a reasonable one in certain situations, we have reservations about its use. The empty head rule produces undesirable effects in an IS-A hierarchy for some of the collective words (that Chodorow et al. treat as empty): set, group, class etc. Our response to the empty head phenomenon is to process them in the same way, but limiting this processing to a much smaller set; that is, to those heads that are truly empty -- the set containing {one, any, kind, type}.

Consider the LDOCE definition:

> **canteen** - (British English) a set of knives, forks and spoons, usu. for 6 or 12 people

Since "set" is one of the empty heads for Chodorow et al., their procedure would create IS-A links to

"knives", "forks" and "spoons", and this again would violate the inheritance properties that should be preserved via IS-A links. Our response to the collective heads, {set, group, collection, class of, family of} (which we maintain are not truly empty, simply disturbed), is to form a taxonomic link to the correct sense of "set," "group," or "class" etc. and to form a HAS_MEMBER link to the noun or nouns which describe the elements of the collective (as found in the differentia of the headword definition). Further, we propose that definitions in which the genus term is plural be treated in the same way as those which begin with "a set of".

In general, our view is that the disturbed heads should be grouped in the sense of Nakamura & Nagao (1988), and that additional links (like HAS_MEMBER, IS_PART_OF, etc.) should be created whenever they are appropriate. However, it is our position that IS-A links should also be created for every word sense given in the dictionary. Moreover, in order to maintain inheritance and transitivity in the IS-A network, a strict "subset of" definition of IS-A should be maintained.

Unlike Nakamura & Nagao (1988), we propose that "member of" definitions should not be grouped with the "set of", "group of" definitions. All but one "member of" definition in LDOCE uses "member of" to mean "person who is a member of". We recommend that in this case, a link be created from the headword to "person", and that the appropriate MEMBER-OF link is constructed. The exceptional case, where "member of" does not refer to a person, is in the definition of **feline** : "a member of the cat family." This case must be treated separately, since it is impossible to identify the correct sense of the word "member" here, given that all these senses, in LDOCE, are marked as referring to a human or a part of the human body.

The difficulty of these many varieties of special cases (and they are not so special, since there are hundreds of them in the dictionary), is that they call into question certain of the long held assumptions about the taxonomic structure of dictionaries. The conventional wisdom has always been that dictionary definitions contained a genus term (a term more general than the one being defined), and that this term could almost invariably be found in the first phrase of the definition text. Further, the exceptions to this convention, the "empty heads" like "one of" or "any of", have been viewed as being similarly well-behaved. Our investigations lead us to conclude that things are not so simple as they once appeared; and the question of what to do with these troublesome cases is far from resolved.

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. NSF Technical Report. (MCS77-01315).

Amsler, Robert A. (1980). The Structure of the Merriam-Webster Pocket Dictionary. Technical Report. (TR-164). University of Texas at Austin. Ph.D. Thesis.

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, pp. 299-304. Chicago, IL.

Markowitz, Judith, Thomas Ahlswede, and Martha Evens (1986). Semantically Significant Patterns in Dictionary Definitions. Proceedings of the 24th Annual Meeting of the Association for Computational Linguistics, pp. 112-119. New York.

MPD (1964). *The New Merriam-Webster Pocket Dictionary.* Pocket Books, New York.

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

Procter, Paul et al. (1978). *Longman Dictionary of Contemporary English (LDOCE).* Harlow, Essex, UK: Longman Group Ltd.

Slator, Brian M. (1988a). Constructing Contextually Organized Lexical Semantic Knowledge-bases. *Proceedings of the Third Annual Rocky Mountain Conference on Artificial Intelligence (RMCAI-88).* Denver, CO, June 13-15, pp. 142-148.

Slator, Brian M. (1988b). Lexical Semantics and a Preference Semantics Analysis. Memoranda in Computer and Cognitive Science. *(MCCS-88-143).* Las Cruces, NM: Computing Research Laboratory, New Mexico State University. (Doctoral Dissertation).

Slator, Brian M. and Yorick A. Wilks. (1987). Towards Semantic Structures from Dictionary Entries. *Proceedings of the Second Annual Rocky Mountain Conference on Artificial Intelligence (RMCAI-87)* Boulder, CO, June 17-19, pp. 85-96.

Slator, Brian M. and Yorick A. Wilks (Forthcoming - 1990). Towards Semantic Structures from Dictionary Entries. *Linguistic Approaches to Artificial Intelligence.* Edited by Andreas Kunz and Ulrich Schmitz. Frankfurt: Peter Lang Publishing House. (Revision of *RMCAI-87*).

W7 (1967). *Webster's Seventh New Collegiate Dictionary.* C. & C. Merriam Company, Springfield, MA.

Wilks, Yorick A., Dan C. Fass, Cheng-Ming Guo, James E. McDonald, Tony Plate, and Brian M. Slator (1988). Machine Tractable Dictionaries as Tools and Resources for Natural Language Processing. *Proceedings of the 12th International Conference on Computational Linguistics (COLING-88),* pp. 750-755. Budapest, Hungary. Aug. 22-27

Wilks, Yorick A., Dan C. Fass, Cheng-Ming Guo, James E. McDonald, Tony Plate, and Brian M. Slator (1989). A Tractable Machine Dictionary as a Resource for Computational Semantics. *Computational Lexicography for Natural Language Processing.* Edited by Bran Boguraev and Ted Briscoe. Harlow, Essex, UK: Longman and New York: Wiley and Sons. pp. 193-228.

Wilks, Yorick A., Dan C. Fass, Cheng-Ming Guo, James E. McDonald, Tony Plate, and Brian M. Slator (Forthcoming - 1990). Providing Machine Tractable Dictionary Tools.*Computers and Translation.* Also to appear in *Theoretical and Computational Issues in Lexical Semantics (TCILS).* Edited by James Pustejovsky. Cambridge, MA: MIT Press.