自然言語処理 94−11
（1993. 3. 19）

日本語分類語彙表からの韓国語分類語彙表の作成

黄 道三　　長尾 眞　　佐藤 理史*
京都大学工学部 電気工学第二教室
*北陸先端科学技術大学院大学 情報科学研究科

要旨

シソーラスは自然言語処理の研究において重要なデータあるいは知識情報としての役割を果たしてきた。しかし、シソーラスを作成するには人手によるしか方法がなかったので、長期間かけて少なくない開発費とマンパワーを投入して開発されてきた。そこで、本稿では、あらかじめ作成されている日韓機械辞書と日本語の分類語彙表を用いて、韓国語の分類語彙表を半自動的に作成した結果とその評価を示す。特に、意味属性の分類番号への変換と意味属性間の類似性を用いて、分類番号を与える方法を示した。

Construction of a thesaurus for Korean from a thesaurus for Japanese

Dosam Hwang　　Makoto Nagao　　Satoshi Sato*
Department of Electrical Engineering, Kyoto University
*School of Information Science, Japan Advanced Institute of Science and Technology, Hokuriku

Abstract

Thesaurus has traditionally played an important role as a source of knowledge in Natural Language Processing research. However, the cost of manual construction of a thesaurus is very high. This fact has urged attempts to automate the process. We present in this paper a method to produce a thesaurus for Korean using a machine translation dictionary and a thesaurus for Japanese. More specifically, we transferred the semantic markers to the classification number of the thesaurus, not only by direct correspondence, but also by the similarity among the semantic markers. We constructed a thesaurus for Korean using a Japanese to Korean machine translation dictionary and Bunruigoihyou, a thesaurus for Japanese, and evaluated the result.

1 はじめに

19 世紀中頃に Roget によって作成されたシソーラスは最近の自然言語処理の研究になくてはならない重要なデータあるいは知識情報として利用されるようになってきている。日本では、1960 年代に分類語彙表が作成され、1980 年代の機械翻訳研究の発展に多く寄与したと思われる。今でも、意味解析の研究の基礎資料として多く活用されている。

しかし、今日までシソーラスを作成するには人手によるしか他に方法がなかった。シソーラスを持たない言語に関して研究をしようとすると、その言語のシソーラスを作らねばならず、そのためには膨大な開発費とマンパワーを投入しなければならないし、かなりの時間がかかる。

そこで、本稿では、あらかじめ開発されている日韓機械辞書 [1] と日本語の分類語彙表 [2] を用いて、韓国語の分類語彙表を半自動的に作成する方法を示す。ここでは、主に、対応および類推によって韓国語意味辞書 [3] の意味属性を日本語分類語彙表の分類番号に変換する方法を用いた。また、その実験として、韓国語の分類語彙表を作成し、その結果を評価した。

2 機械辞書と分類語彙表

2.1 日韓対訳辞書

日韓対訳辞書は機械翻訳システムのために開発されたもので、図 1のように日本語の見出し語、構文属性、韓国語の対訳語の三つの部分からなっている。見出し語が多義性を持っている場合には、対訳語が最大三つまで記述されている。約 45,000 語の見出し語に対して約 48,000 語の対訳語が登録されている。ここでは、連語、熟語、特殊文字などの翻訳のために登録されている単語を取り除いた約 34,700 語とその対訳語の 36,500 語を対象にした。

$Jkey$	$SynA$	$Ktran_1$	$Ktran_2$	$Ktran_3$

Jkey: 見出し語 *SynA*: 構文属性 *Ktran*: 対訳語

図 1: 日韓対訳辞書のレコード構成

2.2 韓国語の意味辞書

韓国語の意味辞書には図 2のように、韓国語の見出し語、概念記号、意味属性、格情報および機能属性、隣接情報が記されている。概念記号とは単語が持つ意味を表現するもので、例えば、「나(私)」、「간다(行く)」「학교(学校)」の場合、各々「I」、「go」、「school」のように英語で表されている。また、意味属性は大きく 2 つのサブ属性から構成されていて、第一番目の文字は名詞、動詞、形容詞、副詞などの構文属性を、残りの文字列は意味属性を表していて、名詞 99 個、動詞 53 個、形容詞 31 個、副詞 28 個の意味属性が分類されている。

