Appendix

To investigate whether our approach is effective for English NER, we conducted experiments on the CoNLL-2003 dataset. As the FLAT model used in the paper incorporates lexical information and is specifically designed for Chinese, we used the most common BERT as our baseline method. Table 1 shows the experimental results on English dataset CoNLL-2003. We can observe a slight performance gain after applying our method to BERT in English. However, since the improvement is not evident, we don't have enough confidence to say it does work to improve English sequence labeling NER.

The reason why our method may have varying degrees of impact could be attributed to the natural distinctions between the two languages. In English, the hierarchy to compose an entity is "letter-word-entity", where a word is its minimal semantic unit with space separators explicitly indicating their word boundaries. However, for Chinese, there's no indicator between characters to form word boundaries in its hierarchy "character-word-entity". Therefore, Chinese NER is more difficult due to multiple possible combinations of characters and their resulting ambiguities. Taking "南京市长江大桥" as an example, since it's a character sequence without explicit word boundaries, Chinese NER usually suffers from such ambiguity of "Nanjing Yangtze River Bridge" and "Jiang Daqiao, Mayor of Nanjing". However, if written in English, "Bridge" or "Mayor" can explicitly determine the sentence meaning and accurate NER results. Apart from that, the same character may have totally different semantics when its word boundary changed, e.g. "长" meaning "Long" and "Mayor" in different character combinations of the example. Since our first learning stage could help to learn better representations sharing label-semantics and thus forming clearer entity representation boundaries, it may be especially effective in Chinese.

Model	CoNLL-2003		
	Pr.	Rec.	F1
BERT	89.41	**90.73**	90.07
Nerco(BERT)	**89.88**	90.68	**90.28**

Table 1: Results for English NER on CoNLL-2003.