Not for its own sake: knowledge as a byproduct of natural language processing

William B. Dolan
billdol@microsoft.com

Stephen D. Richardson
steveri@microsoft.com

Microsoft Research
Microsoft Corporation
Redmond WA 98052
(206) 882-8080

Abstract

Recent claims in the literature as well as current research trends suggest that a long-held goal of natural language processing, the ability to map automatically from machine readable dictionaries into structured knowledge bases that can be used for various artificial intelligence tasks may be impossible. This paper argues to the contrary, describing an extremely large and rich lexical knowledge base which has been created almost as a byproduct of ordinary morphological, syntactic and semantic processing within a broad-coverage NLP system. This approach to mapping between text and knowledge representation appears to offer the only viable hope for creating knowledge bases large and rich enough to support complex tasks like machine understanding and domain-independent reasoning.

Keywords: lexical knowledge base, machine readable dictionaries, knowledge representation, natural language processing

1. Introduction

This paper describes what we believe to be the largest scale effort presented to date aimed at automatically mapping natural language (NL) text into a finely structured knowledge base (KB). In addition to presenting an overview of this work, we argue that broad efforts to exploit NL text as a source of knowledge will be most effective when knowledge acquisition is treated as a byproduct of ordinary processing within a broad coverage NLP system, rather than as an end unto itself.

The goal of broad-coverage NL understanding is inextricably tied to the availability of rich KBs providing information about how words are related to one another and to the world around us. Exactly how these KBs should be built, though, has remained problematic. WordNet (Miller 1990) demonstrates that it is possible to build a useful and impressive LKB by hand. However, we do not believe that hand-coding efforts will ultimately be capable of providing high quality semantic information in the volume (and in the range of languages) necessary to support complex tasks like word sense disambiguation, machine translation and the complex reasoning needed to identify relationships between discourse entities. Deciding what knowledge to represent and how to represent it are problems that become more difficult once we move away from the relatively concrete field of identifying semantic relationships among words, and into more abstract territory, and so WordNet's success cannot be interpreted as evidence that it is feasible to hand-code arbitrarily large KBs.

Among the problems that we believe render hand-built knowledge resources inadequate are the cost and time required to build such resources, and quality control issues. In our view, the resources required to hand-code large-scale KBs could be better devoted to developing strategies for automating knowledge acquisition, and specifically to developing NL processing tools for this purpose. An approach to knowledge acquisition which relies on NLP has a number of advantages over hand-coding approaches. One is its flexibility: hand-coding weds a project to decisions about KB structure that often prove troublesome later in the life of a project. If, however, the knowledge base is dynamically generated from NL text, it is relatively trivial to change the engine to modify the way that the knowledge is represented.

Other advantages of an approach to knowledge representation which blurs the distinction between these representations and NL text are more obvious. A vast amount of world knowledge is stored in NL, and the ability to process this text will arguably lead to the development of KBs which contain enough information to support the broad goals of Artificial Intelligence.

2. Extracting structured information from natural language

The Microsoft NLP group is involved in a long-term effort to construct an engine capable of mapping arbitrary English sentences and texts into structured semantic representations. One aspect of this work has been the development of a very large lexical knowledge base (LKB), which has been automatically derived from two machine-readable dictionaries (MRDs), the Longman Dictionary of Contemporary English (LDOCE) and the American Heritage Third Edition (AHD3).

The idea that dictionary definition text might be exploited to produce structured lexical resources rich enough to support broad coverage NLP systems is a very attractive one. Many ambiguities in NL, including anaphora, polysemy and certain types of syntactic ambiguities, cannot be resolved without access to large amounts of real world knowledge, and dictionaries have long tantalized researchers as ready-made sources of exactly this sort of knowledge. Throughout the 1970s and 1980s, research on extracting semantic information from MRDs flourished (Amsler 1980, Byrd et al. 1987, Klavans et al. 1990, Markowitz et al. 1986 and Wilks et al. 1989). The research described in this paper is based in particular on work in this area begun by Jensen and Binot (1987), continued by Montemagni & Vanderwende (1992) and further refined and extended by Dolan et al. (1993), Vanderwende (1996) and Richarson (1996).

