自 然 言 語 処 理　130−10
（1999．3．5）

WordNet名詞データの日本語化とその利用

林 良彦[1]
[1] NTTサイバースペース研究所
Email: hayashi@nttnly.isl.ntt.co.jp

概要

英語の語彙データベースであるWordNet 1.6の名詞部分を日本語へ手動で翻訳した．翻訳にあたっては，英語の語彙的概念ノードであるsynsetを構成する英語単語を翻訳するだけでなく，その定義文(gloss)をも日本語へと翻訳した．その結果として，語彙概念にインデックスされた日英の対訳コーパスが得られた．本報告では，この日本語化されたWordNet名詞データの基本特性について報告し，本データが日英語の対照研究における良いリソースであること示す．また，synset単語の対応や上記の対訳コーパスからのバイリンガル辞書の抽出について説明し，得られるバイリンガル語彙データの多言語情報検索への適用可能性について検討する．

Translating WordNet Noun Part into Japanese for Cross-Language Natural Language Applications

Yoshihiko Hayashi[1]
[1]NTT Cyberspace Laboratories
Email: hayashi@nttnly.isl.ntt.co.jp

Abstract

The noun part of the WordNet 1.6 English lexical database was translated into Japanese by human translators. In the course of the translation work, elemental English words/collocations in a synset were translated into Japanese counterparts, as well as the "gloss", which defines the English lexical concept. This translation work turns out to give us an aligned and conceptually indexed pararell corpus. This paper shows the basic characteristics of the "Japanized" noun part of the WordNet. We argue that the resulted data might be a good resource for comparative linguistics. In addition to these, we examine how the derived bilingual lexical data can be applied to cross-linguistic natural language applications, such as cross-language information retrieval.

1 はじめに

WordNet[Miller, 1993]は，Princeton Universityにおけるプロジェクトによる英語の語彙データベース(Lexical Database)である．彼らによれば，WordNetは，そもそもは心理言語学的な検討を行うために実験的に構築した語彙データベースであった．しかし，版を重ねる(公開されている最新版は1.6)につれ収録語彙も増えたこと，自然言語処理の立場からは一種のシソーラスとみなせること，自由に利用できる[1]ことなどから，自然言語処理や情報検索の研究においても広く用いられるようになってきており，一種のdefacto standardとなりつつある．

WordNetを利用した研究は様々な領域にわたる[Fellbaum ed., 1998]が，自然言語処理の観点からは，例えば次のように分類することができるだろう．

1. 曖昧性解消のための知識源として利用する:[Leacock, 1998], [Stetina, 1998]など．
2. 知識ベース獲得のために用いる:[Harabagiu, 1998]など．
3. 他の言語資源との結合や統合を目指す:[Utiyama, 1997], [Ogino, 1997], [Farres, 1998]など．
4. 情報検索における質問拡張などに適用する:[Vorhees, 1998], [Mandara, 1998]など．

我々のグループでは，従来より多言語(Cross-Language)情報検索に取り組んできている([Hayashi, 1997], [Kikui, 1998]など)．この観点から，上記の3や4に分類される研究に特に注目してきた．3に分類される従来のWordNet利用研究では，既存の言語資源(例えばバイリンガル辞書やシソーラスなど)をWordNetへ(自動的に)結びつけることに重点が置かれているのに対し，本研究では，WordNetのデータ自体をまず別の言語に(手動であっても)翻訳し，そのデータを基に様々な検討を進めていこうとするという点が根本的に異なる．当然ながら，我々のアプローチは，初期においてlabor intensiveであるという問題点があるが，得られたデータは，品質も高く，既存の言語資源を相補的に用いることにより，多くのアプリケーションに適用可能であると考えられる．

我々は，このような考えに基づいて，手始めとしてWordNet 1.6の名詞データの日本語化を行った．本報告では，得られた日本語データの特性について述べるとともに，多言語情報検索などにおける利用可能性について検討する．

2 WordNet名詞データの翻訳

2.1 WordNetにおける名詞データ

WordNetの構造は，まず最上位のレベルで品詞別(名詞，動詞，形容詞，副詞)に分割されている．本検討では，情報検索などへの応用を考え，まず名詞部分を検討対象とした．WordNetにおいて基本となるのは，synset (synonym set)と呼ばれる単位である．これはある文脈において置き換え可能な，すなわち同義な単語の集合によって定義される．WordNetにおいては，このsynsetを基本ノードとして，ノード間に様々な関係のリンクが張られている．また，各synsetにはglossと呼ばれる語釈文が付与されている．リンクを省略した各ノードにおける情報は，例えば図1のように整理して示すことができる．

```
Offset: 00292171
LexFile: noun.act
Synset: computer game:1 video game:1
Gloss: a game played against a computer
```

