**Supplemental file 1.** Size of HUNER sub-corpora. Each line contains a single token. Entities were quantified by counting B tags.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Sub-corpus** | **Training set lines** | **Development set**  **lines** | **Test set lines** | **Training set entities** | **Development set entities** | **Test set entities** |
| Cell | 578083 | 98337 | 279770 | 3062 | 555 | 1572 |
| Chemical | 2972895 | 486630 | 1495133 | 114579 | 18391 | 56597 |
| Disease | 559063 | 89250 | 270410 | 18113 | 2978 | 8896 |
| Gene | 2882762 | 474002 | 1406244 | 93280 | 15823 | 47135 |
| Species | 893834 | 150478 | 393562 | 7098 | 1050 | 3683 |