Language Models and Semantic Relations: a Dual Relationship

LREC-COLING 2024

Olivier Ferret
CEA-List, LASTI Laboratory

Contact: olivier.ferret@cea.fr

Context

Global perspective
- Injection of lexical semantic relations into BERT-like Pretrained Language Models (PLMs)

Lexical semantic relations
- Present in manually-built resources such as dictionaries (ex.: WordNet)
  - Incompleteness + limited availability for specialised domains and most languages
- Lexical relations: can be extracted from static and contextual PLMs

Objectives

Research question
Is it possible to
- enrich semantically BERT-like PLMs
- by only relying on lexical semantic relations automatically extracted from static and contextual PLMs?

PLMs ? **Semantic lexical relations**

Relation extraction: principles

Focus
- Paradigmatic relations: synonymy, hypernymy…
  - semantic similarity vs. semantic relatedness (Budanitsky & Hirst, 2006)

k-nearest neighbor (*k*-NN) word similarity graph
- For each word w_i → k closest semantic neighbors
- Set $\{w_i\}$ → neighborhood graph (directed graph)
- Hypothesis (Ferret, 2012; Claveau et al., 2014)
  - symmetry of semantic similarity relations → if w_1 and w_2 are similar, $w_1 \in$ {closest neighbors of w_2} and vice versa
  - ➢ selection of relations to extract based on reciprocity relations in the neighborhood graph

Definition of semantic neighbors
- Words similar to the target word w_i according to a language model

Static language models (Skip-gram)
- Semantic neighbors of w_i = k words w_j most similar to w_i according to the Cosine measure between the representations of w_i and w_j

Contextual language models (BERT-like)
- Semantic neighbors of w_i = k first substitutes of w_i provided by a BERT-like language model in masked language modeling task

Relation extraction with BERT-like models

Masked language modeling task

$$\underbrace{\text{LREC is a conference about}}_{\text{prompt}}\ \underbrace{\text{[MASK].}}_{\text{target}} \Rightarrow$$

technology	0.021
science	0.019
biology	0.017
computers	0.016
…	

Using nominal compounds as prompts for relation extraction
- Semantic similarity: distributionally linked to short contexts → compounds
- Extraction of nominal compounds from Wikipedia with the method from (Mikolov et al., 2013) → 3 main structures
  - TERM = ADJ NOUN | NOUN NOUN | NOUN PREP NOUN

Prompt structure
- Basic element: TERM_MSK → all variants of TERM in which NOUN (head or modifier) is masked
- (Qiang et al., 2020): better results by conditioning the sequence with the word to predict by the full sequence
- ➢ General form of prompts

TERM . [SEP] TERM_MSK

TERM = ADJ NOUN → ADJ NOUN . [SEP] ADJ [MASK]
TERM = civil defence → civil defence . [SEP] civil [MASK]

Prompt variants
- Variants of the conditioning part (before [SEP]): w/ or w/o determinant, full sentence or compound only
- Variants of the conditioned part (after [SEP]): idem conditioning part + w/ or w/o TERM

P0 *TERM* . [SEP] *TERM_MSK* .
P1 this is a/an *TERM* . [SEP] this is a/an *TERM_MSK* .
P2 *TERM* . [SEP] this is a/an *TERM_MSK* .
P3 a/an *TERM* . [SEP] a/an *TERM_MSK* .
P4 *TERM* . [SEP] a/an *TERM_MSK* is a kind of TERM .
P5 *TERM* . [SEP] a/an *TERM_MSK* is a/an *TERM* .
P6 *TERM* . [SEP] a/an *TERM* is a/an *TERM_MSK* .
P7 *TERM* . [SEP] a/an *TERM_MSK* and a/an *TERM* .
P8 *TERM* . [SEP] a/an *TERM_MSK* or a/an *TERM* .