図1: synsetノードの情報．

"Offset"は一種のIDと考えれば良い[2]．"LexFile"とは，WordNetにおけるLexicographer Fileのことであり，WordNetの名詞部においては，26種に分かれている．それぞれは，名詞を上位レベルでカテゴライズしたものに相当し，"unique beginner"などとも呼ばれる．この例では，"computer game"という語は，"act"というカテゴリに属することになる．"Synset"は，このlexicalized conceptを定義する語の集合である．各語の後ろに付与されている番号は，その語形(word form)に対する語義番号である．

WordNet 1.6の名詞部においては，94,496語の英語名詞(複合名詞は1語と数える)が，66,025個のsynsetノードに割り付けられている．なお，synset部分に現れる総単語数は，115,639語であり一語あたりの平均語義数は，1.22である．

2.2 翻訳作業

上記のような規模を持つ名詞データを日本語へ翻訳することとした．人手による翻訳作業はlabor intensiveであるが，結果として得られるデータからは次のようなことが期待できる．

- synset部分から，概念によってインデックスされた，バイリンガル辞書を直接抽出することができる．
- gloss部分から，同じく概念によってインデックスされたパラレルコーパスが得られ，様々な形態での利用が想定できる．
- そもそもは英語を対象とした語彙的概念体系を日本語に置きかえることにより，日英語間での相違などが明らかになり，対照関係を検討するための基礎データとして利用できる．

図1のように表現されるWordNetのsynsetノードは，図2に示すような対応する日本語データに翻訳される．

[1] http://www.cogsci.princeton.edu/~wn/

[2] 実際には辞書ファイルのバイトオフセットであり，これをIDとして利用するのは誤りである．

```
Offset: 00292171
LexFile: noun.act
Synset: コンピュータ・ゲーム:1 ビデオ・ゲーム:1
Gloss: コンピュータと遊ぶゲーム
```

図 2: 翻訳された synset ノードの情報.

翻訳作業においては，あえて厳密な作業基準は示さず，以下のような注意を与えたほかは，比較的自由に作業を行ってもらった．

- synset に現れる各英語単語に日本語訳を一つずつ付与する．
- ある synset において，英語単語が複数存在する場合に，日本語側でそれらの訳し分けが困難な場合は，同じ日本語単語を訳語として付与して良い．
- gloss の翻訳においては，できるだけ両言語間の対応 (句や単語) が取れるような翻訳 (直訳調) を行う．
- 全体にわたっての訳語の統一 (同一の英語単語が同一の語義で使用される場合に同一の日本語訳を与える) はしなくても良い．

以下では，抽出したバイリンガル辞書の特性，パラレルコーパスの利用について説明し，さらに，これらから得られるデータの (多言語) 情報検索への適用可能性について検討する．

3 バイリンガル辞書の特性

3.1 基本データ

図 1 と図 2 における “Synset” フィールドの対応関係から，(“computer game”, “コンピュータ・ゲーム”) や (“video game”, “ビデオ・ゲーム”) という対訳関係を直接抽出することができる．このようにして英日方向の対訳辞書を抽出すると，その見出し語数は 94,496 語となる (元データにおける異なり単語数と同じ)．一方，この辞書を invert して日英方向の対訳辞書を生成すると，その見出し語[3]の数は 72,971 語となった．また，オリジナルの英語データにおける 66,025 個のノードの中で，複数の語により定義されるものは，31,987 個 (48.4%) であるのに対し，日本語データにおいては，14,924 個 (22.6%) にとどまった．前節で説明したように，訳し分けが困難な場合に複数の英語単語に同一の日本語訳語を付与することを許したため，これは十分に予想された結果である．