つまり、図 3のように、意味属性の体系はレベル 1 に品詞分類項目を、レベル 2 に意味分類項目を持ち、レベル 3 から意味属性が細分類されていく階層構造になっている。一つの単語には複数の意味属性が記されている。例えば、「나(私)」には「NANI(動物)」、「NCON(具体物)」、「NCRE(生物)」、「NHUM(人間)」などの四つの意味属性が記述されている。また、構文情報としては格情報と機能属性があり、格情報は用言に対してどんな格要素でどのようなパターンに表わされるかが格属性で記されていて、機能属性の部分には単語の文法的機能によって分類されている属性が記されている。

$Kkey$	CW	$SemA_1$	⋯	$SemA_n$	CI	AI

Kkey: 見出し語 *CW*: 概念記号
SemA: 意味属性 *CI*: 格情報・機能属性 *AI*: 隣接情報

図 2: 韓国語意味辞書のレコード構成

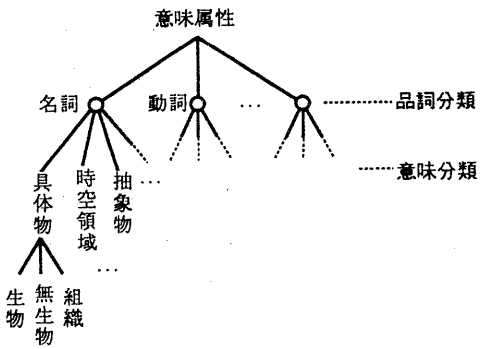

図 3: 意味属性の体系

2.3 日本語の分類語彙表

ここで使った日本語の分類語彙表はオンライン版で、その分類番号は図 4のように、5 桁の分類番号、段落番号、段落内番号の三つに分けられている。ここでは、意味によって明確に分類されている分類番号だけを扱う。これは、意味属性を段落番号と段落内番号までには変換ができないからである。

分類番号の左から第 1 桁は大分類として主に品詞によって体の類、用の類、相の類、その他など四つの類に

$IN_1 \mid \cdots \mid IN_5 \mid TN \mid SN \mid Jkey$

IN: 分類番号　*TN*: 段落番号　*SN*: 段落内番号　*Jkey*: 見出し語

図 4: 日本語の分類語彙表のレコード構成

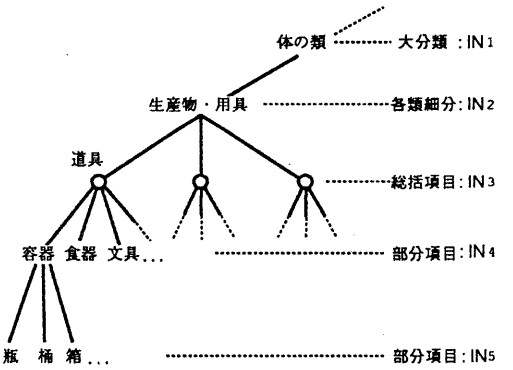

図 5: 分類番号の体系

分けられている。次の第 2 桁は各々の類に対して、意味的に細分されていて、第 3 桁から第 5 桁まではもっと細かく意味分類された細分意味番号が付いている。そして、図 5のように、深さ 5 の階層構造で示すことができる。また、分類番号は総計 798 項になっている。

3　韓国語分類語彙表の作成方法

3.1　単純検索による方法 (方法 1)

日本語と韓国語には意味上の使い方がほとんど完全に一致すると思われる単語が多い。我々は、

- 2 文字以上の日本語単語で 1 文字以上の漢字を含み、
- 日韓対訳辞書で唯一の韓国語単語が対訳語として与えられている

単語については意味上の使い方が日韓でほとんど同じであると仮定し、分類語彙表でその日本語単語が何ヶ所かに表れ、複数個の分類番号をもつ場合もそれら全ての分類番号を対応する韓国語単語に与えるようにした。これを方法 1 と呼ぶ。4 章の表 2 に示すように、この対応関係は非常に精度の高いものであることが分かり、我々の仮定が妥当なものであることが明らかとなった。

