Mindnet/mnex:意味関係データベースの自動構築と解析のためのツール

鈴木 久美　　Gary Kacmarcik　　Lucy Vanderwende　　Arul Menezes

Microsoft Research
One Microsoft Way, Redmond WA 98052 USA
{hisamis, garykac, lucyv, arulm}@microsoft.com

概要

Mindnet(マインドネット)は通常のテキストデータから、単語間の相互意味関係を直接的・自動的に抽出したデータベースであり、mnex(ネックス)はこれをさまざまな観点から表示・探索するためのウェブ・ツールである。通常のシソーラスの機能に加え、ひとつの単語の意味関係だけでなく、2単語間の意味関係を、その意味関係のタイプを指定して表示することができる。Mindnetの意味関係の抽出は、人手によることなく、構文解析による意味関係の同定と、その重みづけの2段階に分けて行われる。重みづけは、語と語を結びつける意味関係のパスに対して行われ、有用な意味関係のみを抽出することを目的としている。Mindnetは、辞書と百科事典をソースデータに用いて、まず英語で開発されたが、現在までに日本語でも辞書と百科事典テキストにもとづいたMindnetが構築されている。

1　はじめに

Mindnet(マインドネット)は、パーサを用いて解析したテキストデータから、単語間の相互意味関係を直接的・自動的に抽出したデータベースであり、mnex(ネックス)はこれをさまざまな観点から表示・探索するためのウェブ・ツールである。Mindnetは、辞書と百科事典をソースデータに用いて、まず英語で開発された[7]が、現在までに日本語でも辞書と百科事典にもとづいたMindnetが構築されている。本稿ではMindnetの構築のプロセスと、それに格納されているデータについて、日本語版Mindnetの構築から得た経験も交えて述べる。あわせて、新しく作られたウェブ・ツールであるmnex(ネックス)も、6節で紹介する。

意味関係のデータベースはこれまでにもいくつか提案されてきたが、Mindnetの一番の特徴は、自動的にテキストデータから構築されるということである。西欧言語では、WordNet[4]やEuroWordNet[12]などのデータベースがよく知られているが、これらは人手で構築されたデータベースである。日本語に関しても、意味関係のデータベースはいくつか存在し、さまざまな言語処理関連のタスクに使用されている。たとえば、EDR電子化辞書の概念体系辞書[2]や、NTTの日本語語彙体系[3]は、日本語の語彙を階層的に集めた大規模なデータベースである。また、用言に対しての格フレームを集めたデータベースも存在し[3]、これらのなかにはテキストから自動的に構築されたものもある[1]。これらのデータベースに対し、Mindnetでは上位・下位概念による階層的な体系(シソーラス)と、用言とその項の関係を記述する格フレームを別に扱わず、単一のデータベースに格納する。このことは語と語の関係をより幅広く記述することに役立っている。

2　テキスト解析

Mindnetの構築はまず使用するテキストを解析することから始まる。テキストの解析には、現在NLPWin parser[9,10]を使用している。辞書からMindnetを構築する場合は、辞書の定義文と例文を、百科事典からの場合には百科事典記事本文を解析の対象とする。これらのテキストを構文解析した結果がlogical formであるが、これは表層の構文関係から用言とその項構造を中心に抽出した構造であり、機能語などの要素はノードの素性やノード間関係のラベルに置き換えられている。logical formの例を図1に示す。logical formはMindnetの構築に特殊なものではなく、機械翻訳システム[8]のトランスファーなどにも使われている。

```
科挙は隋の文帝によって587年ごろにはじめられた。
はじめる1 (+Past +Pass)
 |_Dsub--文帝1 (+Sing +PrprN +Anim +Humn)
          |_の--隋1 (+Sing +PrprN +Cntry)
 |_Dobj--科挙1
 |_TmeAt-587_年1 (+Tme +Year)
          |_Mods---ごろ1
```

図1: logical formの例

Mindnet構築の過程では、logical formをさらにルールによって、最終的にデータベースに格納される意味関係構造(semantic relation structure)に変形する。このような変形の例としては、辞書の定義文を見出し語と上位・下位関係で結びつけること、定義文に使われている分類用語(科、属、種など)を上位・下位概念に置き換えること、などがある。

3　意味関係

この節ではMindnetに格納されているデータについて述べる。Mindnetデータの基礎単位は前述のテキスト解析を介して得られる2語間の関係であり、これを「意味関係」

(semantic relationあるいはsemrel)と呼ぶ(データ抽出の詳細は[11]を参照)。

3.1 関係のタイプ

Mindnetに格納されている意味関係には、すべて有向の関係タイプのラベルがついている。たとえば、「食べる→Tsub→さかな」はひとつの意味関係であり、「さかな」が「食べる」の主体(Tsub)であることを示している。これは「食べる→Tobj→さかな」とは別の意味関係を構成し、単なる共起情報とは区別される。現在使用されている意味関係のタイプの例を図2に示す。

