

Model	NonOOV	OOV	Overall
Word NMT + UNK replacement	27.61	21.57	26.17
Hybrid model	<b>29.36</b>	25.92	28.49
Nested Attention Hybrid Model	29.00	<b>27.39</b>	<b>28.61</b>

Table 5:  $F_{0.5}$  results on the CoNLL-13 set of main model architectures, on different segments of the set according to whether the input contains OOVs.

source	This greatly <b>violets</b> the rights of people .
gold	This greatly <b>violates</b> the rights of people .
word NMT + UNK replacement	This greatly <b>violets</b> the rights of people .
Nested Attention Hybrid Model	This greatly <b>violates</b> the rights of people .

Table 6: An example sentence from the OOV segment where the nested attention hybrid model improves performance.

Table 6 shows an example where the nested attention hybrid model successfully corrects a misspelling resulting in an OOV word on the source, whereas the baseline word-level system simply copies the source word without fixing the error (since this particular error is not observed in the parallel training set).

## 5.2 Impact of Nested Attention on Different Error Types

To analyze more precisely the impact of the additional character-level attention introduced by our design, we continue to investigate the OOV segment in more detail.

The concept of *edit*, which is also used by the official M2 score metric, is defined as a minimal pair of corresponding sub-strings in a source sentence and a correction. For example, in the sentence fragment pair: “Even though there is a risk of causing **harms** to someone, people still **are prefers** to keep their pets without a leash.” → “Even though there is a risk of causing **harm** to someone, people still **prefer** to keep their pets without a leash.”, the minimal edits are “harms → harm” and “are prefers → prefer”. The  $F_{0.5}$  score is computed using weighted precision and recall of the set of a system’s edits against one or more sets of reference edits.

For our in-depth analysis, we classify edits in the OOV segment into two types: *small changes* and *large changes*, based on whether the source and target phrase of the edit are orthographically similar or not. More specifically, we say that the target and

Model	Performance		
	P	R	$F_{0.5}$
<b>Small Changes Portion</b>			
Hybrid model	43.86	16.29	32.77
Nested Attention Hybrid Model	48.25	17.92	36.04
<b>Large Changes Portion</b>			
Hybrid model	32.52	8.32	20.56
Nested Attention Hybrid Model	33.05	8.11	20.46

Table 7: Precision, Recall and  $F_{0.5}$  results on CoNLL-13, on the “small changes” and “large changes” portions of the OOV segment.

source phrases are orthographically similar, iff: the character edit distance is at most 2 and the source or target is at most 8 characters long, or  $edit\_ratio < 0.25$ , where  $edit\_ratio = \frac{character\_edit\_distance}{\min(len(src), len(tar)) + 0.1}$ ,  $len(*)$  denotes number of characters in  $*$ , and  $src$  and  $tgt$  denote the pairs in the edit. There are 307 gold edits in the “small changes” portion of the CoNLL-13 OOV segment, and 481 gold edits in the “large changes” portion.

Our hypothesis is that the additional character-level attention layer is particularly useful to model edits among orthographically similar words. Table 7 contrasts the impact of character-level attention on the two portions of the data. We can see that the gains in the “small changes” portion are indeed quite large, indicating that the fine-grained character-level attention empowers the model to more accurately correct confusions among phrases with high character-level similarity. The impact in the “large changes” portion is slightly positive in precision and slightly negative in recall. Thus most of the benefit of the additional character-level attention stems from improvements in the “small changes” portion.

Table 8 shows an example input which illustrates the precision gain of the nested attention hybrid model. The input sentence has a source OOV word which is correct. The hybrid model introduces an error in this word, because it uses only a single source context vector, aggregating the character-level embedding of the source OOV word together with other source words. The additional character-level attention layer in the nested hybrid model enables the correct copying of this long source OOV word, without employing the heuristic mechanism of the word-level NMT system.