「関係」タグ付きコーパスの作成

河原 大輔
京都大学 情報学研究科
kawahara@pine.kuee.kyoto-u.ac.jp

黒橋 禎夫
東京大学 情報理工学系研究科
科技団 さきがけ研究21
kuro@kc.t.u-tokyo.ac.jp

橋田 浩一
産業技術総合研究所 CARC
科技団 CREST
hasida.k@aist.go.jp

1 はじめに

本稿では、文章中の単語間の様々な関係をタグ付けしたコーパスの作成について述べる。このコーパスで対象とする関係は、用言・サ変名詞に対する格関係、名詞間の関係、および共参照である。このような関係を計算機で自動的に認識することが、機械翻訳、情報検索、自動要約などの言語処理システムを高度化していくために必要である。

文章中の単語間の関係にはどのようなものがあって、それをどのようにタグ付けすればよいかというスペックは、実際のデータに基づいてある程度大規模に考える必要がある。しかし、これまで大規模な調査は行われてなく、関係を付与したコーパスとしても、これらの関係の一部を付与したものが存在する程度である[1, 2, 3]。本論文では、これらの関係を付与したコーパス作成において策定したスペックと明らかになった問題を報告する。

2 作成するコーパスの概要

文章中の単語間の様々な関係タグを京都大学テキストコーパス[4]に付与する。京都大学テキストコーパスとは、毎日新聞約40,000文に構文情報が付与されたコーパスである。今回のタグ付けは、格・省略解析を行った結果を修正することによって行う。タグ付けツールは、京都大学テキストコーパスの作成で用いられたツールを、様々な関係を付与できるようにしたものを用いる(図1)。

タグ付けは記事単位で行う。タグを付与する対象の単位は単語である。京都大学テキストコーパスは文節を単位としているので、それを単語単位に分割し、単語間の係り受けは隣の単語に係るとした。単語間の係り受けが誤っていれば、作業者が修正を行う。

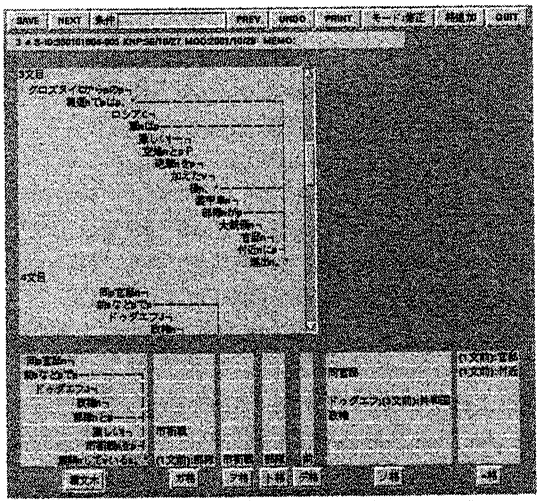


図1: タグ付けツール

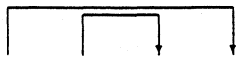
タグは、関係をもつ単語とその関係名からなり、格関係、名詞間の関係、共参照の3種類である。それぞれについて以下で説明する。

2.1 格関係

格関係とは、用言と格要素との間の関係である。この関係は、ガ格、ヲ格、ニ格といった表層格によって表す。関係を詳細に分類した深層格を用いることにすると、タグを付与する際に、最適な関係を選ぶことが難しかったり、最適な関係がない場合に関係の分類の見直しが必要となり、タグ付け作業を進めることが難しくなるためである。

格関係は、格要素が用言に対して直接係り受け関係をもっているものと、格要素が省略されているものがある。格要素が用言に対して直接係り受け関係をもっ

ているものは、2種類あり、ひとつは格が表層に現れているもので、もう一方は格が現れていないものである。格が表層に現れない格要素は、被連体修飾詞、および、「は」「も」などの係助詞が付属する格要素である。



(1) 太郎は 新聞を 読んで 捨てた。

読んで ⇐ ガ格：太郎
ヲ格：新聞
捨てた ⇐ ガ格：太郎
ヲ格：新聞

この例では、まず、「読んで」に「ヲ格:新聞」というタグをつける(自動解析結果についている)。「太郎」が「捨てた」に係っているので、「太郎」は「捨てた」と関係をもつが、係助詞「は」があるために格が表層に現れていない。この場合、「太郎が 捨てた」ので、「捨てた」に「ガ格:太郎」とタグをつける。「読んだ」のガ格、「捨てた」のヲ格は省略されており、「読んだ」に「ガ格:太郎」、「捨てた」に「ヲ格:新聞」とタグをつける。これらのタグ付けは、自動解析結果が誤っていれば修正するという作業になる。省略の指示先として、同じ対象を表す表現が複数あるときは、構造的にもっとも近いものを選択する。