3.2 日本語訳語の特性:品詞列パターンの分布

言うまでもなく WordNet は英語の語彙データベースであり，その名詞部分は，英語名詞 (複合名詞や collocation を含む) を収録している．これらの名詞を日本語に翻訳した結果が，同様の構造を持つとは限らないことは十分に予想される．そこで，日本語データにおける各 synset の構成語 (異なり語数: 72,971 語) に対して形態素解析 (茶筌 1.5[4]を利用．) を行い，その品詞列パターンを調べることとした．その結果，3,835 個もの異なりパターンが生じた．また，累積で 90%の訳語がカバーされる範囲においても 194 個のパターンが得られた．そこで，名詞類の連続を 1 つのかたまり (“NSEQ” と書く) としてパターンを縮退させた結果，21 個のパターンによって 90%の訳語がカバーされた．表 1 にその結果を示す．この表から分かるように，いわゆる「AのB」の形式の名詞句のほかに，動詞類，形容詞類などの使用がかなり見うけられる．このことは，英語における nominal な概念が，日本語においては単純な名詞句では表現しきれない場合 (複合的な概念) がかなり存在する (約 20%) ことを示している[5]．

そこで，さらに上位レベルの意味分類と考えられる unique beginner ごとのパターンの分布を調べた．その結果を表 2 に示す．一番右のカラムは，品詞パターンの多様性を示す指標として考えることができる．すなわち，少ないパターンで多くの訳語をカバーするほどその値は大きくなる．ある程度予想できるとおり，act や attribute といった抽象的な概念を多く含むであろうカテゴリにおいて値は小さくなっている．act では “〜すること” のような動詞を含むパターンが多く，attribute では “〜であること” といった形容表現を含むパターンが多い．このことから，語彙概念に関する品詞間のリンク[6]を自動設定できる可能性がある．一方，plant, animal, location といった concrete な概念カテゴリにおいては値は大きくなっており，しかも，ほとんどが単純な構造しか持たないことが分かった．

3.3 既存の MRD におけるカバー

翻訳者によって付与された訳語が既存の MRD でどの程度カバーされるかの一例として，日英機械翻訳システム ALT-J/E[7]における日英対照辞書[Ikehara, 1997]の見出し語とのマッチングを調査した．その結果，マッチが見られたものは 18,610 語 (約 26%) であったのに対し，マッチしなかったものは 54,361 語 (約 74%) にのぼった．また，マッチが成功した見出し語の品詞パターンのほとんどが単独名詞または単純な名詞類の連続であったのに対し，マッチしなかったものの品詞パターンは多岐にわたった (累積 90%のカバーまでで縮退パターンで 34 パターン)．このことも，日英語におけるギャップを裏付けるデータとしてとらえることができ，得られた日本語化データを有効に利用する上では，このギャップに対して何らか

[3]いわゆる単語だけでなく構造を持つ名詞句を含む．

[4]http://cactus.aist-nara.ac.jp/lab/nlt/chasen.html

[5]翻訳と形態素解析が妥当なものと仮定してであるが．

[6]これが欠けていることが WordNet の大きな欠点であるという批判がある．

[7]NTT コミュニケーション科学基礎研究所にて研究開発．

表 1: synset 訳語の品詞パターン (累積カバー率 90%まで).

パターン	表現例	頻度	累積カバー率
NSEQ	存在	52042	71.3
NSEQ/接尾辞:名詞性名詞接尾辞:*	動物相	2643	74.9
NSEQ/助詞:名詞接続助詞:*/NSEQ	別れのあいさつ	2597	78.5
NSEQ/接尾辞:名詞性名詞接尾辞:*/NSEQ	心理的特色	1148	80.1
形容詞:*:語幹/NSEQ	生物体	1142	81.6
NSEQ/特殊:記号:*/NSEQ	ネーキッド・オプション	845	82.8
接頭辞:名詞接頭辞:*/NSEQ	再積荷	597	83.6
形容詞:*:語幹	親切	538	84.4
NSEQ/助詞:格助詞:*/動詞:*:基本形/NSEQ	碇を下ろすこと	495	85.0
NSEQ/特殊:空白:*/NSEQ	ウーマンズ　リブ	476	85.7
動詞:*:基本形/NSEQ	ある物	442	86.3
動詞:*:基本連用形/NSEQ	買い受けオプション	437	86.9
動詞:*:基本連用形	置き換え	417	87.5
形容詞:*:タ列基本連体形/NSEQ	物質的な物体	341	87.9
形容詞:*:基本形/NSEQ	軽い接触	327	88.4
NSEQ/動詞:*:基本形/NSEQ	競技すること	288	88.8
NSEQ/動詞:*:基本連用形	動機づけ	236	89.1
動詞:*:語幹/NSEQ	落球	227	89.4
NSEQ/接尾辞:名詞性名詞接尾辞:*/助詞:名詞接続助詞:*/NSEQ	全体性の変化	209	89.7
NSEQ/動詞:*:基本連用形/NSEQ	会社乗っ取り対抗策	166	89.9
接頭辞:ナ形容詞接頭辞:*/NSEQ	不実行	154	90.1