Dictionaries would seem to be the perfect testbed for exploring knowledge-extraction schemes. For a variety of reasons, they appear to present a much simpler environment for carrying out this task than does free text:

- Dictionary definitions frequently consist of short sentence fragments, with the syntactic identity of this fragment being predictable. (Typically, for instance, noun definitions are expressed by noun phrases and verb definitions by verb phrases.) In theory, at least, this should simplify syntactic analysis.
- Since lexicographers explicitly aim to define words as thoroughly and economically as possible, dictionaries present none of the problems of truth value or relevance that other corpora might: if information is included in a definition, we can assume that it is both accurate and highly salient.
- The types of rhetorical and discourse devices encountered in dictionaries are much more restricted than in free text. For example, only a small number of standardized devices are used to relate definitions within an entry. Typically this involves simple intersentential anaphora:

  sculpture, noun (LDOCE)
  1. the art of shaping solid figures...
  2. (a piece of) work produced by ***this art***

  peepul, noun (AHD)
  1. a fig tree native to India...
  2. ***it*** is regarded as sacred by Buddhists
- Organizational conventions within dictionaries should ease their mapping into knowledge bases. For example, Guthrie & Bruce (1992) report that LDOCE lexicographers restricted themselves to the first sense of definition words 70% of the time. This suggests that the word sense disambiguation problem, generally regarded as one of the more difficult problems for free text NL understanding, can be much simpler within these restricted corpora.

3. A Waste of Time?

Despite the apparent advantages of working on this restricted genre of text, interest in MRD research has faded significantly during this decade, to the point that a recent paper (Ide & Veronis 1993) could be provocatively titled "Extracting knowledge bases from machine readable dictionaries: have we wasted our time?" The authors cite a number of reasons for the decline in MRD research, including excitement over recent developments in statistical methods for extracting information from large free text corpora, and some of the practical problems involved in transforming a publisher's tape into an MRD.

More generally, though, Ide & Veronis raise two serious questions about the goal of automatically mapping MRDs into useful KBs. First of all, they claim that the *IsA* hierarchies that have been extracted from these sources are of limited utility, being relatively flat and uninformative. Secondly, they argue that even the relatively constrained linguistic problems posed by dictionaries are simply too difficult to allow the automated extraction of more than simple word-association information, and that more ambitious efforts are doomed to failure.

Ide & Veronis show that even within the "constrained" environment of a single dictionary, there can be a large range of ways of expressing even relatively simple semantic relationships. As an example they consider the variety of paraphrases used to express the fact that a "handle" is a part of a utensil in various definitions in the Collins English Dictionary. These include "usually having a handle", "container with a handle", "a long-handled spoon", "container...carried by means of a handle", and "consisting of a pointed metal spiral attached to a handle". Ide & Veronis worry that mapping such varied constructions into the same knowledge representation might require a nearly open-ended set of rules. They conclude that the goal of automatically extracting knowledge bases from existing dictionaries is hopeless, and they suggest that it is simpler to build a knowledge base by hand.

While Ide & Veronis' critique is specifically aimed at MRD research, this second issue is equally relevant to other efforts aimed at extracting structured semantic information from NL prose. In fact, once we move away from the environment of MRDs and into free text, the paraphrase problem is surely much more difficult; dictionary language, after all, represents a subset of this larger problem. If even dictionaries cannot be handled in this way, then what hope can there be for more complex types of prose? More broadly, accepting Ide & Veronis' conclusion entails abandoning the idea that NL can ever be treated as a knowledge representation language, since it suggests that it is impossible to build a general engine that can manipulate it for computational purposes. In our view, this further implies that applications like machine understanding and domain-independent reasoning are equally infeasible, given our conviction that it will be impossible to hand-construct KBs rich enough to support them.

4. MindNet

There are a number of reasons to regard with suspicion Ide & Veronis' claim that dictionaries are not useful targets for automatic information extraction schemes, and that "the difficulties are so considerable that such methods are unlikely to succeed" (p.262). First of all, dictionaries are no more than peculiarly limited and interesting textual corpora, and to claim that no structured information can be gleaned from them is tantamount to claiming that any level of natural language understanding is unattainable. This is clearly too pessimistic. While it may be true that efforts narrowly aimed at extracting *IsA* terms from dictionary definitions can involve more effort than they are worth, it is our experience that mapping from dictionary text into structured semantic representations that are far richer than simple *IsA* relationships is both feasible and extremely productive, within the context of a broad coverage NL understanding system.