格関係は用言だけでなく、サ変名詞、動作的な名詞にもつける。

(2) ロシア側は 首都 [制圧] の最終段階に入った。

制圧 ⇐ ガ格：ロシア側
ヲ格：首都

例(3)では、「違い」を「違う」という用言で考えることができ、この文の解釈は「製品が他社と違う」となるので、タグは次のようにつける。

(3) わが社の 製品 は 他社 とあまり [違い] がない。

違い ⇐ ガ格：製品
ト格：他社

2.2 名詞間の関係

関係をもつ名詞間にタグを付け、その関係名は「ノ格」とする。

(4) 太郎 は背が低い。しかし、[妹] は背が高い。

妹 ⇐ ノ格：太郎

「妹」は「太郎の妹」という意味なので、このようなタグをつける。日本語のノ格は多くの意味をもつが、表層格を関係として用いる理由と同様に、このプロジェクトではすべてノ格でタグ付けする。

上の例の「妹」は典型的な関係名詞で、他の名詞(この場合「太郎」)との関係によって自分自身の意味が成り立っている。関係名詞ほど強い関係でないにしろなんらかの関係をもつ名詞間にもタグを付与する。例えば、「車の ハンドル」「窓の カーテン」のような名詞間の関係をノ格でつける。ただし、「赤色の 車」「3個の みかん」など修飾的な関係をもつ名詞間にはタグをつけない。

2.3 共参照

代名詞、指示詞などの共参照表現にタグをつける。関係名は「=」とする。

(5) 太郎 は太っている。[彼] はいつも何か食べている。

彼 ⇐ =：太郎

代名詞、指示詞だけでなく、まったく同じ表記を含め、共参照している名詞にタグをつける。

(6) 奈緒美 が来た。あの [子] はいつも長居する。

子 ⇐ =：奈緒美

2つの名詞が、上位/下位関係や総称/非総称などの関係をもち、完全に同一でない場合は、= ではなく「≒」でタグ付けする。

(7) 車 の販売台数をみると、[自家用車] は…

自家用車 ⇐ ≒：車

「車」と「自家用車」は上位/下位関係にある。

(8) 小さな パソコン$_1$ が売れている。太郎の パソコン$_2$ は小さい。しかし、花子の パソコン$_3$ は古くて大きい。

パソコン$_2$ ⇐ ≒：パソコン$_1$
パソコン$_3$ ⇐ ≒：パソコン$_1$
パソコン$_3$ ⇐ ≒：パソコン$_2$

パソコン$_1$ は一般的なパソコンを意味し、パソコン$_2$ とパソコン$_3$ は具体的なものを表しているので、総称/非総称の関係となっている。パソコン$_2$ と パソコン$_3$ は具体的なものとしては異なるが、このパソコンに関する文章中で関係しているので、これも ≒ でタグ付けする。

2.4 スケジュール

このプロジェクトは2001年9月から始め、現在まで、2人の作業者が同じ記事にタグ付けを行いながら、タグ付けのスペックの策定を行ってきた。現在、85%程度の agreement がとれ、スペックがほぼ定まったので、本格的に作業を始めたところである。これまで、1200文へのタグ付けが終わっており、1時間で11文程度のタグ付けが可能となっている。現在、作業は2人で行っており、京都大学テキストコーパス全文のタグ付けを目標としている。

3 タグ付けの複雑な問題

タグ付けを行う際の複雑な問題を以下に挙げる。

3.1 タグ付け単位の問題

タグをつける単位は単語であるが、日本語では単語の概念が曖昧で、1語の長さが捉え方によって異なる。例えば、「訪日」は1語であったが、「日(本)を 訪(れる)」のような関係があるのでタグをつける必要がある。このような場合は単語を分割してタグをつけることにした。「訪日」の場合は、「訪」「日」に分割し、「訪」に対して「ヲ格:日」とタグ付けする。ほかに単語を分割した例として、「朝鮮人」→「朝鮮」「人」、「日伯」→「日」「伯」などがある。

3.2 タグの AND/OR

格要素が並列である場合は、並列になっている要素をすべて記述する。

(9) 太郎 と 花子 が学校から 帰った 。

帰った ⇐ ガ格：太郎, 花子 [and]

「太郎」と「花子」は並列であるので、"and" で両方の要素をとることを示す。

並列ではなく、用言がト格をとる場合は、上記のようには扱わない。

(10) 花子 と 太郎 が 結婚した 。

結婚した ⇐ ガ格：太郎
　　ト格：花子

この例では「結婚した」がト格をとる。

関係をもっている要素を一意に決められず、いくつかあるうちのいずれにもとれる場合には、そのすべてを記述する。

(11) 高知県 の 橋本知事 は国籍条項を 撤廃する 方針だ。

撤廃する ⇐ ガ格：高知県, 橋本知事 [or]

「撤廃する」のガ格は「高知県」、「橋本知事」のいずれにも解釈できるので、"or" でどちらでもよいことを示す。ただし、この "or" は解釈の "or" を意味し、文中の要素間の論理的関係を表しているわけではない。

(12) 田園調布 か 国立 に 住みたい 。

住みたい ⇐ ニ格：田園調布, 国立 [and]