Attributive (Attrib)	Part
Cause	Possessor (Possr)
Goal	Purpose (Purp)
Equivalent (Equiv)	Result
Hypernym (Hyp)	Source
Intensifier (Intnsifs)	Synonym (Syn)
Location (Locn)	Time
LogicalOperator(LOps)	TypicalIndirectObj (Tind)
Manner	TypicalObject (Tobj)
Means	TypicalSubject (Tsub)

図2: 現在使用されている意味関係タイプの例

関係のタイプにはどのような種類のものが必要かについては、これを使用するアプリケーションによって異なってくることが予想される。したがって図2も将来的に変化していくことが考えられるが、その際には意味関係構造の生成のルールの変更で対応することができる。

意味関係にラベルを付与することによって、品詞にとらわれず語間関係が一括して扱えるようになっている。たとえば、従来のシソーラスでは、名詞と動詞には別々の体系が与えられており、名詞と動詞の語間関係は格フレーム辞書など、さらに別の体系で記述されることが多かったが、Mindnetにはこのような区別は存在しない。また、品詞が同じ2語間の関係で上位・下位関係に当てはまらないもの、たとえば「机」と「鉛筆」など(机はものを書く場所であり、鉛筆はその道具である)も、後述するが意味関係をつなげで拡張することにより、カバーすることができる。

3.2 意味関係構造

意味関係は、テキスト解析を介して得られる2語間の関係であり、Mindnetデータの最小単位であるが、これに対し、あるテキスト(辞書の定義文や例文、百科事典の記事など)から得られた意味関係の全体を、「意味関係構造」(semrel structure)とよぶ。意味関係構造は、定義文なり例文なりの意味関係を全体として捉えたものであり、Mindnetに格納されているデータの中心をなすものである。たとえば、辞書における「漁」の定義文からは図3のような意味関係構造が得られる。

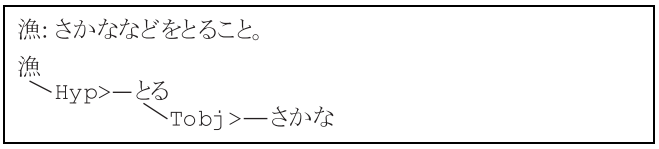

図3:「漁」の定義文の意味関係構造

この意味関係構造を得るには、まず構文解析により定義文をlogical formに解析する。つぎに、ルールを用いてこのlogical formの最上位のノードを辞書の見出し語「漁」の上位概念に設定する。logical formから意味関係構造を得る方法は英語版のMindnet構築においても同様であり、変形のルールも共有されているものが多い。

3.3 意味関係構造の倒置

Mindnetには、上述した意味関係構造だけでなく、倒置された意味関係構造も格納されている。この倒置された意味関係構造には、倒置以前のもとの意味関係構造に含まれている情報とまったく同じ情報が含まれているが、構造の最上位のノードが構造内の別の語で置き換えている点で異なる。たとえば、「漁師」の定義文から得られる倒置前の意味関係構造は図4のようなものである。

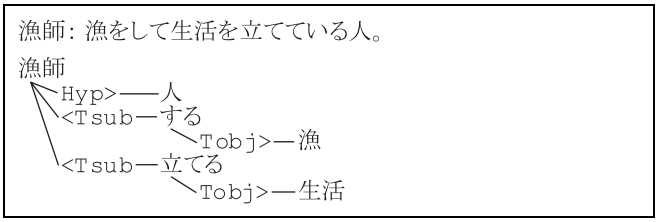

図4:「漁師」の定義文の意味関係構造

この意味関係構造は、見出し語である「漁師」だけでなく、ほかの語についての情報も含んでいる。たとえばこの構造を「漁」という語に着目して倒置すると、図5のような倒置された意味関係が得られる。

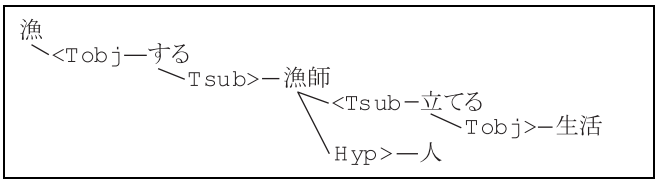

図5:「漁師」の定義文から倒置して得られる「漁」の意味関係構造

この倒置された意味関係構造には、「漁」の観点から見た意味関係が明示されている。倒置は意味関係構造に含まれる語のすべてに対して、Mindnet構築の際に自動的に行われる。倒置された意味関係構造をMindnetに格納することにより、特定の語の意味関係を網羅的に検索する際の効率化に大きな役割を果たしている。

