MindNet: an automatically-created lexical resource

Lucy Vanderwende, Gary Kacmarcik, Hisami Suzuki, Arul Menezes
Microsoft Research
Redmond, WA 98052, USA
{lucyv, garykac, hisamis, arulm}@microsoft.com

Abstract

We will demonstrate MindNet, a lexical resource built automatically by processing text. We will present two forms of MindNet: as a static lexical resource, and, as a toolkit which allows MindNets to be built from arbitrary text. We will also introduce a web-based interface to MindNet lexicons (MNEX) that is intended to make the data contained within MindNets more accessible for exploration. Both English and Japanese MindNets will be shown and will be made available, through MNEX, for research purposes.

1 MindNet

A MindNet is a collection of semantic relations that is automatically extracted from text data using a broad coverage parser. Previous publications on MindNet (Suzuki et al., 2005, Richardson et al., 1998, Vanderwende 1995) have focused on the effort required to build a MindNet from the data contained in Japanese and English lexicons.

Semantic Relations

The semantic relations that are stored in MindNet are directed, labeled relationships between two words; see Table 1:

Attributive	Manner	Source
Cause	Means	Synonym
Goal	Part	Time
Hypernym	Possessor	TypicalObject
Location	Result	TypicalSubject

Table 1: A sampling of the semantic relations stored in MindNet

These semantic relations are obtained from the Logical Form analysis of our broad coverage parser NLPwin (Heidorn, 2000). The Logical Form is a labeled dependency analysis with function words removed. We have not completed an evaluation of the quality of the extracted semantic relations. Anecdotally, however, the quality varies according to the relation type, with Hypernym and grammatical relations TypicalSubject and TypicalObj being reliable, while relations such as Part and Purpose are less reliable. By making MindNet available, we solicit feedback on the utility of these labeled relationships, especially in contrast to simple co-occurrence statistics and to the heavily used hypernymy and synonymy links. Furthermore, we solicit feedback on the level of accuracy which is tolerable for specific applications.

Semantic Relation Structures

We refer to the hierarchical collection of semantic relations (*semrels*) that are automatically extracted from a source sentence as a *semrel structure*. Each semrel structure contains all of the semrels extracted from a single source sentence. A semrel structure can be viewed from the perspective of each unique word that occurs in the structure; we call these *inverted structures*. They contain the same information as the original, but with a different word placed at the root of the structure. An example semrel structure for the definition of *swallow* is given in Figure 1a, and its inversion, from the perspective of *wing* is given in Figure 1b:

```
swallow              wing
 Hyp bird              PartOf bird
      Part wing                Attrib small
      Attrib small             HypOf swallow
```

Figure 1a and b: Figure 1a is the semrel structure for the definition of swallow1, Figure 1b the inversion on wing.

2 MNEX

MNEX (MindNet Explorer) is the web-based interface to MindNet that is designed to facilitate browsing MindNet structure and relations. MNEX displays paths based on the word or words that the

[1] *Swallow*: a small bird with wings (LDOCE). Definition abbreviated for purposes of exposition.

Proceedings of HLT/EMNLP 2005 Demonstration Abstracts, pages 8–9,
Vancouver, October 2005.

user enters. A path is a set of links that connect one word to another within either a single semrel structure or by combining fragments from multiple semrel structures. Paths are weighted for comparison (Richardson, 1997). Currently, either one or two words can be specified and we allow some restrictions to refine the path search. A user can restrict the intended part of speech of the words entered, and/or the user can restrict the paths to include only the specified relation. When two words are provided, the UI returns a list of the highest ranked paths between those two words. When only one word is given, then all paths from that word are ranked and displayed. Figure 2 shows the MNEX interface, and a query requesting all paths from the word *bird*, restricted to Noun part of speech, through the **Part** relation:

Figure 2: MNEX output for "bird (Noun) Part" query

3 Relation to other work

For English, WordNet is the most widely used knowledgebase. Aside from being English-only, this database was hand-coded and significant effort is required to create similar databases for different domains and languages. Projects like EuroWordNet address the monolingual aspect of WordNet, but these databases are still labor intensive to create. On the other hand, the quality of the information contained in a WordNet (Fellbaum et al., 1998) is very reliable, exactly because it was manually created. FrameNet (Baker et al., 1998) and OpenCyc are other valuable resources for English, also hand-created, that contain a rich set of relations between words and concepts. Their use is still being explored as they have been made available only recently. For Japanese, there are also concept dictionaries providing semantic relations, similarly hand-created, e.g., EDR and Nihongo Goi-taikei (NTT).

The demonstration of MindNet will highlight that this resource is automatically created, allowing domain lexical resources to be built quickly, albeit with lesser accuracy. We are confident that this is a trade-off worth making in many cases, and encourage experimentation in this area. MNEX allows the exploration of the rich set of relations through which paths connecting words are linked.

4 References

Baker, Collin F., Fillmore, Charles J., and Lowe, John B. (1998): The Berkeley FrameNet project. in Proceedings of the COLING-ACL, Montreal, Canada.

Fellbaum, C. (ed). 1998. WordNet: An Electronic Lexical Database. MIT Press.

Heidorn, G. 2000. Intelligent writing assistance. in R.Dale, H.Moisl and H.Somers (eds.), A Handbook of Natural Langauge Processing: Techniques and Applications for the Processing of Language as Text. New York: Marcel Dekker.

National Institute of Information and Communications Technology. 2001. EDR Electronic Dictionary Version 2.0 Technical Guide.

NTT Communications Science Laboratories. 1999. Goi-Taikei - A Japanese Lexicon. Iwanami Shoten.

OpenCyc. Available at: http://www.cyc.com/opencyc.

Richardson, S.D. 1997, Determining Similarity and Inferring Relations in a Lexical Knowledge Base. PhD. dissertation, City University of New York.

Richardson, S.D., W. B. Dolan, and L. Vanderwende. 1998. MindNet: Acquiring and Structuring Semantic Information from Text, In *Proceedings of ACL-COLING*. Montreal, pp. 1098-1102.

Suzuki, H., G. Kacmarcik, L. Vanderwende and A. Menezes. 2005. Mindnet and mnex. In Proceedings of the 11th Annual meeting of the Society of Natural Language Processing (in Japanese).

Vanderwende, L. 1995. Ambiguity in the acquisition of lexical information. In Proceedings of the AAAI 1995 Spring Symposium Series, symposium on representation and acquisition of lexical knowledge, 174-179.