MindNet: acquiring and structuring semantic information from text

Stephen D. Richardson, William B. Dolan, Lucy Vanderwende
Microsoft Research
One Microsoft Way
Redmond, WA 98052
U.S.A.

Abstract

As a lexical knowledge base constructed automatically from the definitions and example sentences in two machine-readable dictionaries (MRDs), MindNet embodies several features that distinguish it from prior work with MRDs. It is, however, more than this static resource alone. MindNet represents a general methodology for acquiring, structuring, accessing, and exploiting semantic information from natural language text. This paper provides an overview of the distinguishing characteristics of MindNet, the steps involved in its creation, and its extension beyond dictionary text.

1 Introduction

In this paper, we provide a description of the salient characteristics and functionality of MindNet as it exists today, together with comparisons to related work. We conclude with a discussion on extending the MindNet methodology to the processing of other corpora (specifically, to the text of the Microsoft Encarta® 98 Encyclopedia) and on future plans for MindNet. For additional details and background on the creation and use of MindNet, readers are referred to Richardson (1997), Vanderwende (1996), and Dolan et al. (1993).

2 Full automation

MindNet is produced by a fully automatic process, based on the use of a broad-coverage NL parser. A fresh version of MindNet is built regularly as part of a normal regression process. Problems introduced by daily changes to the underlying system or parsing grammar are quickly identified and fixed.

Although there has been much research on the use of automatic methods for extracting information from dictionary definitions (e.g., Vossen 1995, Wilks et al. 1996), hand-coded knowledge bases, e.g. WordNet (Miller et al. 1990), continue to be the focus of ongoing research. The Euro WordNet project (Vossen 1996), although continuing in the WordNet tradition, includes a focus on semi-automated procedures for acquiring lexical content.

Outside the realm of NLP, we believe that automatic procedures such as MindNet's provide the only credible prospect for acquiring world knowledge on the scale needed to support common-sense reasoning. At the same time, we acknowledge the potential need for the hand vetting of such information to insure accuracy and consistency in production level systems.

3 Broad-coverage parsing

The extraction of the semantic information contained in MindNet exploits the very same broad-coverage parser used in the Microsoft Word 97 grammar checker. This parser produces syntactic parse trees and deeper logical forms, to which rules are applied that generate corresponding structures of semantic relations. The parser has ***not*** been specially tuned to process dictionary definitions. All enhancements to the parser are geared to handle the immense variety of general text, of which dictionary definitions are simply a modest subset.

There have been many other attempts to process dictionary definitions using heuristic pattern matching (e.g., Chodorow et al. 1985), specially constructed definition parsers (e.g., Wilks et al. 1996, Vossen 1995), and even general coverage syntactic parsers (e.g., Briscoe and Carroll 1993). However, none of these has succeeded in producing the breadth of semantic relations across entire dictionaries that has been produced for MindNet.

Vanderwende (1996) describes in detail the methodology used in the extraction of the semantic relations comprising MindNet. A truly broad-coverage parser is an essential component of this process, and it is the basis for extending it to other sources of information such as encyclopedias and text corpora.

4 Labeled, semantic relations

The different types of labeled, semantic relations extracted by parsing for inclusion in MindNet are given in the table below:

Attribute	Goal	Possessor
Cause	Hypernym	Purpose
Co-Agent	Location	Size
Color	Manner	Source
Deep_Object	Material	Subclass
Deep_Subject	Means	Synonym
Domain	Modifier	Time
Equivalent	Part	User

Table 1. Current set of semantic relation types in MindNet

These relation types may be contrasted with simple co-occurrence statistics used to create network structures from dictionaries by researchers including Veronis and Ide (1990), Kozima and Furugori (1993), and Wilks et al. (1996). Labeled relations, while more difficult to obtain, provide greater power for resolving both structural attachment and word sense ambiguities.

While many researchers have acknowledged the utility of labeled relations, they have been at times either unable (e.g., for lack of a sufficiently powerful parser) or unwilling (e.g., focused on purely statistical methods) to make the effort to obtain them. This deficiency limits the characterization of word pairs such as ***river/bank*** (Wilks et al. 1996) and ***write/pen*** (Veronis and Ide 1990) to simple relatedness, whereas the labeled relations of MindNet specify precisely the relations ***river***—**Part**→***bank*** and ***write***—**Means**→***pen***.