3.2　意味属性の分類番号への変換による方法 (方法 2)

3.2.1　分類語彙表の作成の流れ

日韓対訳辞書の見出し語 (日本語) をキーとして、分類語彙表からその単語の分類番号を検索する。ところで、多くの単語は複数の意味を含んでいるため、一つの見出し語に対して分類語彙表には複数の分類番号が付いている。また、両言語の意味上の使い方が一対一

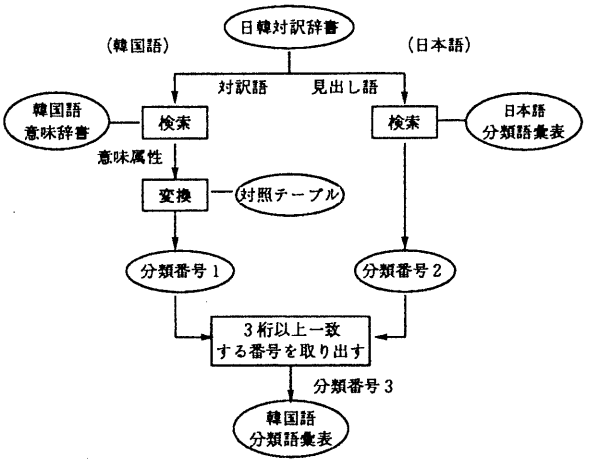

図 6: 韓国語分類語彙表の作成の流れ

に完全には一致しないため、日韓対訳辞書にも一つの日本語単語は複数の韓国語単語に対訳されているものもある。したがって、これらの分類番号を韓国語の対訳語に単純に付与することはできない。例えば、日韓対訳辞書では日本語の「本」が韓国語の「책 (book)」, 「권 (volume)」, 「본 (this)」などに対訳されていて、分類語彙表では「本」に「11960(単位)」, 「13160(文献)」, 「14590(帳)」, 「31000(こそあど)」など複数の分類番号が付いている。

この分類番号を各々意味的に一致する韓国語に付与するために図 6に示す方法をとった。すなわち、韓国語に対する意味属性が必要となるので、韓国語の意味辞書を用意した。そして日韓対訳辞書の対訳語をキーとして、韓国語の意味辞書からその意味属性を抽出する。ところが、この意味属性と分類番号との分類体系が違っているので、それらを対照した対照テーブルを作って、図 6に示すように対応される分類番号 1 を生成する。次に、この分類番号 1 と分類語彙表より検索された分類番号 2 とを比較して類似性の高い分類番号 3 だけを抽出して韓国語の分類語彙表を作成する。この比較は多くの場合分類番号全体で一致がとれず、上位 3 桁以上の一致が得られた分類番号のみを取り出した。

3.2.2　意味属性の分類番号への変換

意味属性を分類番号に変換するためには、両分類項目の対照テーブルが必要になる。しかし、韓国語の意味辞書の意味属性と分類語彙表の分類番号とは分類体系が違っているし、それぞれの分類の数が多いので、人手で

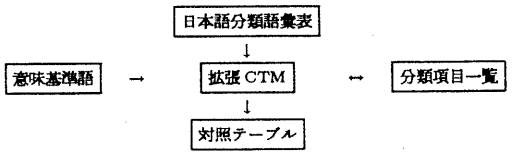

図 7: 対照テーブルの作成

それらの対応づけをするのは大変であり、また間違う可能性が非常に高い。ところで、韓国語の意味辞書と分類語彙表の分類項目一覧には、各々分類の基準となる見出し語が日本語で表 1のように付いている。そこで、図 7のように韓国語意味属性の分類のための基準語 (以下意味基準語) を入力キーとし、分類語彙表の分類のための基準語 (以下分類基準語) と文字列同士の比較を行い、最も似ている分類基準語を取り出した。これは文献 [4] の方法をシソーラスによって拡張した方法によっている*1。こうすることによって入力キーに意味的に似ている単語が検索されて出てくる。こうして検索された分類基準語を人手で検討して意味的に一番近い分類基準語の分類番号だけを選んで、意味属性と分類番号との対照テーブルを作った。こうすることによって効率的に意味基準語に対応する分類基準語を探すことができた。

