2021 年 6 月

「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ます。このときの抽出された対応数は 2020 年の場合 74424 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	11759	28284	45294	57312	63494	65249
2007 年	13292	31044	49215	61724	68480	70395
2008 年	12166	29303	47662	60557	67262	69185
2009 年	12228	28631	45388	56504	62301	64091
2010 年	10087	23152	36166	44880	49455	50939
2011 年	13021	27829	42194	51566	56392	57819
2012 年	15059	31279	46164	55320	59711	60904
2013 年	16248	33859	49912	59792	64498	65838
2014 年	18820	39333	57938	68981	74357	75877
2015 年	25756	51208	72949	85715	91921	93634
2016 年	12468	23537	32972	38624	41527	42368
2017 年	24108	43981	59922	68932	73423	74794
2018 年	22933	43035	60089	69707	74043	75266
2019 年	18453	34497	48032	55724	59094	60065
2020 年	24216	46241	64304	74424	79099	80425
合計	250614	515213	758201	909762	985057	1006849