5 Semantic relation structures

The automatic extraction of semantic relations (or *semrels*) from a definition or example sentence for MindNet produces a hierarchical structure of these relations, representing the entire definition or sentence from which they came. Such structures are stored in their entirety in MindNet and provide crucial context for some of the procedures described in later sections of this paper. The semrel structure for a definition of ***car*** is given in the figure below.

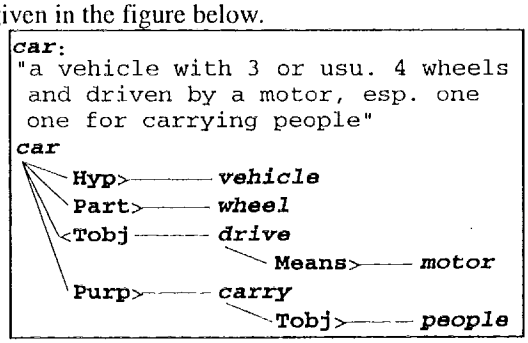

Figure 1. Semrel structure for a definition of ***car****.*

Early dictionary-based work focused on the extraction of paradigmatic relations, in particular **Hypernym** relations (e.g., ***car***—**Hypernym**→***vehicle***). Almost exclusively, these relations, as well as other syntagmatic ones, have continued to take the form of relational triples (see Wilks et al. 1996). The larger contexts from which these relations have been taken have generally not been retained. For labeled relations, only a few researchers (recently, Barrière and Popowich 1996), have appeared to be interested in entire *semantic structures* extracted from dictionary definitions, though they have not reported extracting a significant number of them.

6 Full inversion of structures

After semrel structures are created, they are fully inverted and propagated throughout the entire MindNet database, being linked to every word that appears in them. Such an inverted structure, produced from a definition for ***motorist*** and linked to the entry for ***car*** (appearing as the root of the inverted structure), is shown in the figure below:

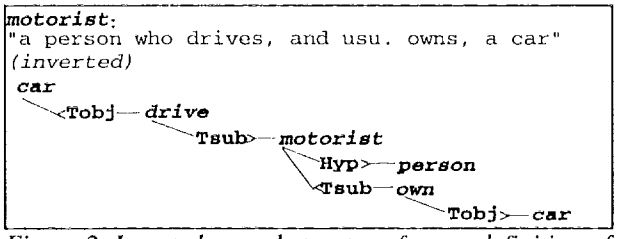

Figure 2. Inverted semrel structure from a definition of ***motorist***

Researchers who produced spreading activation networks from MRDs, including Veronis and Ide (1990) and Kozima and Furugori (1993), typically only implemented forward links (from headwords to their definition words) in those networks. Words were not related backward to any of the headwords whose definitions mentioned them, and words co-occurring in the same definition were not related directly. In the fully inverted structures stored in MindNet, however, all words are cross-linked, no matter where they appear.

The massive network of inverted semrel structures contained in MindNet invalidates the criticism leveled against dictionary-based methods by Yarowsky (1992) and Ide and Veronis (1993) that LKBs created from MRDs provide spotty coverage of a language at best. Experiments described elsewhere (Richardson 1997) demonstrate the comprehensive coverage of the information contained in MindNet.

Some statistics indicating the size (rounded to the nearest thousand) of the current version of MindNet and the processing time required to create it are provided in the table below. The definitions and example sentences are from the *Longman Dictionary of Contemporary English* (LDOCE) and the *American Heritage Dictionary, 3rd Edition* (AHD3).

