BERT-based Lexical Substitution: Supplementary Notes

Wangchunshu Zhou[1] * **Tao Ge**[2] **Ke Xu**[1] **Furu Wei**[2] **Ming Zhou**[2]
[1]Beihang University, Beijing, China
[2]Microsoft Research Asia, Beijing, China
zhouwangchunshu@buaa.edu.cn, kexu@nlsde.buaa.edu.cn
{tage, fuwei, mingzhou}@microsoft.com

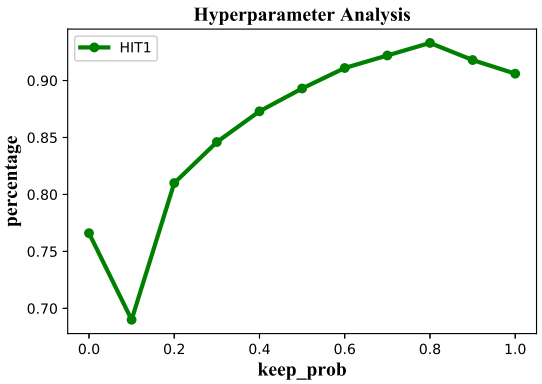

Figure 1: Performance of substitute candidate proposal with respect to different dropout rates.

α	LS07	LS14
our approach (α = 0.01)	**60.5**	**57.6**
α = 0.05	59.8	56.8
α = 0.10	59.1	55.9
α = 0.005	59.4	56.6
α = 0.001	58.5	55.8

Table 1: GAP scores (the higher, the better) on the substitute ranking task with respect to the choice of hyperparameter α.

A Supplemental Material

Using BERT to propose substitute candidates suffers from two major problems. First, as we extract the language model output of a "real" word instead of the `[MASK]` symbol, the probability distribution naturally concentrates at the origin target word and all other words have a probability close to zero. A substitute candidate may have an probability of 0.1%. Second, as described in BERT paper, there is a random replacement rate of 1.5% of all tokens during pre-training. For this 1.5% of probability, the distribution is more uniform and the substitute probability of a gap-filler word is thus on par with valid substitutes. These problems make proposal process unstable by proposing either variants of target word, gap-filler word or random words rather than substitutes. We propose embedding dropout to make the probability distribution higher for valid substitutes, thus benefits the proposal process.

For substitute candidate proposal, we compare our substitute candidate proposal method to the methods based on lexical resources like WordNet and PPDB in terms of substitute candidate proposal. For each instances, we propose 50 candidates. We define HIT score to evaluate the quality of the proposed candidates as follows:

$$\text{HIT} = \sum_{i=1}^{M} \frac{f(H_i, G_i)}{M} \tag{1}$$

$$f(H_i, G_i) = \begin{cases} 1 & \text{if } |H_i \cap G_i| > 0 \\ 0 & \text{else} \end{cases} \tag{2}$$

where H_i and G_i denote the set of the proposed candidates and the set of gold substitutes for the i^{th} test instance respectively.

It is easy to understand HIT. For a test instance, if a system's top 50 hypothesized candidates include a gold substitute, the system will get credit. Table 2 shows the results of HIT of

Resource	LS07	LS14
WordNet	64.7	58.9
PPDB 2.0 XXXL (top 50)	91.2	92.3
ours (top 50)	**94.2**	**96.1**
BERT (Keep target, top 50)	90.3	92.3
BERT (Mask target, top 50)	79.5	81.3

Table 2: HIT results of substitute proposal, higher is better.

*This work was done during the first author's internship at Microsoft Research Asia.

The field is not much different, only a liitle bit *brighter*.		
Valid Substitute	balAddCos	Our Approach
luminous	**light**	**light**
clear	sharp	**clear**
light	smart	brilliant
-	clever	**luminous**
That's not a very high *bar*.		
Valid Substitute	balAddCos	Our Approach
marker	indicator	**hurdle**
level	**hurdle**	**barrier**
barrier	**barrier**	indicator
hurdle	pub	**level**
it should not *take* that long.		
Valid Substitute	balAddCos	Our Approach
last	include	**be**
be	assume	**last**
-	happen	get
truth can be reached *only* through the comprehension of opposites.		
Valid Substitute	balAddCos	Our Approach
solely	just	**solely**
exclusively	merely	merely
purely	**solely**	**purely**
uniquely	**barely**	**exclusively**
what do you see yourself doing five or ten years from *now*?		
Valid Substitute	balAddCos	Our Approach
today	lately	**today**
-	recently	nowadays
-	currently	currently
morgan told them to *find* an object that he wants.		
Valid Substitute	balAddCos	Our Approach
discover	obtain	**discover**
locate	**locate**	get
seek	get	**locate**

Table 3: Instances of the lexical substitution task, the target word is in *itshape* and valid substitutes are bolded.

various approaches. Our 50-best predictions include at least one gold substitute candidate for about 95% instances, which significantly outperforms lexical resources such as WordNet, and is comparable with the carefully built PPDB paraphrase database.

We use LS07 trial set to decide the hyperparameters including embedding dropout rate and weight α of proposal score $score_p$. Figure 1 and Table 1 shows the influence of these hyperparameters on the performance of our approach. From Figure 1 we could see that the proposed embedding-level dropout method benefits substitute proposal with dropout rate range from 0.1 to 0.4. Specifically, dropout rate of 0.3 seems to work the best. For weight α, we find that $\alpha = 0.01$ is able to balance well the proposal score $score_p$ and validation score $score_v$.

In addition, to illustrate the performance of our approach, we provide here 6 instances of the lexical substitution task and compare the widely used balAddCos model with our BERT based lexical substitution approach.