

## 「日英新聞記事対訳コーパス一式」仕様書

2018年10月31日 読売新聞東京本社

◇本データは「読売新聞」の日本語と「The Japan News」の英語の新聞記事データに対し、自動的に記事対応を取り、さらに自動的に対応記事内の文対応を取ったものである。全て自動処理のため、対応付けに誤りが含まれることに注意。

◇文字コードは UTF-8、改行コードは LF となっている。

◇データは 2 種類ある。

1. 記事 ID 対応データ (ファイル名: \*.id\_pair.tsv)  
英語記事に対応する日本語記事を ID で示したデータ。
2. 文対応データ (ファイル名: \*.parallel.txt)  
対応する記事内の文対応を抽出したデータ。

◇記事対応データのフォーマットは以下の通り。

英語記事 ID(C0 タグの値)<TAB>日本語記事 ID(C0 タグの値)

◇文対応データのフォーマットは以下の通り。

# 英語記事 ID(C0 タグの値)=日本語記事 ID(C0 タグの値)- JE-数字 score=文対応スコア  
equivalent=0/1  
en: 英語文  
ja: 日本語文

\* 「-JE-数字」について

記事内の対応ごとに連番で付与される。

\* 文対応スコアについて

文対応スコアは対応の程度を表す指標で、「1」に近いほど対応が高い。0 のものは対応する文が存在しないことを示しており、この場合 en もしくは ja のどちらかが含まれてい

ない。

\* 複数文対応について

日本語、英語とも連続する複数文が対応すると判断した場合には、ja1, ja2 や en1, en2 などのように示してある。不連続なものは出現しない。この場合の score は複数の文をつなげたものの数値となる。各言語最大で 3 文までつなげられる可能性がある。

\* equivalent について

文の対応付けは自動で行なっているため、文対応スコアが高くても対訳になっていないケースも存在する。そこで機械学習を用いて各対応の内容が同一かどうかを判定した。その結果、同一と判定されたものには equivalent=1 を、そうでないものには 0 を付与した。なお equivalent は文対応スコアが 0.2 以上のもののみで付与されている。

例)

# 20101127TDY06T13= 2 0 1 0 1 1 2 6 T Y M 1 0 A F            0 0 3 -JE-18  
score=0.476744186046512 equivalent=1

en1: In contrast, domestic shipments of mobile phones have been declining, after hitting a peak in 2007.

en2: In 2014, they are predicted to drop as low as about 60 percent of the peak.

ja: 一方で、国内の出荷台数は 07 年をピークに縮小傾向にあり、14 年はピーク時の約 6 割に落ち込む見込みだ。

◇本コーパスは京都大学情報学研究科と読売新聞との共同研究において、読売新聞が保有する新聞記事データから自動的に抽出され、構築されたものである。