Dictionaries used	LDOCE & AHD 3
Time to create (on a P2/266)	7 hours
Headwords	159,000
Definitions (N, V, ADJ)	191,000
Example sentences (N, V, ADJ)	58,000
Unique semantic relations	713,000
Inverted structures	1,047,000
Linked headwords	91,000

Table 2. Statistics on the current version of MindNet

7 Weighted paths

Inverted semrel structures facilitate the access to direct and indirect relationships between the root word of each structure, which is the headword for the MindNet entry containing it, and every other word contained in the structures. These relationships, consisting of one or more semantic relations connected together, constitute *semrel paths* between two words. For example, the semrel path between ***car*** and ***person*** in Figure 2 above is:

***car*←Tobj—*drive*—Tsub→*motorist*—Hyp→*person*.**

An *extended semrel path* 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 ***car*—Hyp→*vehicle*** and ***vehicle*←Hyp—*truck***, each from a different semrel structure, at the word ***vehicle***, the semrel path ***car*—Hyp→*vehicle*←Hyp—*truck*** results. Adequately constrained, extended semrel paths have proven invaluable in determining the relationship between words in MindNet that would not otherwise be connected.

Semrel paths are automatically assigned weights that reflect their salience. The weights in MindNet are based on the computation of *averaged vertex probability*, which gives preference to semantic relations occurring with middle frequency, and are described in detail in Richardson (1997). Weighting schemes with similar goals are found in work by Braden-Harder (1993) and Bookman (1994).

8 Similarity and inference

Many researchers, both in the dictionary- and corpus-based camps, have worked extensively on developing methods to identify similarity between words, since similarity determination is crucial to many word sense disambiguation and parameter-smoothing/inference procedures. However, some researchers have failed to distinguish between *substitutional* similarity and general relatedness. The similarity procedure of MindNet focuses on measuring substitutional similarity, but a function is also provided for producing clusters of generally related words.

Two general strategies have been described in the literature for identifying substitutional similarity. One is based on identifying direct, paradigmatic relations between the words, such as **Hypernym** or **Synonym**. For example, paradigmatic relations in WordNet have been used by many to determine similarity, including Li et al. (1995) and Agirre and Rigau (1996). The other strategy is based on identifying syntagmatic relations with other words that similar words have in common. Syntagmatic strategies for determining similarity have often been based on statistical analyses of large corpora that yield clusters of words occurring in similar bigram and trigram contexts (e.g., Brown et al. 1992, Yarowsky 1992), as well as in similar predicate-argument structure contexts (e.g., Grishman and Sterling 1994).

There have been a number of attempts to combine paradigmatic and syntagmatic similarity strategies (e.g., Hearst and Grefenstette 1992, Resnik 1995). However, none of these has completely integrated both syntagmatic and paradigmatic information into a single repository, as is the case with MindNet.

The MindNet similarity procedure is based on the top-ranked (by weight) semrel paths between words. For example, some of the top semrel paths in MindNet between ***pen*** and ***pencil***, are shown below:

pen*←Means—*draw*—Means→*pencil
pen*←Means—*write*—Means→*pencil
pen*—Hyp→*instrument*←Hyp—*pencil
pen*—Hyp→*write*—Means→*pencil
pen*←Means—*write*←Hyp—*pencil

*Table 3. Highly weighted semrel paths between **pen** and **pencil***

In the above example, a pattern of semrel symmetry clearly emerges in many of the paths. This observation of symmetry led to the hypothesis that similar words are typically connected in MindNet by semrel paths that frequently exhibit certain patterns of relations (exclusive of the words they actually connect), many patterns being symmetrical, but others not.

Several experiments were performed in which word pairs from a thesaurus and an anti-thesaurus (the latter containing dissimilar words) were used in a training phase to identify semrel path patterns that indicate similarity. These path patterns were then used in a testing phase to determine the substitutional similarity or dissimilarity of unseen word pairs (algorithms are described in Richardson 1997). The results, summarized in the table below, demonstrate the strength of this integrated approach, which uniquely exploits both the paradigmatic and the syntagmatic relations in MindNet.