The remainder of this section describes the procedures we have used to create a LKB from two machine-readable dictionaries, LDOCE and AHD3. This LKB, called MindNet, is created through automatically extracting and structuring information contained in the dictionary in the following steps:

1. Parsing dictionary definitions and example sentences with a broad-coverage parser that produces deep dependency-type structures,
2. Extracting hierarchical structures of semantic relations of the form "word1 *relation* word2" from the parse structures,
3. Automatically disambiguating word senses so that relations are now of the form "sense1 *relation* sense2",
4. Inverting the semantic relation structures and propagating them across all entries in the LKB, thereby producing a highly connected network of semantic relations,
5. Assigning weights, based on a distribution-adjusted frequency, to each semantic relation and sequence (path) of relations in the semantic relation structures.

The project architecture involves bootstrapping on many levels, and from the outset, one of our goals has been to move beyond dictionaries, to a point where the system can be used to map freely between arbitrary strings of text and their semantic representations. The idea of using the knowledge stored in books and on-line corpora for various AI tasks is almost inevitable in the context of a project which aims at broad coverage NL understanding: in our experience, the sorts of representations needed to capture an appropriately abstract level of linguistic description have turned out to be well-suited for the sorts of inferencing required for anaphora resolution and discourse processing.

4.1 Extracting Semantic Relations

This section describes how we extract semantic relation information from dictionary definitions and example sentences. The process has been described in detail elsewhere (Montemagni & Vanderwende 1992, Vanderwende 1996), but an overview is given here to provide background information for the subsequent steps in the creation of the LKB.

The first step in the extraction process is to parse the definitions and example sentences with a broad-coverage parser of English. The parser produces approximate syntactic parses of general English text segments (sentences

as well as fragments) by adhering to principles and organization similar to those outlined by Jensen (1993). It also produces deeper, dependency-type structures, called logical forms, that identify deep subjects and deep objects of passive verbs and anaphoric references of pronouns, and that resolve the scope of various coordinated structures.

Once a definition or example sentence is parsed, semantic relations are identified in the parse structures produced. A semantic relation consists of a unidirectional link, labeled by a semantic relation type, between two words. For example, in the LDOCE definition for "fruit" shown below, the verb phrase "is used for food" is mapped into a **Purpose** relation between "fruit" and "food", i.e., "fruit has a **Purpose** of food", or "fruit--**Purpose**-->food".

fruit:
"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"

LDOCE definition for "fruit"

Each semantic relation type (e.g., **Purpose**) has associated with it one or more structural patterns that are matched against the parse structures produced by the parser for each definition during extraction. This allows for the identification of semantic relations even when there are long distance dependencies or syntactic variations in the definition. In the definition of ***fruit*** above, the passive verb ***eaten*** does not have a grammatical object in the traditional sense. The logical form of this definition, however, shows that "fruit" is the deep object of "eat". This enables the structural pattern matching to identify a **TypicalObject** relation between "eat" and "fruit".

Approximately 15 types of semantic relations (and their inverses) are currently identified by structural pattern matching. These include **Cause, Domain, Hypernym** (*IsA*), **Location, Part, Material** and **Typical_Subject**. Contrary to the concern voiced by Ide & Veronis, it has been our experience that, within dictionaries, the number of structural patterns required to capture the various English paraphrases of a relation like **Material** is relatively small. In part this is due to the fact that these structural patterns apply to logical forms, which already abstract away from certain types of syntactic paraphrases. Thus, for example, the same pattern can be used to identify **Material** relationships expressed by both active and passive constructions. Just four patterns capture the various English paraphrases used to express this relationship within LDOCE and AHD3. Although the range of potential paraphrases is certainly greater in free text, it seems likely that most intended **Material** relationships expressed in this domain can be handled by a sharply limited set of structural patterns.