表 2: カテゴリごとの訳語の特性(累積カバー率 90%).
[N:synset 数, X:異なり語数, Y:パターン数]

カテゴリ	N	X	Y	X/Y
artifact	9810	9517	17	559.8
plant	7872	9151	1	9151.0
animal	7294	6866	5	1373.2
person	6409	6591	15	439.4
act	5372	5677	29	195.8
communication	4547	4090	19	215.3
attribute	2633	2691	129	22.4
state	2549	2289	49	46.7
substance	2391	2423	4	605.8
food	2377	2337	7	333.9
cognition	2260	2245	19	118.2
location	2123	2409	2	1204.5
group	1831	1803	5	360.6
body	1592	1502	5	300.4
quantity	1104	1132	12	94.3
object	1050	926	8	115.8
pocession	907	858	15	57.2
time	874	911	18	50.6
event	850	753	41	18.4
phenomenon	523	487	7	69.6
process	521	467	11	43.5
feeling	393	356	36	9.9
relation	369	380	29	13.1
shape	299	290	12	24.2
motive	40	49	9	4.4
Tops	35	50	5	10.0

の対処を行う必要性があることを示している.

4 パラレルコーパスの利用:対訳データの抽出

gloss 部分の翻訳により得られるパラレルコーパスは，概念によってインデックスされているという特徴を持つ．そのため，このようなコーパスには幅広い利用可能性が考えられるが，今回は訳された日本語 gloss に出現する日本語単語に対する英語訳語の抽出を試みた．なお，後述する処理を施したあとに抽出される異なり単語 (token) 数は，英語側が 38,080 語，日本語側が 31,467 語であった[8].

4.1 抽出実験の設定

以下のような設定で抽出実験を行った.

1. 他の言語資源を使用せずに本パラレルコーパスのみから対訳データを抽出し，既存のバイリンガル辞書とのマッチングによって結果を評価する．(実験 1)
2. 既存のバイリンガル辞書を援用し，そこに存在しない，すなわち，未知の対訳データを抽出する．既存の辞書の利用法としては，
   (a) 前節で抽出したバイリンガル辞書において，日英，英日の両方向で一致するペアを抽出対象から除外する．(強い制約条件：実験 2-a)
   (b) 前節で抽出したバイリンガル辞書と前述の ALT-J/E の日英対照辞書を併用し，いずれかの辞書で一致するペアを抽出対象から除外する．(弱い制約条件：実験 2-b)

   の二通りを行なった.

4.2 対訳データの抽出処理

今回の実験の目的は，対訳データの自動抽出のアルゴリズムを追及することではないので，とりあえずは既存の方法[Utsuro, 1995]と同様の抽出処理を適用した．また，あえて[Kitamura, 1997]や[Oomori, 1997]で提案されているような複合語や連語の扱いは行な

[8]gloss は意外に少ない語彙で記述されていることが分かる.

わず，どのようなデータが抽出されるかをみることとした．

抽出処理の流れを簡単にまとめると以下のようになる．

1. 頻度のカウント: 各ノードペアについてについて以下を行う．
   (a) 英語 gloss について前処理を行い token を抽出する．
   (b) 日本語 gloss について前処理を行い token を抽出する．
   (c) 英語側の token の頻度 ($F(Et)$)，日本語側の token($F(Jt)$) の頻度をカウントアップする．
   (d) 日本語 token と英語 token の組み合わせの頻度 (共起頻度 $F(Jt, Et)$) をカウントアップする．
2. Dice 係数の計算: 各日本語 token に対して以下を行う．
   (a) この日本語 token との共起頻度があらかじめ定められた閾値を超える英語 token との組み合わせについて，Dice 係数を計算する．Dice 係数は，良く知られているように次式で定義される．

$$D(Jt, Et) = \frac{2 \times F(Jt, Et)}{F(Jt) + F(Et)}$$

   (b) $D(Jt, Et)$ が最大となる英語 token をこの日本語 token に対する対訳データとして選択する．

ここで，英語側の前処理とは次のようなものである．

1. space, punctuation で token へ分割する．
2. token について以下を行う．
   (a) いわゆるストップワードに該当する機能語や closed class の語を除去する
   (b) WordNet に付属するツールで使用されている形態素解析 **morph** を用いて，token を名詞と仮定して，原型 (lemma) へ戻す．