例えば、図 8に示すように意味基準語の「方向」を入力すると、それと意味的に近い単語が類似点数の大きい順にその点数とともに「方向」、「上下」、「方面」、「左右」などの分類基準語とその分類番号が検索されてくる。類似点数の計算の仕方は次節で述べるが、日本語分類語彙表の分類番号の一致度を用いて求めた。対照テーブルの一部分を表 1に示した。

3.2.3 一致度による分類番号の選択

3.2.1 で説明したように分類語彙表より検索された分類番号 2 の中で対照テーブルを用いて意味辞書より生成された分類番号 1 と一致する番号だけを選ぶ必要がある。だが、両番号には完全に一致するのもあれば、部分的にしか一致しないものもある。特に、部分一致するのは複数個存在することが多い。ところで、図 5をみるとレベル 3 まで一致すると、相当に意味的に近いことが分かる。そこで、少なくともレベル 3 まで一致する分類番号だけを選ぶことにした。すなわち分類番号の上位から少なくとも 3 桁以上一致するものを選ぶのである。

例として、図 9のように「나(私)」の場合には、意味辞書より「NCON」, 「NCRE」, 「NHUM」などの意味属性が検索され、対照テーブルによる分類番号への変換によって「14000(人間活動の生産物)」, 「15500(生物)」, 「12020(人間)」などの分類番号 1 が生成される。一方、分類語彙表からは「12000(われ)」, 「12500(公私)」などの分類番号 2 が検索されて出てくる。これらの番号の間の比較の結果、一致度が 5 である「12000」が選ばれることになる。この時、一致度は次のような数式で求めている。

$$Md = \{k | BN1[1,\ldots,k] \equiv BN2[1,\ldots,k] \wedge BN1[k+1] \neq BN2[k+1]\} \quad (1)$$

Md: 一致度 *BN1*: 分類番号 1 *BN2*: 分類番号 2

*1 この方法については別の機会に発表する。

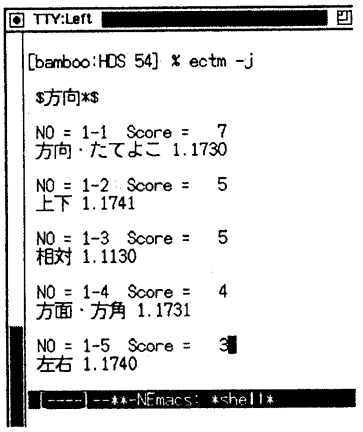

図 8: 拡張 CTM の検索結果の例

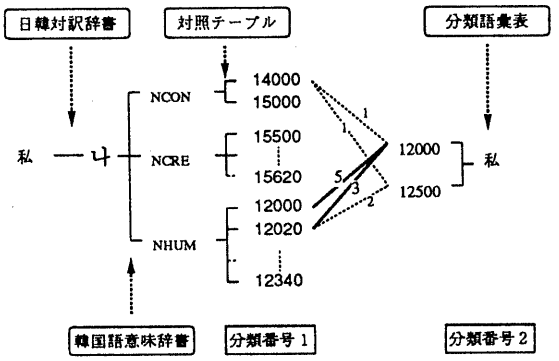

図 9: 一致度による分類番号の選択

3.3 類推による分類番号の生成 (方法 3)

上記の二つの方法で約 12,000 個の分類番号を韓国語の単語に付与することができた。それは日韓対訳辞書の対訳語のみで、対訳語として登録されていない単語については分類番号を付与することができない。そこで日韓対訳辞書には載っていないけれども、意味辞書には載って