Through the matching of structural patterns to the parse structures and logical forms of definitions, vast numbers of semantic relations are extracted automatically across the entire dictionary and added to each entry in the LKB, dubbed *MindNet*. From over 165,000 (single word noun and verb definitions) of the approximately 241,000 LDOCE definitions, over 300,000 semantic relations have been extracted and added to entries in the LKB. Richardson (1993) reports that the overall accuracy of the semantic relations is 78%, with the accuracy of certain relations for LDOCE being as high as 87%.

Several improvements have been made recently, especially in using a bootstrapping process to vet and modify relations that were extracted on earlier passes over the dictionary (Vanderwende 1996). This possibility is especially important for improving the accuracy of particular semrels such as **Part** and **Material**, which are less reliably associated with particular syntactic/lexical configurations than are values for relations like **TypicalObject and Hypernym**. For instance, nothing about the syntactic construction "a *noun* of *noun*" tells us that ***a necklace of gold*** should yield the semrel ***necklace*—Matr→*gold***, but ***necklace of pearls*** should yield ***necklace*—Part→*pearls***). A two pass strategy makes this possible, though, with the initial pass assigning a default likelihood) that ***gold*—Hypernym→*metal***. This fact is used on a second pass to alter the erroneous ***Part*** relation in ***necklace*—Part→*gold*:** "metal" belongs to a small set of words (including "wood", "cloth". "stone") which likely reflect a ***Material*** relation.

The advantage of building the LKB in a fully automated manner, rather than by hand, was made clear in the summer of 1995, when the AHD3 data were first processed. Overnight, the size of our LKB tripled, with relatively minor degradation in the overall quality of the data. New structural patterns have since been added to handle some of the novel syntactic and lexical constructions used in this larger dictionary, but the vast bulk of the machinery used to process LDOCE remained unchanged. Adding this information by hand would have been a massive undertaking

Perhaps most importantly, the additional work needed to improve the quality of information extracted from AHD3 is entirely reusable within the context of our system. As noted above, dictionary language is merely a subset of general English text, our ultimate processing target. Since any work devoted to normalizing paraphrases within LDOCE or AHD has immediate benefits for our broad coverage NLP system, this work is amortized over the whole life of the system and its use in processing e-mail, encyclopedias, instruction manuals, and so on. For this reason, it is always profitable for us to produce general solutions for processing problems

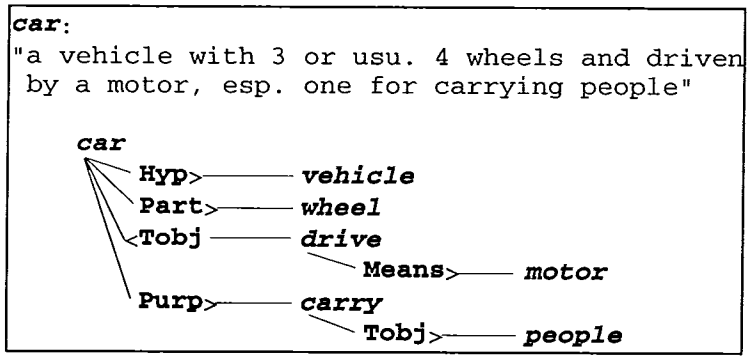

Figure 1: Semrel structure for a definition of **car**

that might affect only one or two definition structures within the dictionary. In contrast, efforts whose primary focus is the automated extraction of knowledge from dictionaries could more efficiently resort to hand-coding in such cases.

4.2 Inverting Semantic Relation Structures

The cluster of semantic relations extracted from a given definition or example sentence actually forms a hierarchical structure that has as its root node the headword of the LKB entry containing the definition. The *semantic relation structure* (or semrel structure) for the definition of "car" is shown in Figure 1. Directionality of the relations is indicated by arrow-like brackets <> on either side of the relation type, and the relations may be read by inserting the words "has a" before a left-to-right relation and the words "is a" and "of on either side of a right to left relation. Thus, "car has a **Hypernym** vehicle" and "car is a **TypicalObject** of drive" are two of the relations in the structure above.[1] These representations have something of the flavor of Case Grammar (Fillmore, 1968), and are also similar in spirit to Sowa's Conceptual Graphs.

