Formulaic sequence annotation guidelines, Japanese supplement

Tim Baldwin

September 9, 2016

アノテーションの対象は「Formulaic Sequence」（FS）ですが、大まかに説明すると一般的にネーティブ話者にとって聞き覚えがありそうな複数の形態素からなる表現です。複合表現（MWE）、すなわち何かしらの非合成ある表現がその一部で、「機械翻訳」、「出来上がる」、「腕を上げる」などが全部FSとなります。「named entity」（NE）は、一般的に知られているものであればFSとなります。例えば「東京オリンピック」、「代々木公園」、「村上春樹」などはFSで、「山田英彦」「宇佐美公園」などはFSにはなりません。それ以外では、「じゃあね」、「　を傍らに」、「黒ごま」、「　ざるを得ない」もFSになります。FSかどうかの判定の手がかりの一つとしては、英語などへの翻訳において、英語には類似表現がないか、直訳ではおかしなものになるか、英訳も定形した表現になるか、のいずれかが見られがちです。

アノテーションのしかたとしては、FS候補をlemmaとUniDicによる複数の形態素として表示し、その下には、ランダムにサンプリングされた5つ程の例文を表示します。FSとして認定するためには、少なくとも5つの例文の中の1つが実際のFSの使われ方にならなくてはいけません。アノテーションは以下の4通りになります。

- 「Is a canonical formulaic sequence」
  FS候補がFSに完全一致した場合
- 「Recalls a formulaic sequence, but not canonical」
  FS候補が部分的にFSに一致しているか、FSの一部になっている場合
- 「Is not formulaic sequence」
  FS候補がFSあるいはFSの一部になっていない場合

- 「Encoding error」
  FS候補が文字化けしているため、FSかどうかの判定が不可能な場合

  FS候補が 2 種類ありますが、まず一つ目は「contiguous FS」（連続FS）で、普通の使われ方ではFSが連続して、一つにまとまったものです。 2 つ目は「gappy FS」で、FSの中には名詞句など含まれるものです。gappy FSの例としては「決して　ない」、「総合〜会議」が上げられます。それぞれのFSのアノテーションには個別にインタフェースを用意しています。

  FSの定義とアノテーションの過程を詳しく書いたものを次のように用意しましたが、英語のみになっているので、あらかじめご了承ください。