Data selection: The first experiment

I select the first 1 million sentences for training. The corpus is small enough for my first experiment with EMS (Experiment Management System). The corpus is also big enough, 17% of the whole corpus, for training an SMT system.

	Lang8-big.head1M	Lang8-big	NUCLE	Lang8-big + NUCLE
Size(sentence)	1M	5.75M	57K	-
Size (Word)	11.9M	70.9M	1.16M	-
OOV	-	-	-	-
Precision	0.2481	0.2323	0.3195	0.5419
Recall	0.2997	0.3194	0.2261	0.2290
F-0.5	**0.2570**	**0.2457**	0.2951	0.4256

I have not examined the OOV rate of the training data yet. #TODO

The result shows that the performance of the first 17% is on par (slightly better) with the performance of the whole lang8-big corpus. In my opinion, if we select a subset in a clever way, we could outperform the whole dataset.

Nevertheless, the large corpus is preferable to combine with NUCLE corpus, as it covers a wider range of vocabulary. Hence, this experiment needs a lot of analysis for a final conclusion. #TODO

Next, I write a naïve filter, which filters out all the obvious noise in the corpus (all sentence pairs with length ratio greater than 2 or overlapping ratio smaller than 0.2). It filters out 50K sentences of the selected corpus. The percentage of sentences being filtered is 5%. The results are as follows:

Precision : 0.2559

Recall : 0.3093

F_0.5 : **0.2651**

From my point of view, this is a promising result. However, we need futher analysis for a clear explanation of the improvement. #TODO