京都大学格フレーム Ver 2.0

2019/01/21

概要

格フレームとは、述語とそれが格関係をもつ語(項)を記述したものです。京都大学格フレーム Ver 2.0 は、ウェブから収集した日本語 100 億文から自動構築した格フレームです。格フレームの自動構築は、[林部+ 2015]の手法を上記のウェブテキストに適用することによって行っています。約 11 万個の述語が含まれており、述語あたり平均 5.1 個の格フレームが構築されています。

配布ファイル

README.txt : このファイル
kyoto-univ-web-cf-2.0.xml.gz : 格フレームファイル
(XML 形式; エンコーディング: utf-8; gzip 圧縮: 元サイズは 4.2GB)

格フレームファイルのフォーマット

<caseframedata> : 格フレームファイル全体
<entry> : 1 つの述語のデータ
  headword : 述語表記
  predtype : 述語タイプ
<caseframe> : 1 つの格フレーム (1 つの述語に 1 つ以上存在)
  id : 格フレームの ID
  frequency : 格フレームを構成する用例の頻度
<argument> : 項
  case : 格の名前
  frequency : 格を構成する用例の頻度
<component> : 項を構成する 1 つの用例の表記
  sm : "1"ならば、意味属性に汎化されている
  frequency : 用例の頻度

述語と項の仕様

・述語

述語の情報は、述語タイプと述語表記からなります。

述語タイプは述語の種類を表しており、次の 3 種類のいずれかです。

動 : 動詞
形 : 形容詞
判 : 名詞+判定詞

述語表記は、形態素解析器 Juman++の代表表記で表されます。たとえば、動詞「認める」に対して代表表記「認める/みとめる」が述語表記となります。

「れる」「られる」「せる」「させる」など、格交替を起こす可能性のある付属語が動詞に後続している場合には、それらの付属語が述語表記の一部となり、述語表記が複数の単語の代表表記からなります。また、格交替以外にも、形容詞に「なる」が後続する場合のように動詞化する場合などに、述語表記が複数の単語からなります。これらの場合には、"+"記号によって各単語の代表表記を連結しています。

コーパス中の元の表記に対応する単語が一意に決まらない場合には、"?"記号によってその可能な単語を連結しています。

例）認める/みとめる　　「認める」「みとめる」のどちらの表記でもこの代表表記になる

認める/みとめる+られる/られる	2 つの単語からなる場合
行く/いく+せる/せる	2 つの単語からなる場合
美しい/うつくしい+なる/なる	2 つの単語からなる場合
弾く/ひく?弾く/はじく	曖昧性がある場合

・項

項の情報として、各格フレームの格ごとに用例が記述されています。

格は、ガ格、ヲ格、ニ格、ヘ格、ト格、デ格、カラ格、ヨリ格、マデ格のような表層格です。それ以外の格として、「ガ２」は二重主語構文の外のガ格、「外の関係」は被連体修飾名詞が述語に対して外の関係をもつことを表しています。「修飾」は、副詞などの修飾的な連用要素、「時間」は時間要素をとる格を表しています。また「について」「によって」「として」のような複合辞も格として扱っています。項に係るノ格も収集しており、たとえば「ノ格~ニ格」はニ格の項に係るノ格、「ノ格~判」は述語の「名詞+判定詞」に係るノ格を表しています。

用例は、頻度 2 以上のものを記述しており、基本的には文節内の主辞単語の代表表記で表しています。主辞単語が一文字の漢字である場合には、その直前の単語も連結しています。述語表記と同様に、複数の単語からなる場合には"+"記号、曖昧性がある場合には"?" 記号で連結しています。

用例が数量、時間、補文を表している場合には、意味属性として汎化し、<数量>、<時間>、<補文>と記述しています(XML データ中では、"<"と">"は"<"と">"にエスケープされており、sm 属性は"1"です)。さらに、数量に助数辞がついている場合には、「<数量>個/こ」のように助数辞の代表表記を付加しています。

例) 子供/こども　　「子供」「子ども」「こども」のいずれの表記でもこの代表表記になる

交差/こうさ+点/てん	2 つの単語からなる場合
ボタン/ぼたん?牡丹/ぼたん	曖昧性がある場合
駐車/ちゅうしゃ+場/じょう?場/ば	2 つの単語からなり、曖昧性もある場合

参考文献

河原大輔, 黒橋禎夫.
格フレーム辞書の漸次的自動構築,
自然言語処理, Vol.12, No.2, pp.109-131, 2005.

河原大輔, 黒橋禎夫.
高性能計算環境を用いた Web からの大規模格フレーム構築,

情報処理学会 自然言語処理研究会 171-12, pp.67-73, 2006.

Daisuke Kawahara and Sadao Kurohashi.
Case Frame Compilation from the Web using High-Performance Computing,
In Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC2006), pp.1344-1347, 2006.

林部祐太, 河原大輔, 黒橋禎夫.
格パターンの多様性に頑健な日本語格フレーム構築,
情報処理学会 自然言語処理研究会 224-14, pp.1-8, 2015.

更新情報

・Ver 1.0 からの変更点

- 格フレーム構築のために用いるコーパスを 16 億文から 100 億文に増やした。

- 格フレーム構築手法として[林部+ 2015]を用い、クラスタリングの精度を向上させた。

- 受身や使役の述語を表すために、態タイプ(P, C など)を使っていたが、これを廃止し「れる/れる」「せる/せる」などを述語表記に付加することによって表現することにした。

- ノ格について、どの格に係っているノ格かを区別するようにした。

謝辞

京都大学格フレーム Ver 2.0 の構築には、科学技術振興機構 CREST 研究領域「ビッグデータ統合利活用のための次世代基盤技術の創出・体系化」「知識に基づく構造的言語処理の確立と知識インフラの構築」（研究代表者: 黒橋禎夫)の助成を受けました。

連絡先

本格フレームに関するご意見、ご質問は nl-resource@nlp.ist.i.kyoto-u.ac.jp
宛にお願いいたします。

----------以上