Training: over 100,000 word pairs from a thesaurus and anti-thesaurus produced 285,000 semrel paths containing approx. 13,500 unique path patterns.		
Testing: over 100,000 (different) word pairs from a thesaurus and anti-thesaurus were evaluated using the path patterns.	Similar correct	Dissimilar correct
	84%	82%
Human benchmark: random sample of 200 similar and dissimilar word pairs were evaluated by 5 humans and by MindNet:	Similar correct	Dissimilar correct
Humans:	83%	93%
MindNet:	82%	80%

Table 4. Results of similarity experiment

This powerful similarity procedure may also be used to extend the coverage of the relations in MindNet. Equivalent to the use of similarity determination in corpus-based approaches to infer absent n-grams or triples (e.g., Dagan et al. 1994, Grishman and Sterling 1994), an inference procedure has been developed which allows semantic relations not presently in MindNet to be inferred from those that are. It also exploits the top-ranked paths between the words in the relation to be inferred. For example, if the relation ***watch*—Means→*telescope*** were not in MindNet, it could be inferred by first finding the semrel paths between ***watch*** and ***telescope***, examining those paths to see if another word appears in a **Means** relation with ***telescope***, and then checking the similarity between that word and ***watch***. As it turns out, the word ***observe*** satisfies these conditions in the path:

watch*—Hyp→*observe*—Means→*telescope

and therefore, it may be inferred that one can ***watch*** by **Means** of a ***telescope***. The seamless integration of the inference and similarity procedures, both utilizing the weighted, extended paths derived from inverted semrel structures in MindNet, is a unique strength of this approach.

9 Disambiguating MindNet

An additional level of processing during the creation of MindNet seeks to provide sense identifiers on the words of semrel structures. Typically, word sense disambiguation (WSD) occurs during the parsing of definitions and example sentences, following the construction of logical forms (see Braden-Harder, 1993). Detailed information from the parse, both morphological and syntactic, sharply reduces the range of senses that can be plausibly assigned to each word. Other aspects of dictionary structure are also exploited, including domain information associated with particular senses (e.g., ***Baseball***).

In processing normal input text outside of the context of MindNet creation, WSD relies crucially on information from MindNet about how word senses are linked to one another. To help mitigate this bootstrapping problem during the initial construction of MindNet, we have experimented with a two-pass approach to WSD.

During a first pass, a version of MindNet that does not include WSD is constructed. The result is a semantic network that nonetheless contains a great deal of "ambient" information about sense assignments. For instance, processing the definition ***spin 101: (of a spider or silkworm) to produce thread...*** yields a semrel structure in which the sense node ***spin101*** is linked by a **Deep_Subject** relation to the undisambiguated form ***spider***. On the subsequent pass, this information can be exploited by WSD in assigning sense 101 to the word ***spin*** in unrelated definitions: ***wolf_spider 100: any of various spiders...that...do not spin webs***. This kind of bootstrapping reflects the broader nature of our approach, as discussed in the next section: a fully and accurately disambiguated MindNet allows us to bootstrap senses onto words encountered in free text outside the dictionary domain.

10 MindNet as a methodology

The creation of MindNet was never intended to be an end unto itself. Instead, our emphasis has been on building a broad-coverage NLP understanding system. We consider the methodology for creating MindNet to consist of a set of general tools for acquiring, structuring, accessing, and exploiting semantic information from NL text.

Our techniques for building MindNet are largely rule-based. However we arrive at these representations, though, the overall structure of MindNet can be regarded as crucially dependent on statistics. We have much more in common with traditional corpus-based approaches than a first glance might suggest. An advantage we have over these approaches, however, is the rich structure imposed by the parse, logical form, and word sense disambiguation components of our system. The statistics we use in the context of MindNet allow richer metrics because the data themselves are richer.

Our first foray into the realm of processing free text with our methods has already been accomplished; Table 2 showed that some 58,000 example sentences from LDOCE and AHD3 were processed in the creation of our current MindNet. To put our hypothesis to a much more rigorous test, we have recently embarked on the assimilation of the entire text of the Microsoft Encarta® 98 Encyclopedia. While this has presented several new challenges in terms of volume alone, we have nevertheless successfully completed a first pass and have produced and added semrel structures from the Encarta® 98 text to MindNet. Statistics on that pass are given below:

Processing time (on a P2/266)	34 hours
Sentences	497,000
Words	10,900,000
Average words/sentence	22
New headwords in MindNet	220,000
New inverted structures in MindNet	5,600,000

Table 5. Statistics for Microsoft Encarta® 98

Besides our venture into additional English data, we fully intend to apply the same methodologies to text in other languages as well. We are currently developing NLP systems for 3 European and 3 Asian languages: French, German, and Spanish; Chinese, Japanese, and Korean. The syntactic parsers for some of these languages are already quite advanced and have been demonstrated publicly. As the systems for these languages mature, we will create corresponding MindNets, beginning, as we did in English, with the processing of machine-readable reference materials and then adding information gleaned from corpora.