Although the definition and corresponding semrel structure above provide a useful set of information about "car", it is by no means a complete or even near complete set. Definitions often fail to express many basic facts about word meanings, facts that should obviously be included in an LKB that would be useful for broad-coverage NL understanding. Although the information in this example is indeed useful, it mentions only one of potentially dozens or even hundreds of generally recognized "car" parts, only one action performed with a "car", and one purpose of "cars".

In order to overcome this ostensible limitation on the information that can be extracted, one might access the semrel structures associated with the words in the above structure (e.g., "vehicle", "wheel") in a way that is similar to the forward spreading activation in the networks of Veronis & Ide (1990). In this case, however, such a strategy is not especially productive, yielding some information already given by the definition of "car", as well as more general information (about "vehicles", "wheels", etc.), but not many new details specific to "cars".

A more effective solution is to access the semrel structures associated with words that are related to "car" by those structures. This is the same as looking up words in the dictionary whose definitions mention the word "car". There is a great deal of additional information about a given word's meaning that is typically stored not in the entry for that word itself, but rather in the entries for other words that mention that word. For instance, it is relatively unusual to find the words which describe the parts of some object in the dictionary entry for that object. Instead, the relationship between the words for these parts and the larger object is defined generally in the dictionary definitions for the parts themselves. Other parts of the larger object may even be found in definitions other than for the parts themselves. For instance, one source of the fact that an "engine" is **Part** of a "car" is the LDOCE definition

hood, n "the bonnet covering the engine of a car"

It is these backward-linking semantic relations, and other relations in their associated structures, that serve to dramatically increase the amount of information accessible

[1] Note that while the definition of "car" tells us that cars have three or, more typically, 4 wheels, this information is not captured by the corresponding semrel structure. Gaps of this kind are common in our KB, reflecting its status as a work in progress. In this instance, the complex quantifier "3 or usu. 4" is correctly analyzed by our parser, but no rule subsequently maps this portion of the parse structure into a corresponding modification of "wheel". Once such a rule is added to the system, a new version of the LKB will be generated, capturing not only this fact about cars and wheels, but also similarly-expressed facts about other words across the dictoionary.

for a given word. By making these backward-linked relations explicit, together with the already explicit forward-linked relations, the LKB thus becomes a highly interconnected network of semantic information. This network captures both the paradigmatic (e.g. ***Synonym***, ***Hypernym***, ***Hyponymn***) and syntagmatic relations (e.g. ***TypicalObject***, ***TypicalSubject***) contained in dictionary definitions and example sentences in a consistent fashion, thereby squeezing every potential piece of useful information it can out of the MRDs processed during its creation. A similar representational framework, this time used for information extracted from a children's dictionary, is described in Barrière & Popovich (1996).

The full inversion of complete semrel structures in our work is unique in all documented MRD research. Also distinctive is the richness of the semrel structures themselves. This massive LKB network invalidates the criticism leveled against MRD methods by Yarowsky (1992) and Ide and Veronis (1993) that LKBs created from MRDs provide spotty coverage of a language at best.

4.3 Semantic Relation Paths

Within MindNet, it is possible to trace chains of relationships among word senses. These relationships, consisting of one or more semantic relations connected together, constitute *semantic relation paths* (or *semrel paths*).

The semrel paths between "car" and "people", or "person", which were generated from definitions of "hitchhike", "car", "motorist", and "chauffeur", are shown in Figure 2. Paths may be as simple as a single semantic relation or they may consist of multiple semantic relations, as illustrated in these examples. The directionality of the individual semantic relations in a semrel path does not imply a directionality for the entire path. The path is simply a bi-directional linkage between two word senses, and the directionality of the individual semantic relations only pertains to the meaning of the relationships represented by those semantic relations. Hence, one may read the third semrel path below starting at either end of the path: "***a car is driven by a motorist, who is a person***", or "***a motorist is a person who drives cars***". Each reading is a paraphrase of the other, and both indicate the same essential relationship between "car" and "person".

hitchhike:	***car***—Possr→***people***
car:	***car***—Purp→***carry***—Tobj→***people***
motorist:	***car***←Tobj—***drive***—Tsub→***motorist***—Hyp→***person***
chauffeur:	***car***←Tobj—***drive***←Purp—***employ***—Tobj→***chauffeur***—Hyp→***person***

