Contextualized *Context2vec*

Kazuki Ashihara†, Tomoyuki Kajiwara†, Yuki Arase†, Satoru Uchida‡

† Osaka University, ‡ Kyushu University

{ashihara.kazuki, arase}@ist.osaka-u.ac.jp, kajiwara@ids.osaka-u.ac.jp, uchida@flc.kyushu-u.ac.jp

Lexical Substitution Task

Lexical Substitution Task [1, 2]

... and you are required to listen **hard**.

One event in particular hits the platoon **hard** ...

carefully

badly

- The same word might have different meanings.
- Requires considering the meaning of the word in the context.

approach1 : *context2vec* [3]

- ✓ Generates context embeddings using the whole sentence.
- ✗ Uses simple word embeddings.

approach2 : *DMSE* [4]

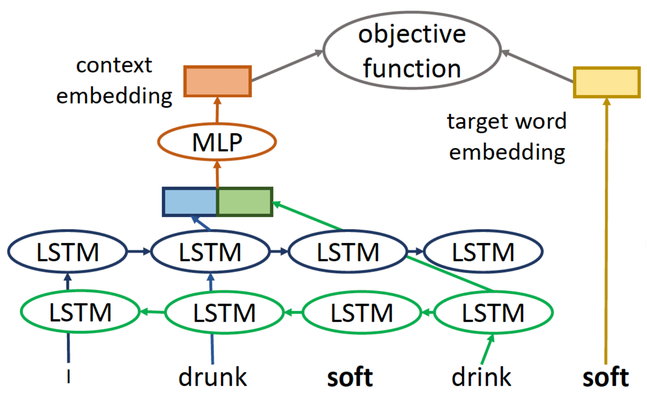
- ✓ Generates contextualized word embeddings by assigning multiple embeddings to one word.
- ✗ Considers only a single word as a context.

Proposed Method

Fusion of *DMSE* and *context2vec*

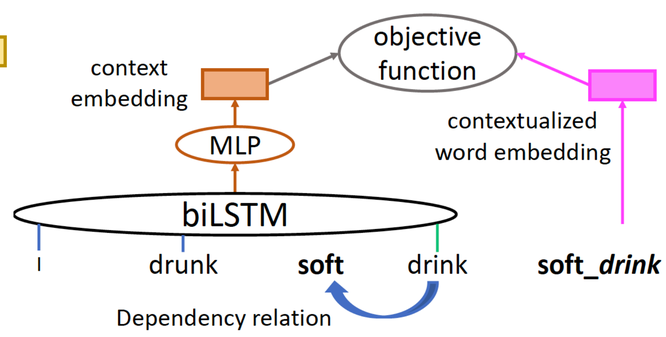
- ✓ Context embedding
- ✓ Contextualized word embeddings

Pre-training



- *context2vec* [3]
- Pre-training of target word embedding.
- Pre-training of parameter of LSTM.