11 References:

Agirre, E., and G. Rigau. 1996. Word sense disambiguation using conceptual density. In *Proceedings of COLING96*, 16-22.

Barrière, C., and F. Popowich. 1996. Concept clustering and knowledge integration from a children's dictionary. In *Proceedings of COLING96*, 65-70.

Bookman, L. 1994. *Trajectories through knowledge space: A dynamic framework for machine comprehension.* Boston, MA: Kluwer Academic Publishers.

Braden-Harder, L. 1993. Sense disambiguation using an online dictionary. In *Natural language processing: The PLNLP approach*, ed. K. Jensen, G. Heidorn, and S. Richardson, 247-261. Boston, MA: Kluwer Academic Publishers.

Briscoe, T., and J. Carroll. Generalized probabilistic LR parsing of natural language (corpora) with unification-based grammars. *Computational Linguistics* 19, no. 1:25-59.

Brown, P., V. Della Pietra, P. deSouza, J. Lai, and R. Mercer. 1992. Class-based n-gram models of natural language. *Computational Linguistics* 18, no. 4:467-479.

Chodorow, M., R. Byrd, and G. Heidorn. 1985. Extracting semantic hierarchies from a large on-line dictionary. In *Proceedings of the 23rd Annual Meeting of the ACL*, 299-304.

Dagan, I., F. Pereira, and L. Lee. 1994. Similarity-based estimation of word cooccurrence probabilities. In *Proceedings of the 32nd Annual Meeting of the ACL*, 272-278.

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* (Vancouver, Canada), 5-14.

Grishman, R., and J. Sterling. 1994. Generalizing automatically generated selectional patterns. In *Proceedings of COLING94*, 742-747.

Hearst, M., and G. Grefenstette. 1992. Refining automatically-discovered lexical relations: Combining weak techniques for stronger results. In *Statistically-Based Natural Language Programming Techniques, Papers from the 1992 AAAI Workshop* (Menlo Park, CA), 64-72.

Ide, N., and J. Veronis. 1993. Extracting knowledge bases from machine-readable dictionaries: Have we wasted our time? In *Proceedings of KB&KS '93* (Tokyo), 257-266.

Kozima, H., and T. Furugori. 1993. Similarity between words computed by spreading activation on an English dictionary. In *Proceedings of the 6th Conference of the European Chapter of the ACL*, 232-239.

Li, X., S. Szpakowicz, and S. Matwin. 1995. A WordNet-based algorithm for word sense disambiguation. In *Proceedings of IJCAI'95*, 1368-1374.

Miller, G., R. Beckwith, C. Fellbaum, D. Gross, and K. Miller. 1990. Introduction to WordNet: an on-line lexical database. In *International Journal of Lexicography* 3, no. 4:235-244.

Resnik, P. 1995. Disambiguating noun groupings with respect to WordNet senses. In *Proceedings of the Third Workshop on Very Large Corpora*, 54-68.

Richardson, S. 1997. Determining similarity and inferring relations in a lexical knowledge base. PhD. dissertation, City University of New York.

Vanderwende, L. 1996. The analysis of noun sequences using semantic information extracted from on-line dictionaries. Ph.D. dissertation, Georgetown University, Washington, DC.

Veronis, J., and N. Ide. 1990. Word sense disambiguation with very large neural networks extracted from machine readable dictionaries. In *Proceedings of COLING90*, 289-295.

Vossen, P. 1995. Grammatical and conceptual individuation in the lexicon. PhD. diss. University of Amsterdam.

Vossen, P. 1996: Right or Wrong. Combining lexical resources in the EuroWordNet project. In: M. Gellerstam, J. Jarborg, S. Malmgren, K. Noren, L. Rogstrom, C.R. Papmehl, Proceedings of Euralex-96, Goetheborg, 1996, 715-728

Wilks, Y., B. Slator, and L. Guthrie. 1996. *Electric words: Dictionaries, computers, and meanings.* Cambridge, MA: The MIT Press.

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