表1に現在構築されているMindnetデータの詳細と、プ

Mindnet辞書版	英語版	日本語版
テキスト解析の所要時間	1:18	0:19
Mindnet構築の所要時間	0:06	0:02
辞書見出し語の数	159,000	35,288
辞書定義文の数(N, V, ADJ)	191,000	119,287
辞書例文の数(N, V, ADJ)	58,000	35,973
Mindnet見出し語の総数	120,023	107,626
意味関係の総数	2,231,839	1,079,730

Mindnet百科事典版	英語版	日本語版
テキスト解析の所要時間	5:39	3:23
Mindnet構築の所要時間	0:19	0:16
記事の総数	32,628	22,476
文の総数	609,673	516,153
Mindnet見出し語の総数	295,389	207,645
意味関係の総数	7,189,908	6,575,361

表1: Mindnetデータ

ロセスに要した時間を示す[1]。辞書は英語版がLongman Dictionary of Contemporary English (LDOCE)とThe American Heritage Dictionary, 3rd Edition (AHD)を、日本語版が岩波国語辞典第5版を使用している。「Mindnet構築の所要時間」には、意味関係構造の倒置と、次節で述べる意味関係の重みづけに要する時間が含まれている。

4 意味関係のパスと拡張パス

Mindnetに格納されている意味関係構造において、その構造内にある、ある語から別の語へのリンクの集合を指して、「意味関係のパス」(semrel path)という。たとえば、図4の意味関係構造で、「漁」から「人」へのパスは、

漁←Tobj←する→Tsub→漁師→Hyp→人

である。ここでは意味関係の非対称性を矢印を使ってしめしているが、パス自体には方向性は付与されていない。言い換えると、上述のパスは「漁は漁師がするもので、漁師は人だ」とも読めるし、「人は漁師の上位概念で、漁師は漁をする」と読んでもよい。

Mindnetの特徴のひとつとして、この意味関係のパスをつなぎ合わせて「意味関係の拡張パス」(extended semrel path)を生成することがあげられる。たとえば、日本語辞書版のMindnetでは、「漁師」と「さかな」の間には、単一の意味関係構造から得られるパスは存在しない。しかし、図4の「漁師」の定義文から得られるパス「漁師←Tsub←する→Tobj→漁」と、図3の「漁」の定義文から得られるパス「漁→Hyp→とる→Tobj→さかな」から、「漁」という語を結び目にすることによって、以下のような、「漁師は漁をし、漁とはさかなをとることだ」という拡張パスを得ることができる。

漁師←Tsub←する→Tobj→漁→Hyp→とる→Tobj→さかな

パスの拡張はMindnetを推論に使っていくためには欠かせない操作であるが、拡張しすぎた場合の弊害は非常に大きいため、細心の注意が必要である。現段階においては、品詞などの情報を使い、結び目の数もひとつまでという制限を設け、パスの拡張を制限しているが、これは今後さらに改良が求められているエリアである。

パスの拡張と並んで重要なMindnetの操作に、意味関係パスの重みづけがある。語と語を結んでいる意味関係のパスは、場合によっては拡張パスも含めて膨大な数になることが予想されるため、重要なパスから順番にスコアを付与してランク付けすることは必要不可欠なことである。現在使われている重みづけの計算には、Richardson[6]によるaveraged vertex probabilityが用いられている。この評価法は、語の頻度に注目し、頻度のきわめて高いものと低いものは有益な意味関係パスの抽出には向かないことから、中程度の頻度の語を重みづけに使用するものである。今後、この評価法をもとに、さらに重みづけのアルゴリズムを改良していく予定である。

5 日本語辞書からのMindnet構築

前述のとおり、Mindnetは当初、英語の辞書と百科事典をソーステキストとして開発が行われてきたが、日本語版を構築するにあたって、ほとんどの部分では既存のシステムをそのまま使用することができた。これはMindnetに格納されている意味構造の基礎となっているlogical formがかなりの部分で、英語と日本語に共通の構造を持っていたこと、辞書や百科事典の構造が両言語で似通っていたことなどが理由にあげられる。とはいえ、両言語で異なる部分もあり、そのなかでもっとも顕著なのが、日本語の表記に関するものである。

