

=====Test Data for NLPCC 2018 Shared Task 2:GEC=====

1. Files in this category:

--source.txt: contains the original sentences. One sentence per line.

--seg.txt: contains the segmented sentences. One sentence per line.

2. Submissions:

Segmented sentences in which all the grammatical errors have been corrected by your system. One sentence per line, aligned to the input sentence.

For example:

Original sentence:

我在家里一个人学习中文。

Segmented sentence:

我 在 家 里 一 个 人 学 习 中 文 。

Submission:

我 在 家 里 自 学 中 文 。

3. Evaluation:

MaxMatch(M2) Scorer

--computes the sequence of phrase-level edits between a source sentence and a system hypothesis that achieves the maximal overlap with the gold standard.

The precision, recall, and F_1 measure between the set of system edits $\{e_1, \dots, e_n\}$ and the set of gold edits $\{g_1, \dots, g_n\}$ for all sentences are computed as following formulas:

$$P = \frac{\sum_{i=1}^n |e_i \cap g_i|}{\sum_{i=1}^n |e_i|}$$

$$R = \frac{\sum_{i=1}^n |e_i \cap g_i|}{\sum_{i=1}^n |g_i|}$$

$$F_1 = 2 \times \frac{P \times R}{P + R}$$

where the intersection between e_i and g_i is defined as:

$$e_i \cap g_i = \{e \in e_i | \exists g \in g_i (\text{match}(e, g))\}$$