Grammatical Framework における語彙データの自動生成

渡邉　秀隆　　中野　圭介
電気通信大学 大学院 情報理工学研究科
hwatanabe@ipl.cs.uec.ac.jp, ksk@cs.uec.ac.jp

1　はじめに

Grammatical Framework [1] は，自然言語を統合的に扱うことを目的として Ranta によって提案された枠組みである．これは，特定の言語に依存しない文法である **abstract syntax** と，個々の言語における見た目を定義する **concrete syntax** の 2 つを書くことで，ある言語から別の複数の言語へ翻訳することができるようになるというものである．ある言語 A の文章から別の言語 B に翻訳する場合，まず，abstract syntax と言語 A の concrete syntax に含まれる文法と語彙を用いて，入力された文章を個々の言語に依存しない，意味に基づく論理構造に変換する．その後，その論理構造を言語 B の concrete syntax の文法と語彙を用いて言語 B へ翻訳する．Grammatical Framework には多数の自然言語の文法用標準ライブラリが用意されているが，いくつかの言語においては実験的に作成されたものに過ぎないため語彙の数が少ないという問題がある．特に日本語の標準ライブラリにおいてはその傾向が顕著である．また，英語のように豊富な語彙データがある自然言語においても，他の言語に対して単語が 1 対 1 に対応するような語彙しか登録されていないため，ある単語から同じ意味を持つ複数の単語に翻訳することができない．

本研究では，Grammatical Framework の使いやすさを向上するため，豊富な語彙を持つ英語の語彙データを利用し，日本語の語彙データを自動生成する手法を提案する．同じ意味を持つ単語があれば，それらの語彙を全て生成し，翻訳する際にはそれらの語彙が同列に扱えるようにする．複数の単語への翻訳を可能にするため，概念辞書である日本語 WordNet を用いて自動生成を行う．日本語 WordNet では「概念」を用いて単語が登録されているため，同じ概念を持つ単語を 1 つにまとめることで，1 つの単語に対して同じ意味を持つ複数の単語が対応するような語彙データの作成が可能となる．

2　Grammatical Framework

Grammatical Framework は，自然言語を統合的に扱うための枠組みである．Grammatical Framework の目的は，言語学者だけでなく自然言語を扱うアプリケーションを記述したいプログラマにも貢献することであり，abstract syntax と concrete syntax の 2 つを用いることでそれを実現している．abstract syntax は言語における論理構造を記述するものであり，例えば形容詞は名詞節を受け取って名詞節になる，といった言語に依存しないことや，単語の概念等を記述する．具体的には，日本語において「かわいい」は形容詞であり，「猫」は名詞であるため，「かわいい猫」という節は名詞節となるというような節の構造や，「物事の進む度合が大きいことを表す単語」は形容詞であるということを abstract syntax に記述する．このような構造は abstract syntax の中では型情報として与えられているため，型検査の枠組みにより，Grammatical Framework では誤った構文が型エラーとして検出される．concrete syntax は自然言語における見た目を記述する．例えば，abstract syntax で定義されている「物事の進む度合が大きいことを表す形容詞」に対して，日本語の concrete syntax では「速い」，英語の concrete syntax では「fast」が割り当てられる．

Grammatical Framework の概要を図 1 に示した．ここでは，言語 A，言語 B，言語 C の間で相互に翻訳

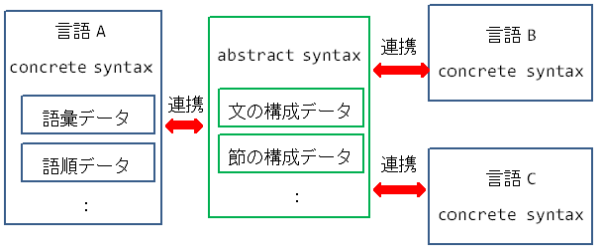

図 1: Grammatical Framework の概要

が可能となるような例を示している．

各言語の concrete syntax には語順や語彙等が記述されている．他の言語に翻訳する際は，abstract syntax に記述されている文の構成データや節の構成データ等を用い，意味に基づく論理構造に変換し，変換された論理構造を目的の言語の concrete syntax を用いて翻訳する．

3　本研究における実装

本研究のシステムは，2 つの部分からなる．1 つ目は，Grammatical Framework の英語の語彙データを入力とし，日本語 WordNet [2] を用いて Grammatical Framework の日本語の語彙データを生成する部分である．2 つ目は，Grammatical Framework の英語，日本語の両方の言語の語彙データを入力として，日本語 WordNet を用いて同じ意味の単語を 1 つにまとめる部分である．

3.1　日本語 WordNet

日本語 WordNet とは，日本語の意味辞書である．独立行政法人情報通信研究機構が開発した．日本語 WordNet は，プリンストン大学で開発された Princeton WordNet [3] や Global WordNet Grid [4] に着想を得て開発された．