また，日本語側の前処理とは次のようなものである．

1. 茶筅 1.5 にかけ，token へ分割する．
2. 各 token について，以下を行う．
   (a) 機能語相当のものを品詞条件によるチェックにより除去する．
   (b) 「する」,「いう」などの機能用言や記号類 (の一部) を除く．

4.3 実験結果と評価

4.3.1 実験 1

実験 1 においては，共起頻度の閾値に対する対訳ペアの抽出率と，抽出したペアの精度の評価を行った．後者に対しては，評価を自動化するため，既存のバイリンガル辞書を利用し，それとのマッチングを試みた．図 3 に結果をまとめる．訳語抽出率とは，与えられた共起頻度閾値のもとで何らかの訳語候補が得られるかを示すものである．ここで，分母は，日本語側の異なり語数 (31,467) である．訳語抽出成功率とは，抽出した訳語の中で既存の辞書 (WNJ：3 節で述べたバイリンガル辞書，ALT:ALT-J/E の日英対照辞書) と完全に一致する (訳語完全一致率) だけでなく，以下に示すいずれかの条件でマッチしたものの割合を示す．

表 3: マッチ条件の分布 (%).

マッチ条件	WNJ	ALT
完全一致	79.4	72.1
部分一致	20.6	27.9
出現形 (as is)	96.2	92.7
lemma	1.2	3.2
stem	2.7	4.1

1. 文字列が部分一致する．抽出した訳語が正解訳語の一部である場合 (複合語の構成要素や派生語の共通部分など) と考え，(準) 正解とする．
2. **morph** によって復元された lemma (原型:品詞は問わない)，または，stemming [Porter, 1980] によって得られる文字列 stem (語幹:不変化部分と想定される) が一致する．

また，新訳語候補語抽出率とは，既存のいずれの辞書にも見出し語として存在しない語に対して何らかの訳語が得られたものの割合である．予想されるとおり，閾値を増すと訳語抽出成功率は低下するが，抽出したものの精度は高くなり，新訳語候補抽出率は低下する．

表 3 に共起頻度閾値を 2 としたときの場合のマッチ条件の分布を示す．ALT の場合の方が lemma または stem によるマッチ率が若干高くなっている．これは，ALT が名詞以外の品詞も含んでいるためと考えられる．

なお，本評価は，上記のような条件により自動的に行った．そのため，本来ならば正解とされるべきものが不正解となる場合 (バイリンガル辞書の見出しとの日本語側での不一致など) が含まれていることに注意されたい．

4.3.2 実験 2

共起頻度閾値が 2 の場合の結果を表 4 に示す．参考のため，実験 1 の同条件の場合の結果も示す．

さて，ここでとりあえず抽出に成功したと考えられる語数の総計を検討してみよう．実験 1 の場合，これは 7,663 語である．実験 2-a においては，すでに WNJ にマッチする語が事前に除かれているので，上記の 3,873 語にこれらの語 (31467-27183=4,824 語) を加算することになり，総抽出語は 8,157 語となる．同様に実験 2-b の場合は，8,492 語となる．このことから，既知の (あるいは信頼性高く抽出された) 対訳ペアをもとのコーパスから除くことにより，次段階では新たな対訳ペアが抽出できるという [Kitamura, 1997] と同様の結果が確認された．

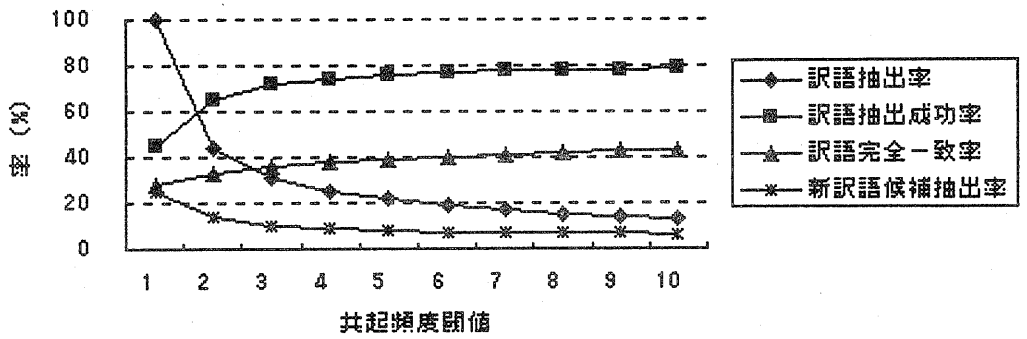

図 3: 対訳ペアの抽出

表 4: 実験 2 の結果 (共起頻度閾値=2).