*Figure 2: Semrel paths from "**car**" to "**people**" or to "**person**", taken from inverted semrel structures*

Semrel paths, as described thus far, are taken only from single semrel structures generated from individual definitions. An *extended semrel path* (figure 3) is a path created from sub-paths in two different inverted semrel structures. For example, "car" and "truck" are not related directly by a semantic relation or by a semrel path from any single semrel structure. However, if one allows the joining of the semantic relations (which may also be considered as single relation sub-paths) in the first two columns of the table below, each from a different semrel structure, at the word "vehicle", the semrel paths in the third column result. Extended semrel paths exhibit a tradeoff between length and accuracy in that paths consisting of several semantic relations can be generated (where the sub-paths themselves consist of multiple semantic relations), and yet there is only one junction point where a higher possibility of inaccuracy exists (usually because of a potential word sense mismatch). As long as they are constrained adequately, extended semrel paths have proven invaluable in determining the relationship

1st semrel sub-path	2nd semrel sub-path	Extended semrel path
car—Hyp→***vehicle***	***vehicle***←Hyp—***truck***	***car***—Hyp→***vehicle***←Hyp—***truck***
car—Purp→***carry***	***carry***←Purp—***truck***	***car***—Purp→***carry***←Purp—***truck***

Figure 3: Creation of extended semrel paths

between certain words in the LKB that would not otherwise be connected.

A number of constraints have been implemented that prevent the formation of certain extended semrel paths, including blocking paths if they contain certain types of redundant cycles, or if the word which allows the subpaths to be joined is extremely high in frequency (e.g. "thing", "make", "give"). Extended semrel paths can also be blocked if the two subpaths would join through different senses of the same word.

4.4 Assigning Weights to Semantic Relation Paths

In querying the information contained in the LKB, to obtain all the semrel paths between two words for example, it is highly desirable to rank the results of a query based on potential usefulness. This is true not only for the needs of similarity determination (Richardson 1996) and matching for lexical and structural disambiguation, but also for the sake of efficient processing. To this end, a methodology for assigning weights to semrel paths, based on weights assigned to individual semantic relations (hereafter shortened to "semrels") in the paths, has been developed. These weights are meant to be indicators of the usefulness of the information contained in the semrel paths, and as such, they attempt to measure both the accuracy and the relevance of that information. Weighting schemes with similar goals are found in work by Braden-Harder (1992), Bookman (1994), and Sumita & Iida (1991).

The use of automatic, rather than manual, methods to create the LKB make a weighting scheme particularly important. In certain cases, because of semrel structure inversion, there may be dozens, hundreds, or even thousands of semrel path connections between specific words. While the first goal of this work was to create a very richly connected LKB in order to achieve as wide a coverage of the language as possible, the increased numbers of connections for more common words actually began to interfere with efficient and useful processing.

In response to this problem, Richardson (1996) uses statistical measures from the field of information retrieval to assign weights to individual semrels within the network. Determining how salient a semrel is involves balancing frequency and uniqueness: it must be frequent enough to be salient, yet not so frequent that it loses that salience; it must also be distinctive, yet not so distinctive that it rarely occurs. This weighting scheme has proven quite effective in determining how salient a particular semrel path is. In addition, a similarity metric (Richardson 1996) which exploits these weighted sets of paths yields results which correlate well with human judgements of conceptual similarity. This metric is the result of training on a thesaurus, collecting information about which patterns of semrel paths consistently link words which are known to be similar in meaning.