WordNet の特徴として，類義語関係のセット (synset) でグループ化している点が挙げられる．1 つの synset が 1 つの概念に対応し，各 synset は上位下位関係等の多様な関係で結ばれている．上位関係にある単語を上位語，下位関係にある語を下位語と呼び，上位語は当該 synset を包含するような synset，下位語は当該 synset に包含されるような synset のことである．例えば，日本語の「犬」という単語の上位語の 1 つは「イヌ科動物」であり，下位語には「小型犬」や「パグ」などがある．

3.2　日本語語彙データの自動生成

Grammatical Framework において，各言語の語彙は他の言語の語彙と対応している必要がある．すなわち，英語語彙と日本語語彙にも対応がある．そのため，英語語彙データの情報を利用して日本語語彙データを生成するようにした．システムはまず，入力された英語の語彙データを一行読み取り，単語，品詞，活用等の情報に分解する．その後，それらの情報を用いて日本語 WordNet にクエリを与え，Grammatical Framework の日本語の語彙データとして出力するために必要な情報を得る．後で述べるように語彙データに必要な情報は日本語の品詞の種類によって異なる．クエリは，必要な情報がすべて得られるまで，日本語 WordNet に与えられる．得られた情報は Grammatical Framework の語彙データの形式に整形され，ファイルに書き込まれる．入力として受け取った英単語に対して，同じ意味の日本語が複数存在する場合は，それぞれの日本語に対して Grammatical Framework の語彙データの形式に整形しファイルに書き込む．日本語の語彙データにおける品詞の分類は，大きく分けて以下の 4 つである．

- 名詞
- 動詞
- 形容詞
- 副詞

日本語において名詞を修飾するものには，形容詞と形容動詞の 2 つがあるが，Grammatical Framework の語彙データにおいてはどちらも形容詞として扱う．以下，各品詞における語彙データの自動生成について説明する．

3.2.1　名詞の自動生成

Grammatical Framework の日本語の語彙データにおける名詞に必要な情報は，単語，動物か非動物かの判定，助数詞，助数詞化の可否，の 4 つである．表 1 にそれぞれの例を示す．

単語　日本語の単語は，日本語 WordNet に英単語の synset を問い合わせることで得られる．

動物か非動物かの判定　動物か非動物の判定については，上位語の概念におけるすべての先祖を用いて判定を行う．この判定は，日本語における存在を表す助詞「ある」と「いる」の使い分けに利用される．日本語 WordNet から，その単語の概念の上位語のすべてを受け取り，その概念の中に動物か人間かいずれかの概念が含まれているかどうかで判定を行う．

助数詞　助数詞とはものを数える際に数量を表す語の後ろにつける接尾語である．助数詞には，人間の数を表す「人」や薄くて平らなものを数えるときに用いる

表 1: 日本語語彙データにおける名詞の自動生成に必要な情報の具体例

必要な情報	英単語の具体例		
	cow	airplane	year
単語	「牛」「雌牛」	「飛行機」「航空機」	「年」「歳」「年間」
動物か非動物の判定	動物	非動物	非動物
助数詞	「頭」	「機」	「年」
助数詞化の可否	否	否	可

「枚」などがある．助数詞の決定は，単語の上位語の先祖の概念の助数詞によって決定している．例えば，「牛」の上位語の先祖の中に「有蹄類」を含むため，助数詞は「頭」を選択する．

助数詞化の可否　助数詞化とは，例えば「3 台の車」という文を「3 車」に置き換えることができるかどうかというものである．助数詞化の可否の判定は，助数詞とその単語が同じ場合のみ可とし，それ以外は不可とした．

3.2.2　動詞の自動生成

Grammatical Framework の日本語の語彙データにおける動詞に必要な情報は，基本的には単語のみである．その単語が他動詞の場合，その他動詞の目的語に必要な助詞も必要である．まず，日本語 WordNet を用いて英単語を日本語に翻訳する．さらに，その動詞が他動詞である場合には助詞として「を」を選択する．目的語が二つ必要な他動詞の場合は，助詞に「に」と「を」を選択する．

3.2.3　形容詞の自動生成

Grammatical Framework の日本語の語彙データにおける形容詞に必要な情報は，単語のみである．Grammatical Framework において「い」で終わる形容詞と「な」で終わる形容動詞は，どちらも名詞を修飾するため同一のものとして扱っており，語尾が「い」または「な」ではない形容詞を含む語彙データはコンパイル時にエラーが出る．日本語 WordNet においても同様に形容詞と形容動詞を同列に扱っているが，形容動詞の場合は語幹だけが単語として登録されている．例えば,「速い」はそのまま「速い」として日本語 WordNet に登録されているが,「ワイドな」は「ワイド」として日本語 WordNet に登録されている．したがって，日本語 WordNet を用いて英単語を日本語に翻訳した際に，日本語 WordNet から得た単語の語尾が「い」で終わっていなければ「な」を付けることで，形容動詞化して Grammatical Framework で扱うことができるようにする．その後，Grammatical Framework の語彙データの形式に整形する．