ひとつの語をさまざまな字種を組み合わせて表記できる日本語の解析では、まずテキスト解析の際に、表記をひらがな表記か漢字表記に正規化している[5]。つぎに辞書の定義と見出し語の処理では、通常見出し語はひらがなであり、語義ごとにそれに対応する漢字表記が付与されているので、Mindnet構築の際には、ひらがなと漢字表記の両方をMindnetの見出し語として、定義文と結びつけている。たとえば、図3の意味関係構造は、「りょう」と「漁」という二つの語の定義として使用される。これは辞書の見出し語という特殊な構造を解析するときにはにのみ行われ、見出し語以外のコンテキストでは行われない。したがって、Mindnetの検索語がひらがなの場合と漢字の場合では通常異なる結果が出てくるが、これは実際の表記の用法を反映している。たとえば、ひらがなの「さかな」でMindnetを検索すると、生物としての魚に加えて「酒のさかな」の語義に関するパスが検索されるが、漢字の「魚」の場合には、生物的な意味の魚に限定される。

辞書の見出し語一語がMindnetでは通常ひらがなと漢

[1] 使用したプロセッサは3.2GHz P4 (2GB RAM)である。

字の二語に相当することは、Mindnetに含まれる意味関係構造が重複することを意味しそれ自体は望ましいことではない。今後Mindnetのデータの効率化の面から見直すことのできる点は見直していきたい。

6 ユーザ・インターフェース

Mindnetデータをさまざまな観点から表示・探索するためのウェブ・ツールをmnex（ネックス=Mindnet Explorerの略）と呼んでいる。このインターフェースは英語・日本語に共通のものである。検索を始める画面(図6)で、パスを表示したい語のペア、あるいはある語の意味構造を単独で検索したい場合にはその語を入力する。その際、ドロップダウン・リストから品詞や関係タイプなど選んで、検索を制限することもできる。検索の結果は意味関係パスとして、重要と判断された順に表示される(図7)。デフォルトではパスを上から10ずつ表示する。ある意味関係パスがどのようなソーステキストに由来するのかを知りたいときには、パスの右横にあるリンク（辞書や百科事典の見出し語にあたる）をクリックすることで、ソーステキストとそこから得られた意味関係構造を参照することができるようになっている。

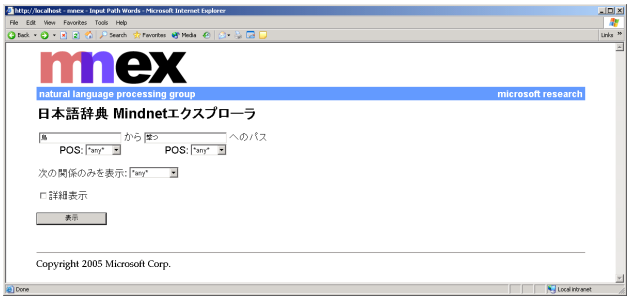

図6: mnex検索画面

図7: mnex検索結果画面

7 おわりに

以上日本語辞書版を中心に、Mindnetの構築・検索の方法とそれに格納されているデータについて述べた。日本語版Mindnetはまだ構築されてから日が浅く、今後改善すべき点が多く見られるが、このようなデータベースの構築に当たって何よりも重要なのは、その具体的な使用を念頭において、研究・開発を進めることである。Mindnetを辞書や百科事典以外のソーステキストから構築することも含めて、このようなアプリケーションからの観点が今後さらに重要になってくると思われる。

参考文献

[1] 河原大輔、黒橋禎夫. 2002. 用言の直前の格要素の組を単位とする格フレームの自動構築. 自然言語処理. Vol.9-1: 3-19.

[2] 日本電子化辞書研究所. 2001. EDR電子化辞書2.0版仕様説明書.

[3] NTTコミュニケーション科学基礎研究所. 1999. 日本語語彙体系. 岩波書店.

[4] Fellbaum, C. (ed.). 1998. *WordNet: An Electronic Lexical Database*. The MIT Press.

[5] Kacmarcik, G., C. Brockett, H. Suzuki. 2000. Robust Segmentation of Japanese Text into a Lattice for Parsing. In *Proceedings of COLING*, pp.390-396.

[6] Richardson, S.D. 1997. *Determining Similarity and Inferring Relations in a Lexical Knowledge Base*. PhD. thesis, City University of New York.

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

[8] Richardson, S.D., W. Dolan, A. Menezes and J. Pinkham. 2001. Achieving Commercial-quality Translation with Example-based Methods. In *Proceedings of MT Summit VIII*, pp.293-298.

[9] Ringger, E.K., R.C. Moore, E. Charniak, L. Vanderwende, H. Suzuki. 2004. Using the Penn Treebank to Evaluate Non-Treebank Parsers. In *Proceedings of LREC*, pp.867-870.

[10] Suzuki, H. 2004. Phrase-Based Dependency Evaluation of a Japanese Parser. In *Proceedings of LREC*, pp.863-866.

[11] Vanderwende, L. 1995. *The Analysis of Noun Sequences Using Semantic Information Extracted from On-line Dictionaries*. Ph.D. thesis, Georgetown University.

[12] Vossen, P. (ed). 1998. *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*. Kluwer Academic Publishers.