Building Static Embeddings from Contextual Ones: Is It Useful for Building Distributional Thesauri?

Olivier Ferret
CEA List, LASTI Laboratory

Contact: olivier.ferret@cea.fr

Context

- Extraction of lexical semantic relations in English
  - Paradigmatic relations and more particularly synonyms
- General method: distributional thesauri
  - For each target word: retrieval of its first neighbors according to the distributional representation of words
  - According to the principles of distributional analysis (Harris, 1954)
    - first neighbors = synonyms of the target words

Objectives

- Adaptation to contextual embeddings (ELMo, BERT…)
  - Static embeddings = word-level representations
  - Contextual embeddings = token-level representations

→ Building of word representations from contextual language models

Existing methods

Averaging of token embeddings (Bommasani et al., 2020)
- Encoding of sentences containing the target word with a contextual model
- Word representation = average of embeddings extracted for the target word in all sentences

Principal Component Analysis (Ethayarajh, 2019)
- Word representation = first principal component of token representations

Token without context (Vulić et al., 2020)
- Encoding a pseudo-sentence made of only one word

Proposed method

Starting point
- For each target word: set of token representations = in-context word representations

General idea
- Exploiting the diversity of token representations for building a more general representation at the word level

3 main steps
- Selection of a focused set of token representations for each target word
  - for limiting the heterogeneity of representations
- Generalization of selected token representations
- Aggregation of generalized token representations

Representation selection

Starting point
- Random selection of *N* sentences from the AQUAINT corpus (news articles) for each target word
- Target words = nouns
- *N* between 10 and 250 according to the frequency of the target nouns

Hypothesis
- (McCarthy et al., 2004): dominant sense for each word in a corpus
- Average token representation = representation of the dominant sense of the associated word

4 selection strategies
- *Random*
- *Closest* to the dominant sense
- *Farthest* to the dominant sense
- *Uniform* distribution in terms of proximity with the dominant sense

→ 10 token representations / word

Representation generalization and aggregation

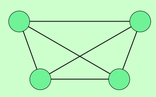
Generalization principles
- Adaptation of the token representations of a word to bring them closer together
- Similar to knowledge injection methods such as *retrofitting* (Faruqui et al., 2015)
  - as if the token representations of a word were linked by similarity relations

2 main variants
- **gen+agg**: generalization, then aggregation
  - generalization
    - pseudo-similarity relation between each pair of token representations
    - application of a retrofitting-like algorithm: PARAGRAM (Wieting et al., 2015)
  - aggregation: averaging of generalized token representations
- **agg+gen**: aggregation, then generalization
  - aggregation: averaging of token representations
  - generation of pseudo-similarity relations (generalization)
    - **agg+gen$_{all}$**: between each pair of token representations, including the aggregate
    - **agg+gen$_{agg}$**: only between the aggregate and the initial token representations
  - application of the retrofitting-like algorithm to all representations

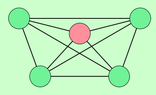
➢ final word representation = aggregate

pseudo-similarity relation
token representation
aggregate

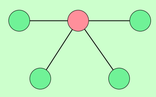
gen+agg



agg+gen$_{all}$



agg+gen$_{agg}$



Experiments and evaluation

Large-scale intrinsic evaluation
- Gold Standard: WordNet's synonyms
- Evaluated target words = 10,305 nouns
  - frequency: large spectrum
- Information Retrieval (IR) paradigm
  - target word ≡ query; neighbors ≡ docs
  - IR measures: MAP, R-precision, precision@{1,2,5}
- Evaluated models
  - BERT uncased: token representation = average representation of its WordPieces
  - CharacterBERT (CBERT): influence of WordPieces
  - fastText: static embeddings (Skip-gram model) from Wikipedia

Evaluation of the proposed method
- Reference methods: *-avg,*-pca, *-iso
  - *-iso < *-avg = *-pca ~ fastText
- Proposed method > reference methods and static embeddings
  - for both BERT and CharacterBERT
- aggregation+generalization > generalization+aggregation
- agg+gen$_{agg}$ ~ agg+gen$_{all}$
  - preference for agg+gen$_{agg}$: requires fewer computations
- Similar improvement for CharacterBERT and BERT

	R_{prec}	MAP	P@1	P@2	P@5
CBERT-avg	15.6	18.0	22.0	15.9	9.2
CBERT-pca	15.6	17.9	22.0	15.9	9.2
CBERT-gen+agg	16.1	18.6	22.6	16.3	9.5
CBERT-agg+gen$_{all}$	**16.3**	**18.9**	**22.8**	**16.5**	9.7
CBERT-agg+gen$_{agg}$	**16.3**	18.8	**22.8**	16.4	9.6
BERT-avg	15.6	17.9	21.8	16.0	9.5
BERT-iso (L0)	14.0	15.8	19.2	14.6	8.7
BERT-agg+gen$_{agg}$	16.2	18.8	22.5	16.1	**10.1**
fastText	15.5	18.4	21.9	15.7	9.2

Table 1

Best layer for CharacterBERT = L4
Best layer for BERT = L5

Random selection for token representations

Evaluation of selection methods for token representations
- Results for agg+gen$_{agg}$
- Equivalence of all strategies
- Limited effect of the contextualization of embeddings

➢ uniform: « logical » choice as a deterministic strategy

	R_{prec}	MAP	P@1	P@2	P@5
random	16.3	18.8	22.8	16.4	9.6
uniform	16.4	**18.9**	**22.9**	**16.7**	9.7
farthest	**16.5**	**18.9**	**22.9**	16.6	**9.8**
closest	16.3	18.8	**22.9**	16.4	9.6

Table 2

Word frequency analysis
- Split of results according to the median frequency of target words
- High-frequency words
  - contextual > static
  - impact of contextualization on polysemy
- Low-frequency words: opposite trend
- Fusion of thesauri (CombSum strategy): significant improvement / static and contextual

	R_{prec}	MAP	P@1	P@2	P@5
uniform$_{high}$	18.0	20.5	26.6	19.8	12.0
uniform$_{low}$	14.8	17.3	19.3	13.4	7.3
fastText$_{high}$	14.5	16.8	22.0	15.9	9.7
fastText$_{low}$	16.4	20.0	21.8	15.4	8.6
CombSum	18.6	21.5	25.9	19.0	11.1

Table 3