表 1: 意味属性と分類番号の対照テーブル

意味辞書		分類語彙表
意味基準語	意味属性	分類番号 (分類基準語)
能力、性向	N.ABI	1.1404(能力) 1.1330(性質)
建築物、施設	N.ARCH	1.3823(建築) 1.4720(その他の土木施設)
方向	N.ARR	1.1730(方向・たてよこ) 1.1741(上下)
		1.1731(方面・方角) 1.1740(左右)
具体物	N.CON	1.4000(人間活動の生産物) 1.5000(自然)
生物	N.CRE	1.5500(生物) 1.2040(男女) 1.5620(鳥)
		1.4323(さかな・鰹節・肉) 1.5510(植物)
人間	N.HUM	1.2020(人間) 1.2000(人間活動の主体) 1.2340(人物)
		1.2050(老少) 1.2040(男女) 1.2000(われ・なれ・かれ・だれ)
社会現象、発生	V.ASPHE	2.1210(出現) 2.1220(成立・発生)
無意志、抽象的状態、優劣	V.DIF	2.1900(過不足・優劣など)
無意志、社会現象、衰退	V.DSPHE	2.1583(強め・衰えなど) 2.1526(進退)
無意志、存在状態	V.EXI	2.1200(存在) 2.1240(残存・消滅)

いる単語については以下の類推による分類番号の付与方法を用いた。この処理概念を図 10に示す。意味辞書の中には上記の処理で分類番号の付与された単語 (図 10の黒い部分) と、分類番号の付与されていない単語 (図 10の白い部分) がある。ところが、両方とも概念記号、意味属性、格情報および機能属性、隣接情報を持っている。そこで、これらの構文および意味情報の一致度を用いて黒い部分の単語から白い部分に意味的に似ている単語を探して、上記の方法で予め構築されている韓国語の分類語彙表からその単語の分類番号を取り出して、白い部分の単語に写す。このようにして、日韓対訳辞書に載っていない単語にも分類番号を与えることができる。しかし、この時にも多義語の場合があるので、検索された分類番号を 3.2.3 のような方法で意味的に一致する番号だけを選ばなければならない。図 10で実線は意味的に強い一致をすることを、点線は意味的に弱い一致しかしないことを表す。この場合の意味的一致度は次のような方法で求めた。

$$Sd = w_1 \times Mcw + w_2 \times Msa + w_3 \times Mci + w_4 \times Mai \quad (2)$$

$$\begin{cases} if\ Icw = Rcw\ then\ Mcw = 1\ else\ Mcw = 0 \\ if\ Isa = Rsa\ then\ Msa = 1\ else\ Msa = 0 \\ if\ Ici = Rci\ then\ Mci = 1\ else\ Mci = 0 \\ if\ Iai = Rai\ then\ Mai = 1\ else\ Mai = 0 \end{cases}$$

Sd : 一致度の合計 M_n : 属性別一致度

w_n : 加重値 I : 白い部分の単語 R : 黒い部分の単語

cw : 概念記号 sa : 意味属性 ci : 格情報・構文属性 ai : 隣接情報

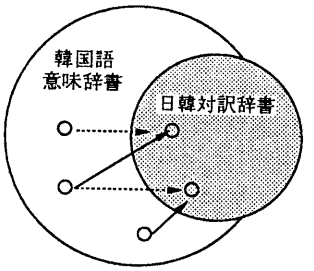

図 10: 類推による分類番号の付与の概念度

4 韓国語分類語彙表の検索および評価

4.1 韓国語分類語彙表の検索

上記の方法で 1 万語あまりの単語について韓国語の分類語彙表を作ることができたので、ここではその検索例を示す。検索方法には完全一致法と末尾最長一致法の二方法があり、検索例を図 11に示す。

4.2 評価

上記の三つの方法によって 9,274 語の韓国語に対して各々 11,497 個、 465 個、 263 個の分類番号を与えることができた。ここで作成された結果に対しては一つ一つの分類番号の正確性を検討しなければ、分類語彙表としての良さを判断することが難しいということを考慮して、すべての単語に対して人手によってその妥当性を調べた。