3.3 連体修飾

連体修飾において、被連体修飾詞が連体修飾節の用言に対して格関係をもたない場合がある。この場合の関係名を「外の関係」とし、用言に対してタグをつける。また、被連体修飾詞に対して「トイウ格」またはノ格のタグをつける。トイウ格は、被連体修飾詞が連体修飾節の内容を表す場合に用い、ノ格は、被連体修飾詞が「の」を使って言い換えることができる場合に用いる。

(13) 政治家が賄賂を 受け取った 事実

受け取った ⇐ 外の関係：事実
事実 ⇐ トイウ格：受け取った

この例で、「事実」は「受け取ったという事実」を意味しており、「事実」と「受け取った」の関係はどのような表層格でも表すことができない。このような関係を外の関係とする。また、「事実」は内容を表しているのでトイウ格のタグをつける。

(14) 花子が旅行に 出かける 前日

出かける ⇐ 外の関係：前日
前日 ⇐ ノ格 ：出かける

「前日」は、「出かける日の前日」という意味であり、「出かける」に対して相対的な関係にある。この場合も外の関係であり、「の」を使って言い換えられるので、「前日」にノ格のタグをつける。

(15) 塩と水を 混ぜた 食塩水

混ぜた ⇐ 外の関係：食塩水
食塩水 ⇐ ノ格 ：混ぜた

「塩と水を混ぜた結果の食塩水」という意味であり、外の関係となる。「食塩水」にはノ格でタグをつける。

3.4 不特定:人

日本語では、用言の動作主体が省略されることが多い。そのなかでも、動作主体の省略が、具体的な対象を指さずに不特定の人々を指している場合が多々ある。そのような場合には、「不特定:人」というタグを付与する。

(16) 用途を限った専用計算機としても、世界最速だと [いう]。

いう ⇐ ガ格：不特定:人

「いう」のガ格は、特に具体的な対象を指しているわけではない。不特定の人々を指している、「いう」「いえる」などのガ格、「言われる」「見える」「見られる」「聞こえる」などのニ格はこのタグを与える。

(17) 地方公務員法では日本国籍がない人の [任用] を禁じる [規定] はない。

任用 ⇐ ガ格：不特定:人
規定 ⇐ ガ格：不特定:人

この例の「任用」、「規定」のガ格はいずれも不特定の人々であると考えられる。また、「規定」は動作的な意味ではなく「規定」という動作の結果を表しているおり、ヲ格について考えることには意味がないので、ヲ格のタグは付与しない。

3.5 一人称

「思う」「思える」などのガ格、「思われる」「考えられる」などのニ格が、著者である場合「一人称」というタグを与える。

(18) 妥当な考えだと [思う]。

思う ⇐ ガ格：一人称

4 タグ付け例

以下に文のタグ付け例を挙げる。例 (19) では、複合名詞「首都制圧」にタグをつける必要がある。例 (20) では、2 文目の「事実」は、前文の内容を受けているので、タグを与える必要がある。

(19) ロシア側は首都制圧の最終段階に入った。

制圧 ⇐ ガ格：ロシア
　　ヲ格：首都
入った ⇐ ガ格：ロシア
　　ニ格：段階

(20) 政治家が賄賂を受け取った。その事実は…

受け取った ⇐ ガ格　：政治家
　　ヲ格　：賄賂
事実 ⇐ トイウ格：受け取った

5 おわりに

本論文では、文章中の様々な関係をタグ付けしたコーパスの作成について述べた。本コーパスで対象とする関係は、用言・サ変名詞に対する格関係、名詞間の関係、および共参照である。このような関係を付与したコーパスは、機械翻訳、情報検索、自動翻訳などの言語処理システムを高度化するために必須のリソースであり、また、GDA (http://i-content.org/GDA/)、MPEG-7 (http://mpeg.telecomitalialab.com/)、Semantic Web (http://www.semanticweb.org/) などによるエンドユーザコンテンツにも利用できる。

謝辞

本コーパスのタグ付け作業に協力してくださいました石川真奈見氏、堀内マリ香氏、玉井陽子氏に心から感謝致します。

参考文献

[1] Daniel Marcu, Estibaliz Amorrortu, and Magdalena Romera. Experiments in constructing a corpus of discourse trees. In *Proceedings of the ACL'99 Workshop on Standards and Tools for Discourse Tagging*, pp. 48–57, 1999.

[2] 竹澤寿幸, 中村篤, 隅田英一郎. ATR の会話音声翻訳研究用データベース. 音声研究, pp. 16–23, 2000.

[3] Massimo Poesio. Annotating a corpus to develop and evaluate discourse entity realization algorithms: issues and preliminary results. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation*, pp. 211–218, 2000.

[4] 黒橋禎夫, 長尾眞. 京都大学テキストコーパス・プロジェクト. 言語処理学会 第 3 回年次大会発表論文集, pp. 115–118, 1997.