	実験 2-a	実験 2-b	実験 1
対象語数	27183	25899	31467
訳語抽出率	34.8	33.6	43.7
訳語抽出成功率	49.5	41.5	64.8
抽出成功語数	3873	2924	7663

次に，抽出した対訳ペアの日本語側と一致する既存の MRD の見出し語が存在するが，英語側一致するものが得られない場合 (ケース X: 形式的には抽出失敗とみなされるもの) と，抽出した対訳ペアの日本語側に一致する見出し語が既存の MRD に存在しない場合 (ケース Y: 形式的には未知の対訳ペア候補を抽出したとみなせるもの) について分析を行う．以下では，実験 2-b の共起頻度閾値が 2 の場合を対象とする．この場合，ケース X と分類される件数は 1,732 件，ケース Y と分類される件数は 1,636 件であった．両ケースについて，共起頻度 10 以上のもの (X:426 件, Y:257 件)，Dice 係数が 0.6 以上のもの (X:411 件, Y:741 件) を抽出し詳細な分類を試みた．まず，この抽出件数で注目すべきことは，ケース X についてはどちらの抽出条件においてもほぼ同程度の件数が得られたのに対し，ケース Y については，低頻度であるが Dice 係数が高いものが多く存在する点である．このことは，日本語側で未知語になるものが多いことを示唆するが，実際，ケース Y についての Dice 係数による抽出の 741 件のうち，カタカナ表記の語は 427 件 (57.6%) を占めている．

分析結果を表 5 にまとめる．各分類の説明は表に示したが，“正解” というのは，チェックに用いた MRD の不備を表している (ケース X の場合は日本語見出し語に対する英語訳語の不足，ケース Y の場合は日本語見出しの不足) とも言える．さて，この結果から言えることは次のとおりである．

1. 抽出の有効率 (二種の誤りを除いたもの) は，ケース X の方が高い．
2. どちらのケースにおいても，頻度条件で抽出したほうが有効率が高い．
3. 特にケース X の頻度条件抽出の場合の抽出単位誤りの率が低い．これは形態素解析の成功率が高いことを示している．
4. ただし，どちらのケースにおいても，“正解” の率は，Dice 係数条件で抽出した場合の方が高い．
5. どちらにケースにおいても，Dice 係数条件による抽出における “抽出単位誤り” が頻度条件の場合に比べて高い．日本語側でカタカナ連続語を一語の未知語とする[9]のに対して英語側では複数語による collocation が対応する場合や，日本語側で動植物の属・科・目などを示す接尾辞が分離されてしまう場合が多く見られた．[Kitamura, 1997]や[Oomori, 1997]におけるような複合語や連語の扱いを導入することがこれらを救済するのに有効であると考えられる．

以上の結果から，ケース X，ケース Y の場合についても，かなり有用な対訳，あるいは日英の何らかの対応関係の情報が得られる可能性があることがわかるが，有用性の基準を明確にして，それに合致したものを自動判別することが重要になってくる．

5 情報検索への適用可能性

シソーラスを情報検索における質問拡張に適用しようという試みは以前より行われてきた．WordNet も一種のシソーラスであると考えられ，WordNet を適用しようという代表的な試みとして[Vorhees, 1998] がある．その結論は，query term の語義が正しく解消されていれば，シソーラスを用いて有用な質問拡張が可能というものである．つまり，語義解消が極めて重要ということになるが，以下では，この問題は別の問題であると考え，本報告で述べたようなデータの情報検索への適用性について検討する．

5.1 Monolingual な質問拡張

WordNet を用いた質問拡張においては，どのような質問拡張を行なうかが問題となる．すなわち，ど

[9]茶筅 1.5 の標準辞書，ルールを使用した．

表 5: ケース X,Y の詳細分析 (%).