Once weights have been assigned to all the semrel paths contained in inverted semrel structures in the LKB, querying the LKB for semrel paths between two words returns a ranked list of paths that indicate the most salient connections between the query words. For example, some of the top semrel paths between "drive" and "car", together with their path weights, are shown below.

semrel path	***path weight***
drive –Typical_Object-> car	2.7164e-006
drive -Typical_Object-> prowl_car -Hypernym-> car	5.1286e-007
drive -Typical_Object-> panda_car -Hypernym-> car	3.0761e-009
drive -Typ_Subj->motorist -Typ_Subj-> own -Typ_Obj-> car	3.8815e-013
drive <-Hypernym- motor -Purpose-> travel -Means-> car	9.749e -014
drive <-Hypernym- run -Typical_Object-> car	1.9953e-033

5. Using MindNet

MindNet is very different in character from taxonomic models of lexical organization (Copestake 1990, Klavans, et al. 1990, Vossen 1991, Bruce and Guthrie 1992), which attempt to characterize lexical relationships strictly in terms of class inclusion. Its structure is also very different from WordNet's, which is organized in terms of relationships between opposing pairs of (synonymous sets of) lexical items. MindNet is instead organized as a set of interlinked "constellations" of complex predications, each extracted from a particular definition or example sentence. The result is that even quite that long chains of lexical links in MindNet can display a high degree of complexity and internal coherence that greatly increases their information value.

MindNet's content and navigation tools are currently used in several ways to serve our NLP engine. The combination of weighted, associative semrel paths and a similarity metric

provides sufficient inferencing power for tasks such as word-sense disambiguation, semantically-driven prepositional phrase attachment, and the tracking of lexical cohesion for the purposes of discourse understanding.

Such tasks are intrinsically tied up with the language itself, of course, and so perhaps it is not surprising that a prose description of the English lexicon serves up just the sort of information which is relevant for them. The larger question of whether MindNet will ultimately provide the type of information needed more complex forms of reasoning (about e.g. causality) remains open. In practice, there appears to be virtually no empirical basis for separating facts about word meanings from facts about the world and how entities interact in it; a dictionary's description of lexical semantics is nothing more than a partial description of human world knowledge.

6. Conclusions

This paper has described the automated procedures we use to map between the NL constructions encountered in dictionaries and structured semantic representations. The resulting KB refutes the common wisdom in the field that NL prose is so complex that even the relatively restricted domain of dictionary language cannot be profitably mined for information using automated methods. What distinguishes our work from other efforts aimed at producing information from MRDs is that processing a dictionary definition or example sentence within our system yields essentially the same sort of semantic representation that would be produced for a free text sentence. In processing an MRD, then, our system is carrying out exactly the same morphological, syntactic and semantic processing that it would if it were processing Wall Street Journal text. The LKB that is derived from LDOCE and AHD3 can thus to a great extent be regarded as a byproduct of this ordinary system processing.

This simple shift in perspective--KB as a byproduct of general processing mechanisms versus KB as the central goal of processing-- has great implications for the feasibility of the task. If we were primarily interested in creating an LKB from one or two dictionaries, hand-coding might be the simplest approach. The amount of effort required to automatically extract even simple *IsA* hierarchies from dictionaries is considerable, and it is not clear that this limited type of information is even particularly useful once extracted. However, given an NLP system whose goal is to map arbitrary strings of text into fully-specified semantic representations, dictionaries can yield up a vast amount of extremely useful information. The work of extracting this information automatically is one which the system must be capable of performing anyway, so any extra effort needed to handle morphological, syntactic or semantic processing problems within the dictionary is useful for the project's broader goals. Each modification brings us closer to our aim of mapping easily between free text and knowledge representation.

Acknowledgments

The authors would like to thank Kathy Hunt and Lucy Vanderwende, Karen Jensen, and the other members of the Microsoft Natural Language Processing Group for their valuable assistance in writing this paper.

References

Amsler, R. A. 1980. *The structure of the Merriam Webster Pocket Dictionary*. Ph. D. dissertation, University of Texas at Austin.

Barrière, Caroline and Fred Popovich. 1996. "Concept Clustering and knowledge integration from a children's dictionary," Proceedings of COLING-96, pp. 65-70.

Bookman, L. 1994. *Trajectories through Knowledge Space: A Dynamic Framework for machine Comprehension.* Kluwer Academic Publishers, Boston, MA.

Braden-Harder, L. 1992. "Sense Disambiguation using an Online Dictionary." In *Natural Language Processing: the PLNLP Approach,* ed. K. Jensen, G. Heidorn, and S. Richardson. Boston: Kluwer Academic Publishers.