Post-training

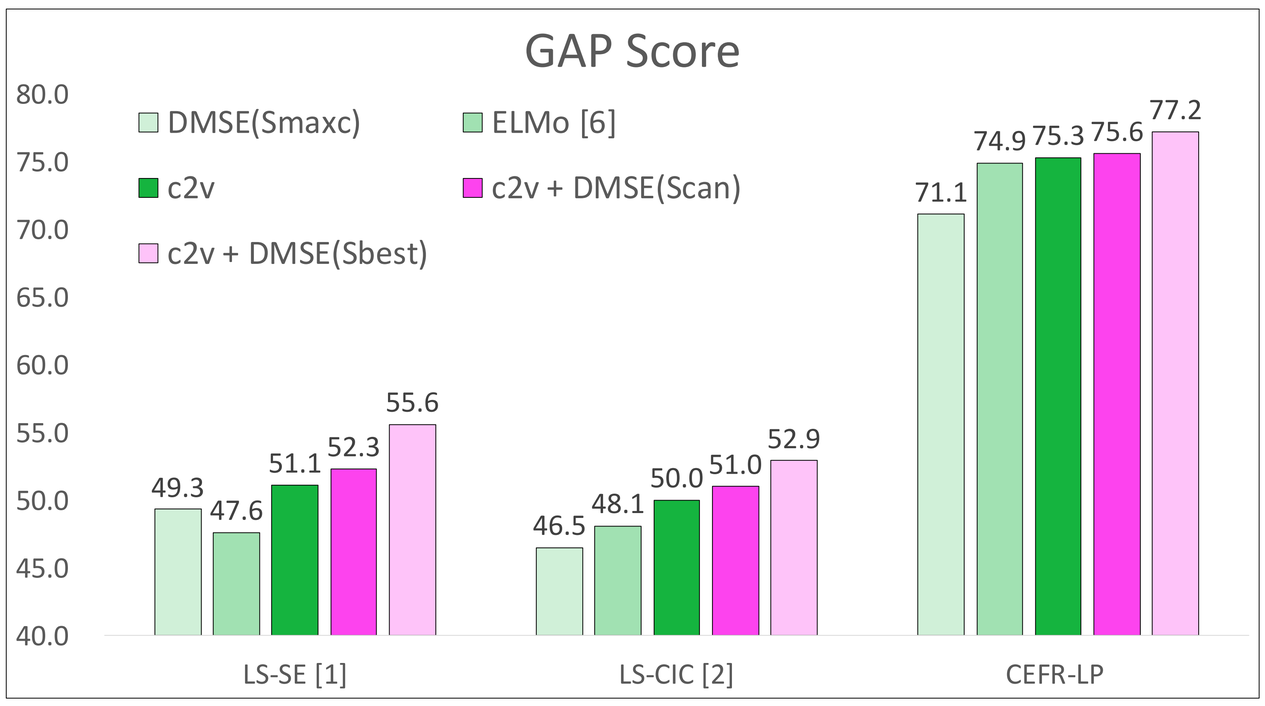


- *DMSE* [4]
- Contextualizes words using dependent words. (*dependency-word*)
- LSTM parameters and pre-trained word embeddings are fixed.

Experiments

Rank in order of cosine similarity of target and candidates.

Results



New Dataset: CEFR-LP

New dataset for Lexical Substitution.
(http://www-bigdata.ist.osaka-u.ac.jp/arase/pj/CEFR-LP.zip)

- Expanded coverage of substitution candidates.
  Extended based on CEFR-LS [5].
- English proficiency levels (CEFR levels).
  A1 (lowest), A2, B1, B2, C1, C2 (highest)

	CEFR-LP	LS-SE	LS-CIC
target word	863	2,010	15,344
candidates	14,259	34,600	601,257
paraphrasable candidates per target	10.0	3.48	6.65

Basic statistics in CEFR-LP.

context	... From alchemy came the historical **progressions** that led to modern chemistry : the isolation of drugs from natural sources , metallurgy , and the dye industry . ...
target	progression [C1]
candidate	block [B1] (0), advancement [B2] (8), break [A2] (1), ...

Example of CEFR-LP.

Examples of Output

Target	**go**
context	... , explain the basic concept and purpose and get it **going** with minimal briefing .
DMSE (S_{maxc})	try (0), move (1), proceed (1), leave (0), be (0), ...
c2v	proceed (1), run (0), start (4), move (1), take (0), ...
c2v + DMSE (S_{can})	start (4), proceed (1), move (1), run (0), take (0), ...

Success example 1: *Dependency-word* is underlined. The numbers in parentheses show candidates' weights.

Target	**tender**
context	Rabbits often feed on young , **tender** perennial growth as it emerges in spring , or on young transplants .
DMSE (S_{maxc})	immature (0), young (0), great (1), soft (4), ...
c2v	delicate (1), immature (0), soft (4), painful (0), ...
c2v + DMSE (S_{can})	soft (4), delicate (1), immature (0), young (0), ...

Success example 2

Target	**hold**
context	A doctor sat in front of me and **held** my hands .
DMSE (S_{maxc})	put (0), lift (1), grasp (3), carry (0), ...
c2v	grasp (3), carry (0), take (1), keep (0), ...
c2v + DMSE (S_{can})	take (1), carry (0), keep (0), lift (1), ...

A Failed example caused by incorrect *dependency-word* (sat) selection.

[References]
[1] McCarthy et al., 2007, "SemEval-2007 Task 10: English Lexical Substitution Task," In Proc. of SemEval, pp. 48-53.
[2] Kremer et al., 2014, "What Substitutes Tell Us - Analysis of an "All-Words" Lexical Substitution Corpus," In Proc. of EACL, pp. 540-549.
[3] Ashihara et al., 2018, "Contextualized Word Representations for Multi-Sense Embedding," In Proc. of PACLIC, pp. 28-36.
[4] Melamud et al., 2016, "context2vec: Learning Generic Context Embedding with Bidirectional LSTM," In Proc. of CoNLL, pp. 51-61.
[5] Uchida et. al., 2018, "CEFR-based Lexical Simplification Dataset," In Proc. of LREC, pp. 3254-3258.
[6] Peters et al., 2018, "Deep Contextualized Word Representations," In Proc. of NAACL, pp. 2227-2237.