分類	X:頻度	X:Dice	Y:頻度	Y:Dice	説明
正解	26.1	29.7	31.9	43.0	そのままで正解と考えてよいもの
派生語	36.4	29.7	30.7	19.7	日本語名詞に対して英語形容詞などが抽出されたもの
連想関係	11.7	6.6	3.5	3.4	格要素や上位・下位概念などが抽出されたもの
派生語 (反意)	0.5	1.2	0.8	0.9	反意の形容詞が抽出されたもの
接辞	5.6	0.9	3.5	0.3	日本語の接辞に対して訳語が抽出されたもの
抽出単位誤り	8.9	23.1	18.7	25.8	日本語の抽出単位の誤りによる誤り
抽出誤り	10.8	8.8	10.9	6.9	抽出単位は正しいが対応する英語側が誤っているもの
抽出有効率	80.3	68.1	70.4	67.3	二種の誤りを除いたものの比率

画家 `artist` (上位-下位関係)
生物 `organism` (全体-部分関係)
クラッシュ `disk` (格関係)
結核 `bacillus` (その他の関連)
アイビーエム `Microsoft` (その他の関連)

図 4: 興味ある対応関係の例.

のような関係を用い，階層構造のどの程度までを拡張の対象とするかなどである[Flank, 1998]. しかし，もっとも単純な拡張方法は，同一の synset に属する他の語を query に追加するという方法であろう. オリジナルの WordNet において，複数の異なり語によって定義される synset の数は，31,987 件 (48.4%) であり，これらについては，(適切な synset が決定されれば) このデータを質問拡張に用いることができる. 一方，日本語化されたデータについていえば，同様な synset の数は，14,924 件 (22.6%) であり，これも直接的な質問拡張に用いることができる. ただし，すでに 3 節で述べたように複合語や句構造を持つような翻訳表現が多く存在するため，それを query 解析時に正しく扱う必要がある. さらに，gloss の翻訳データにおいて，synset を単位とした共起関係を計量すれば，これを用いた質問拡張が可能であるが，いずれにせよ，実際の検索のタスクにおいて評価を行なう必要がある.

5.2 Cross-Language な質問拡張

まず考えられるのは，monolingual の場合と同様に，同一の synset に含まれる相手側言語の語を用いるという方法である. これは，3 節で抽出したバイリンガル辞書を用いて query の翻訳を行なうことを意味する. さらに，前節で示したような，対訳コーパスから自動獲得した対応関係を用いる方法が考えられる. この場合は，“派生語” や “連想関係” と分類された語などは，対訳としては正しくないが，質問拡張としては有効に用いることができる可能性がある. 例えば，前節の実験においては，以下のような “興味ある関係” を抽出している.

このような関係を用いた Cross-Language な質問拡張は，通常のシソーラスを用いた質問拡張と同様の問題 (再現率は向上しても適合率は低下する場合が多い) に加え，両言語間の概念の差異を含むので，実際の利用状況も考慮した評価が必要であると考えられる. また，どのような関係が質問拡張に有効であるかをある程度自動的に判別する手法を開発しておく必要があり，これも今後の重要な課題である. また，このような関係のうち，どの程度が WordNet の階層構造 (正確にはネットワーク構造) によってカバーされるかも日英語の対象研究の観点からは興味あるところであり，今後分析を進めていきたい.

なお，本報告で述べたデータの Cross-Language 情報検索への適用については，[Iwadera, 1998]で述べられている三カ国 (シンガポール:KRDL，韓国:KAIST，日本:NTT) による多言語情報検索サービスに関する共同実験における我々の検索サイトで一部利用される予定である.

6 おわりに

WordNet の名詞部分を日本語へ翻訳したデータの特性について報告し，その利用可能性について検討した. 既存の言語資源と併用することや，より高度な対訳データ抽出アルゴリズムを適用することによって，さらに有益な言語データが抽出可能であると考えられる.

一方，我々の本来の目的は，Cross-Language 情報検索への適用にあるので，そのような言語データを抽出するとともに，実際の利用法についても検討を進めて行きたい. 質問拡張に有効な対訳・対応関係の自動判別は重要な課題である. また，質問拡張などにおいて特に重要な query における語に対する語義解消については，意味タグ付けされたコーパスの利用や，本報告では全く触れなかった WordNet の階層構造の適切な利用について検討していく. さらに，得られた日本語のデータ自体，日英語間の対照研究の観点から興味あるデータと考えられる. 今後はこの問題にもアプローチしていきたい.