Bruce, R. and L. Guthrie. 1992. "Genus disambiguation: a study in weighted preference" in *Proceedings of COLING92*, pp.1187-1191.

Byrd, R.J., N. Calzolari, M.S. Chodorow, J.L. Klavans, M.S. Neff, and O.A. Rizk. 1987. "Tools and Methods for Computational Lexicology" in *Computational Linguistics* 13.3-4. pp. 219-40.

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

Dolan W., L. Vanderwende, and S. Richardson. 1993. "Automatically deriving structured knowledge bases from on-line dictionaries" in *Proceedings of the First Conference of the Pacific Association for Computational Linguistics*, April 21-24, Simon Fraser University, Vancouver, Canada.

Fillmore, C. J. 1968. "The case for case". In Bach, E. & Harms, R. (eds.) *Universals in Linguistic Theory.* New York: Holt, Rhinehart & Winston.

Ide, N. and J. Veronis. 1993. "Extracting knowledge bases from machine-readable dictionaries: have we wasted our time?" KB & KS, December, 1993, Tokyo.

Jensen, K. 1993. "PEG: The PLNLP English Grammar" in *Natural Language Processing: The PLNLP Approach*, ed. Jensen, K., G.E. Heidorn, and S.D. Richardson, Kluwer Academic Publishers, Boston, MA. pp 29-45.

Jensen, K., and J.-L. Binot. 1987. "Disambiguating prepositional phrase attachments by using on-line dictionary definitions" in *Computational Linguistics* 13.3-4: 251-60.

Klavans, J., M. Chodorow, and N. Wacholder. 1990. "From Dictionary to Knowledge Base via Taxonomy" in *Electronic Text Research*, University of Waterloo, Centre for the New OED and Text Research, Waterloo, Canada.

Luhn, H. 1958. "The Automatic Creation of Literature Abstracts" in *IBM Journal of Research and Development* 2.2. pp. 159-165.

Markowitz, J., T. Ahlswede and M. Evens. 1986. "Semantically Significant Patterns in Dictionary Definitions" in *Proceedings of the 24th Annual Meeting of the Association for Computational Linguistics*, June 10-13 1986, pp. 112-119.

Miller, G.A. 1990. "WordNet: an online lexical database". *International Journal of Lexicography*, 3: 235-312.

Montemagni, S., and L. Vanderwende. 1992. "Structural Patterns vs. string patterns for extracting semantic information from dictionaries" in *Proceedings of COLING92*, pp.546-552.

Richardson, S. 1996. Determining similarity and inferring relations in a lexical knowledge base. Ph.D. dissertation, City University of New York.

Salton, G., and M. McGill. 1983. *Introduction to Modern Information Retrieval.* McGraw-Hill Book Company, New York, NY.

Sowa, J. 1984. *Conceptual Stuctures: information Processing in mind and machine*. Addison-Wesley.

Sumita, E. and H. Iida. 1991. "Experiments and Prospects of Example-Based Machine Translation." *Proceedings of the 29th Annual Meeting of the Association of Computational Linguistics,* pp. 185-192.

Vanderwende, L. 1996. *The Analysis of Noun Sequences using Semantic Information Extracted from On-Line Dictionaries.* Ph.D. dissertation, Georgetown University, Washington, D.C.

Veronis, J., and N.M. Ide. 1990. "Word Sense Disambiguation with Very Large Neural Networks Extracted from Machine Readable Dictionaries" in *Proceedings of COLING90*, pp 289-295.

Vossen, P. 1991. "Converting data from a lexical database to a knowledge base." ESPRIT BRA-3030 ACQUILEX WP NO.027.

Wilks, Y., D. Fass, C. Guo, J. McDonald, T. Plate, and B. Slator. 1989. "A Tractable Machine Dictionary as a Resource for Computational Semantics" in Boguraev & Briscoe, eds., *Computational Lexicography for Natural Language Processing*, Longman, London, pp. 193-228.

Yarowsky, D. 1992. "Word-sense disambiguation using statistical models of Roget's categories trained on large corpora" in *Proceedings of COLING92*, pp. 454-460.

Zipf, G. 1949. *Human Behavior and the Principle of Least Effort.* Addison-Wesley Press, Reading, MA.