評価のランクは次のように二つに分けた。

A 分類番号が IN と TN で一致していて、それが適切

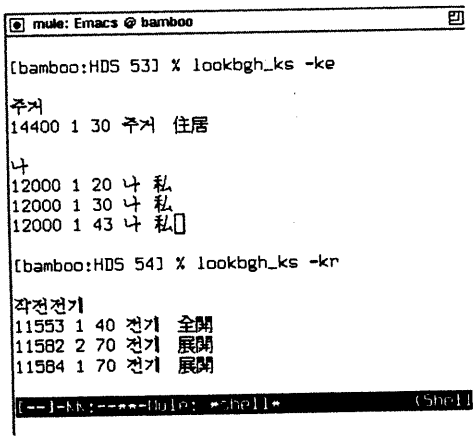

図 11: 検索結果の例

である。

E 与えている分類番号 (IN・TN) が違っている。

評価の結果、韓国語の単語に与えられた分類番号の中で約 98% が正しいことが分かった。表 2は第 3 章に述べた方法 1 から方法 3 までの作成方法別に分けて評価した結果で、表 3は分類番号を品詞別に分けて評価した結果である。また、本評価の後、約61,183 語の見出し語に対して約 67,187 語の対訳語が登録されている別の日韓対訳辞書を用いて、上記の三つの方法を適用して韓国語単語約 2 万 3 千語に対して表 4 のように約 3 万 3 千個の分類番号が与えられて、ある程度大量の韓国語分類語彙表を作ることができた。しかし、この分類語彙表についてはまだ評価を行なっていない。

5 おわりに

本稿では、既存の日韓機械辞書と日本語の分類語彙表を利用して、韓国語の分類語彙表を自動的に作成する方法を示した。特に、韓日語間の対照関係、また日韓機械辞書と日本語の分類語彙表の対応関係に基づいて行なわれた単純検索による方法、意味属性の分類番号への変換による方法、類推による方法で非常に良質の結果が得られることが分かった。

表 2: 作成方法別の評価結果

ランク	方法 1	方法 2	方法 3	合計
A	11,261(98%)	455(98%)	242(92%)	11,958(98%)
E	236(2%)	10(2%)	21(8%)	267(2%)
合計	11,497	465	263	12,225

表 3: 品詞別の評価結果

ランク	体の類	用の類	相の類	その他	合計
A	10,170(98%)	796(92%)	959(92%)	33(87%)	11,958(98%)
E	142(1%)	73(8%)	47(8%)	5(13%)	267(2%)
合計	10,312	869	1,006	38	12,225

表 4: 別の日韓対訳辞書からの韓国語分類語彙表の作成

方法 1	方法 2	方法 3	合計
31,280	1,963	96	33,339

しかし、一般の辞書ではなくて、機械翻訳用の辞書を利用したので、見出し語が単語ではなくて、句または節であるものが相当数あり、うまく変換された単語数が少ないという問題点は残っている。しかし、ここで作成された単語を元にして、人手で単語を追加してゆくことは比較的に容易であると思われる。

参考文献

[1] 第 4 回日韓機械翻訳共同研究報告書、韓国科学技術研究院システム工学研究所 + 富士通株式会社 (1986).

[2] 国立国語研究所、分類語彙表、秀英出版 (1964).

[3] 黄 道三、李 尚起、張 遠: 韓日機械翻訳システム開発に関する研究、韓国富士通 (株) の委託研究報告書、韓国科学技術研究院システム工学研究所 (1991).

[4] Satoshi Sato: CTM:An Example-Based Translation Aid System Using the Character-Based Match Retrieval Method, Vol.4, pp.1259-1263, Proc. of COLING-92, Nantes, August 23-28 (1992).

[5] Dongin Park, Dosam Hwang, Sangki Lee, Won Chang et al.: A Study on the Development of the Korean to Japanese Machine Translation System, Proc. of PRICAI'90, JSAI, Nagoya, November 14-16 (1990).

謝辞

本研究に有益な議論と助言を頂いた長尾研究室の方々に感謝します。ならびに研究の遂行に協力して頂いた韓国科学技術研究院の張 遠氏、京都大学の辻村 郁子氏に感謝します。