謝辞

翻訳作業を担当していただいた日本アイアール株式会社の柴田葉子様，野原剛様，山本 修様ほか皆様に感謝いたします.

参考文献

[Farres, 1998] Xavier Farres, German Rigau, and Horacio Rodriguez. Using WordNet for Building WordNets. *Coling-ACL '98 Workshop on Usage of WordNet in Natural Language Processing Systems*, 1998.

[Fellbaum ed., 1998] Christiane Fellbaum, editor. *WordNet, An Electronic Lexical Database*. The MIT Press, 1998.

[Flank, 1998] Sharon Flank. A Layered Approach to NLP-Based Information Retrieval. *Coling-ACL '98 Workshop on Usage of WordNet in Natural Language Processing Systems*, 1998.

[Harabagiu, 1998] Sanda M. Harabagiu and Dan I. Moldvan. Knowledge Processing on an Extended WordNet. in [Fellbaum ed., 1998], 1998.

[Hayashi, 1997] Yoshihiko Hayashi, Genichiro Kikui, and Seiji Susaki. TITAN: A Cross-Linguistic Search Engine for the WWW. in *Cross-Language Speech and Text Retrieval*, AAAI Technical Report SS-97-05, 1997.

[Kikui, 1998] Genichiro Kikui. Term-list Translation using Mono-lingual Word Co-occurrence Vectors, in *Proc.of Coling-ACL '98*, 1998.

[Leacock, 1998] Claudia Leacock and Martin Chodorow. Combining Local Context and WordNet Similarity for Word Sense Identification. in [Fellbaum ed., 1998], 1998.

[Mandara, 1998] Rila Mandara, Takenobu Tokunaga, Hozumi Tanaka. The Use of WordNet in Information Retrieval. *Coling-ACL '98 Workshop on Usage of WordNet in Natural Language Processing Systems*, 1998.

[Miller, 1993] Geroge A. Miller, Richard Beckwith, Christianne Fellbaum, Derek Gross, Kathrine Miller, and Randee Tengi. *Five Papers on Wordnet.*, Cognitive Science Laboratories, Princeton University. CSL Report 43, 1993.

[Ogino, 1997] Takano Ogino, Hideo Miyoshi, Fumihito Nishino, Masahiro Kobayashi, and Jun'ichi Tsujii. An Experiment on Matching EDR Concept Classification Dictionary with WordNet. *IJCAI-97 Workshop on Ontologies and Multilingual NLP*, 1997.

[Porter, 1980] Porter M. An Algorithm for Suffix Stripping. *Program*, Vol.14(3), 1980.

[Stetina, 1998] Jiri Stetina, Sadao Kurohashi, and Makoto Nagao. General Word Sense Disambiguation Method Based on a Full Sentential Context. *Coling-ACL '98 Workshop on Usage of WordNet in Natural Language Processing Systems*, 1998.

[Utiyama, 1997] Masao Utiyama and Koiti Hasida. Bottom-up Alignment of Ontologies. *IJCAI-97 Workshop on Ontologies and Multilingual NLP*, 1997.

[Vorhees, 1998] Ellen M. Vorhees. Using WordNet for Text Retrieval. in [Fellbaum ed., 1998], 1998.

[Ikehara, 1997] 池原 悟, 宮崎正弘, 白井 諭, 横尾昭男, 中岩浩巳, 小倉健太郎, 大山芳史, 林 良彦 (編集). 日本語語彙大系. 岩波書店, 1997.

[Iwadera, 1998] 巌寺俊哲, 林 良彦, 菊井玄一郎, 小橋喜嗣, Mun-Kew Leong, Key-Sun Choi. 多言語分散情報検索アーキテクチャに関する検討. 情報処理学会自然言語処理研究会, 127-9, 1998.

[Kitamura, 1997] 北村美穂子, 松本裕治. 対訳コーパスを利用した対訳表現の自動抽出. 情報処理学会論文誌, Vol.38, No.4, 1997.

[Oomori, 1997] 大森久美子, 佐藤健吾, 中西正和. 共起関係を利用した対訳コーパスからの連語の対訳表現抽出. 情報処理学会自然言語処理研究会, 122-3, 1997.

[Utsuro, 1995] 宇津呂武仁, 松本裕治. 対訳辞書および統計情報を用いた二言語対訳テキスト照合. コンピュータソフトウェア, Vol.12, No.5, 1997.