3.2.4　副詞の自動生成

Grammatical Framework の日本語の語彙データにおける副詞に必要な情報は，単語のみである．日本語 WordNet を用いて英単語を日本語に翻訳し，Grammatical Framework の語彙データの形式に整形する．

3.3　同じ意味の単語の併合

同じ意味の単語は，日本語 WordNet の synset を用いて一つにまとめる．日本語，英語の語彙データを入力とする．日本語 WordNet で単語を検索し，同じ synset を共有する単語があればそれらをひとまとまりにする．なければ次の単語を検索する．

3.4　実装の現状

Grammatical Framework の標準ライブラリの日本語の語彙の数は 348 であるが，本研究のシステムを用いることで，単語数が 84783 にまで増えた．また，語彙が増えたことにより，翻訳の幅も広がった．例えば Grammatical Framework の標準ライブラリの語彙データにおいて，“airplane” に対する日本語が「飛行機」しか存在しなかったため，“it is an airplane.” の翻訳結果が「これは飛行機です」のみであった．本実装により語彙が増え，また，同じ意味の単語を同列に扱うことができるようになったため,「航空機」や「銀翼」等も扱うことができるようになった．つまり，“it is an airplane.” という文章を翻訳すると,「これは飛行機です」だけではなく,「これは航空機です」や「これは銀翼です」など翻訳結果が増えた．

4 関連研究

Verma らは，WordNet を用いて Universal Network Language と呼ばれる，機械翻訳用の中間言語の語彙を自動的に生成する手法を提案している [5]．これは WordNet を用いて Universal Network Language と呼ばれる，機械翻訳用の中間言語の語彙を自動的に生成するというものである．WordNet を利用することで，synset の上位語を用いて，単語が持つ意味的，構造的な属性を付与した語彙を自動的に生成する．例えば，“crane” という英単語には，動物の「鶴」という意味と機械の「クレーン」の 2 つの意味があり，これらを区別するために WordNet を用いて上位語の概念におけるすべての先祖を検索し，単語に属性を付与する．この研究における対象言語は，英語，ヒンディー語，マラーティー語である．本研究と WordNet の上位語の概念を利用しているところは共通しているが，Grammatical Framework の語彙データを生成すること，また，対象言語が英語と日本語という点で異なる．

また，日本語の語彙を自動生成する研究に，柴田らによる Web テキストから語彙知識を自動獲得する手法 [6] がある．これは Wikipedia や大規模 Web テキストから語彙を生成する手法であり，本研究は同じ意味を持つ異なった語彙を扱うために，日本語 WordNet を用いて語彙を生成している点で異なる．

5 おわりに

本研究では，日本語 WordNet を用いて Grammatical Framework の語彙データを自動生成する手法を提案した．Grammatical Framework の標準ライブラリの日本語の語彙の数はおよそ 350 であるが，本研究のシステムを用いることで単語数が 80000 以上にまで増え，1 つの単語が複数の単語に対応するような語彙データの作成に成功した．

しかし，単語数が増えたことにより，文を翻訳した結果が爆発的に増えるという問題が出てくる．これらは共起辞書等を利用することにより改善できるものと期待されるが，具体的な手法や評価については今後の課題である．

参考文献

[1] Aarne Ranta. Grammatical Framework. *Journal of Functional Programming*, Vol. 14, No. 2, pp. 145–189, March 2004.

[2] Hitoshi Isahara, Fransis Bond, Kiyotaka Uchimoto, Masao Utiyama, and Kanzaki Kyoko. Development of the Japanese WordNet. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, May 2008. European Language Resources Association (ELRA).

[3] George A. Miller. WordNet: A Lexical Database for English. *Commun. ACM*, Vol. 38, No. 11, pp. 39–41, November 1995.

[4] Francis Bond and Ryan Foster. Linking and Extending an Open Multilingual WordNet. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1352–1362, Sofia, Bulgaria, August 2013. Association for Computational Linguistics.

[5] Nirtin Verma and Pushpak Bhattacharyya. Automatic Lexicon Generation through WordNet. *Second International WordNet Conference*, pp. 226–233, 2004.

[6] 柴田知秀, 村脇有吾, 黒橋禎夫, 河原大輔. 実テキスト解析をささえる語彙知識の自動獲得. 言語処理学会 第 18 回年次大会, pp. 81–84, 2012.