	De-En		En-Es	
	BLEU [%]	TER [%]	BLEU [%]	TER [%]
Baseline	30.6	49.6	32.7	52.6
+self-training	31.4	48.1	35.6	49.1
+sent-sBLEU	31.4	48.1	36.0	48.4
+sent-binary	31.6	47.8	36.2	47.6
+chunk-match	32.2	47.0	37.9	45.4
+chunk-lcs	**32.3**	**46.5**	**38.8**	**44.5**

Table 1: Chunk-level feedback compared to sentence-level feedback. *Self-training* is equivalent to having no feedback or setting all $w_i = 1, \forall i$ in the training objective in Eq. (2). *sent-sBLEU* and *sent-binary* are sentence-level methods with sentence BLEU and binary weighting rules, defined as in Section 4.2. *chunk-match* and *chunk-lcs*-level feedback refers to assigning w_i using simple matching or LCS method described in Section 3.

En and En-Es, respectively. We also note a significant improvement of 1.3% and 3.1% in TER. Chunk-based approach based on simple matching also outperforms sentence level methods, but not by as much as lcs-based, which suggests that this method benefits more from consecutive segments, rather than single correct words.

We believe that the success of the partial feedback approach can be explained by the fact that often a sentence can be split into chunks which can be translated independently of the context. Reinforcement of the correct translation of such a chunk in one training example seems to positively affect translations of such chunks in other, different sentences. By focusing on the good and masking out erroneous chunks, partial feedback acts as a precise noise reduction method.

We have also trained the models using fine-tuning (Luong and Manning, 2015) on the reference target in-domain data, which further improved translation by 2% and 3.8% BLEU on De-En and En-Es compared to using chunk-based feedback. We note that by using partial feedback we are able to recover between 30% and 45% of improvements that come from in-domain adaptation.

4.3 Robustness

The proposed artificially generated partial feedback is very precise as it does not introduce any

#		De-En		En-Es	
		BLEU [%]	TER [%]	BLEU [%]	TER [%]
1	Chunk-level feedback	32.3	46.5	38.8	44.5
Under selection ratio:					
2	25%	32.2	47.0	38.9	45.0
3	50%	31.9	47.4	38.1	45.6
4	75%	31.4	47.9	36.7	46.7
Incorrect selection ratio:					
5	10%	32.0	47.2	38.1	44.9
6	25%	31.5	47.9	37.2	46.9
7	50%	30.9	48.8	35.6	50.0
8	#2 + #5	31.6	47.7	38.1	45.5

Table 2: Impact of user errors on the translation performance. *Under selection ratio%* indicates on average what percentage of words in a correct chunk have not been selected in user simulation, but all selected words are correct. *Incorrect selection ratio%* indicates what percentage of words are incorrectly selected, here the total number of marked words is the same as in chunk-level feedback. In the last row, 10% of marked words are actually incorrect and the total number of marked words is 25% less compared to system in row 1.

type of noise in marking of good chunks. For example, on the En-Es dataset artificial methods mark 40% of all words as correct. However, a user might not mark all the correct words in a sentence, but select only a few.

Furthermore, artificially generated partial feedback does not contain noise, given that the reference translation is adequate. However, users may make mistakes in selection. We differentiate two types of errors that a user can make: under selection, when a correct word was not marked; and incorrect selection, when an incorrect word was marked as correct.

To anticipate the impact of these mistakes we experiment with deliberately impairing the feedback in Table 2. We see that randomly dropping 25% of the selection has very little effect on the model, while dropping 50% and more decreases the translation performance significantly, yet still performing at the same level or better than self-training system.

When selection contains noise, the impact already becomes noticeable at 10%. Increasing the amount of noise up to 25% decreases the perfor-