Result aggregation
- Target word: can be present in several compounds → several lists of semantic neighbors
- List fusion: CombSum method + zero-one normalization of scores
  - CombSum: sum of normalized scores

Experiments and evaluation

Relation injection
- **LexFit** (Vulić et al., 2021)
  - transposition of SentenceBERT (Reimers et al., 2019) at the level of words
  - input: pairs of similar words (w_i, w_j)
  - each word encoded as a sentence

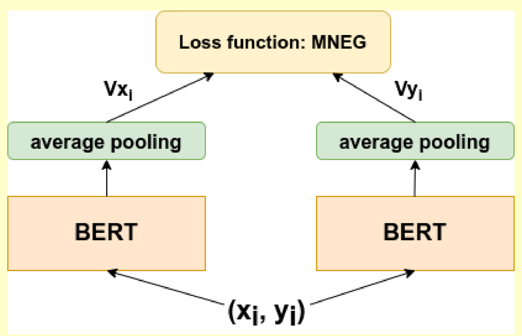

Figure 1

- Siamese network architecture + contrastive loss function
  - MNEG (Multiple Negatives Ranking) loss: inside a batch of (w_i, w_j)
    - maximize the similarity of similar pairs (w_i, w_j)
    - minimize the similarity of dissimilar pairs (w_i, w_k)
      - w_k: a word from another similar pair in the batch

Evaluation of extracted relations
- Gold Standard: WordNet's paradigmatic relations
  - synonyms, hypernyms, hyponyms, cohyponyms
- 2 language models: BERT and CharacterBERT (CBERT)

Prompt	P0	P1	P2	P3	P4	P5	P6	P7	P8
Accuracy	**32.0**	30.3	24.9	31.1	30.9	31.7	26.5	31.7	31.0

Table 2

- Prompt variants
  - best results for P0, the simplest prompt
  - only P2 and P6 having lowest results

	Accuracy	# relations
BERT	13.9	17,007
CBERT	22.9	17,023
CBERT – heads	24.2	**13,465**
CBERT – modifiers	16.0	7,792
CBERT – ADJ NOUN	**26.2**	10,511

Table 1

- Globally, with P0 prompt
  - CBERT > BERT and head > modifier
  - ADJ NOUN: best trade-off between accuracy and # extracted relations

- **Extracted** relations
  - accuracy : static ~ contextual but > x 2 relations for static
- **Selected** relations
  - static > contextual for both accuracy & number
  - fusion: small ↘ for accuracy but many more relations

Relations	Model	Accuracy	# relations
Extracted	Static	30.0	35,246
	Contextual	30.6	15,473
Selected by reciprocity	Static	**44.1**	11,298
	Contextual	42.6	8,558
	Fusion	41.1	**18,430**

Table 3

Evaluation of relation injection
- Evaluation task: semantic similarity of static word embeddings built from a BERT model
  - static embedding: encoding of one occurrence of a word without context → embedding from best layer; average of token embeddings
- Information Retrieval (IR) paradigm
  - target word ≡ query; neighbors ≡ docs
  - IR measures: MAP, R-precision, precision@{1,5,10}
- Gold Standard: idem relation extraction
- Evaluated target words = 10,305 nouns
  - frequency: large spectrum

- **BERT-iso**: base for injection
  - word in isolation
  - < word in context (**BERT-ctxt**)
- **BERT-refsyn**: upper reference ~ LexFit
  - injection of 1 million manual relations (WordNet, Roget…)
- Very significant improvement brought by the injection of extracted relations (/ BERT-iso)
  - slightly higher for contextual / static
  - fusion: best option, especially for P@1
  - fusion < BERT-refsyn but with far fewer relations (55 times), and automatically extracted

	MAP		P@1		P@5		P@10	
fastText-wiki	6.0		36.5		21.3		15.9	
BERT-ctxt	5.7		36.5		22.4		17.0	
BERT-iso	4.4		30.9		19.6		14.6	
BERT-refsyn	12.3		55.7		37.4		29.0	
relations: static	7.4	+3.0	42.2	+11.3	26.1	+6.5	19.7	+5.1
relations: contextual	7.6	+3.2	42.8	+11.9	26.9	+7.3	20.3	+5.7
relations: fusion	**7.8**	+3.4	**44.2**	+13.3	**27.0**	+7.4	**20.4**	+5.8

Table 4