「読売新聞記事　日英文対応コーパス」データ構築条件

1. 各英語記事(★)に対し、最大 60 日前までの全ての日本語記事データの中から、対訳辞書を用いて出現する単語の傾向が似ている日本語記事を最大 20 記事取得しました。

2. 取得した各日本語記事と★との間で文対応付けプログラムを実行し、文対応を抽出すると同時に記事対応スコアを算出しました。

3. 記事対応スコアが最も高い記事ペアと 2 番目に高い記事ペアのスコアの差が 0.5 以上になった場合、スコアが最も高い記事ペアが信頼できると判断し、文対応抽出結果を利用しました。

＊各年度ごとの文対応数の分布

文対応スコアが 0.2 以上のものならば、およそ対訳文として利用可能と判断できるので、抽出された対応数は 729,167 ペアとなります。

なお例えば 2016 年 1 月の英語記事は、対応する日本語記事が 2015 年 12 月の可能性もありますが、英語記事の書かれた日付を元にカウントしてあります。

文対応スコア	0.5 以上	0.4 以上	0.3 以上	0.2 以上	0.1 以上	0 より上
2006 年	25794	45062	59223	68428	73213	74405
2007 年	29050	49099	63826	73348	78252	79488
2008 年	26418	46426	61283	70653	75299	76559
2009 年	26381	44656	57619	65694	70080	71270
2010 年	20502	34743	45085	51495	54972	55925
2011 年	25937	41962	53337	60235	63828	64861
2012 年	29476	45959	56301	61927	64600	65451
2013 年	31853	49274	60680	67154	70401	71386
2014 年	35322	54963	68462	76171	80035	81142
2015 年	45470	68682	84125	92951	97339	98712
2016 年	20254	30300	36934	41111	43210	43883
合計	316457	511126	646875	729167	771229	783082