# USING PSEUDO-SENSES FOR IMPROVING THE EXTRACTION OF SYNONYMS FROM WORD EMBEDDINGS

Olivier Ferret

# CONTEXT AND OBJECTIVES

- **Context**
  - semantic specialization of word embeddings
  - most approaches following Retrofitting [Faruqui et al., 2015]
    - a priori set of lexical semantic relations
    - bring word vectors closer if they are part of similarity relations (synonymy, lexical association ...)
    - move them away from each other if they are part of dissimilarity relations (antonymy …)

- **Objectives of Pseudofit**
  - improving word embeddings for semantic similarity without a priori lexical relations

# PRINCIPLES: GENERAL PERSPECTIVE

- **Theoritical hypothesis**
  - homogeneous corpus C
  - equal split of C in 2 parts: C1 and C2
  - distributional representation of a word w from a corpus C = $distrep_{C}(w)$ = set of contexts
  - $distrep_{C1}(w) = distrep_{C2}(w)$

- **In practice**
  - $distrep_{C1}(w) \neq distrep_{C2}(w)$

- **Hypothesis**
  - differences between $distrep_{C1}(w)$ and $distrep_{C2}(w)$ are contingent
  - bringing $distrep_{C1}(w)$ and $distrep_{C2}(w)$ closer → more general (and better) distributional representation of w

# PRINCIPLES: IMPLEMENTATION

- **Distributional representations**
  - dense representations: Skip-Gram [Mikolov et al., 2013]

- **Notion of pseudo-sense**
  - 2 sub-corpora → 2 representation spaces
    - require projection in a shared space → source of disturbances
  - instead, 1 corpus but 2 pseudo-senses for each word
  - pseudo-sense
    - arbitrarily split the occurrences of a word into two or more subsets

- **Overall process**
  - generation of distributional contexts for pseudo-senses
  - turning pseudo-sense contexts into dense representations
  - convergence of pseudo-word representations → more general word representation

# REPRESENTATIONS OF PSEUDO-WORDS

- **Generation of contexts**
  - 2 successive occurrences of a word → 2 different pseudo-senses
  - 3 representations / word
    - 2 pseudo-senses + word itself → for each occurrence, generation of contexts for the current pseudo-sense + word
    - « frequency trick »: adding the representation of the word → avoiding the impact of having half the occurrences for each pseudo-sense

A $policeman_1$ was arrested by another $policeman_2$.

| TARGET | CONTEXT | TARGET | CONTEXT | TARGET | CONTEXT |
|---|---|---|---|---|---|
| policeman | a | $policeman_1$ | a | $policeman_2$ | another |
| policeman | be | $policeman_1$ | be | $policeman_2$ | by |
| policeman | arrest (x2) | $policeman_1$ | arrest | $policeman_2$ | arrest |
| policeman | by (x2) | $policeman_1$ | by | | |
| policeman | another | | | | |

- **Building of dense representations**
  - `word2vecf` [Levy & Goldberg, 2014]

# CONVERGENCE OF PSEUDO-WORD REPRESENTATIONS

- **Principles**
  - 3 representations / word w: v (word); $v_1$, $v_2$ (pseudo-senses)
  - v, $v_1$ and $v_2$: supposed to be semantically equivalent

  ➔ 3 similarity relations: (v, $v_1$), (v, $v_2$) and ($v_1$, $v_2$)

  - application of a semantic specialization method for word embeddings to v, $v_1$ and $v_2$ with the similarity relations between them
  - final representation for w: v after its « specialization »

- **Implementation**
  - specialization method: PARAGRAM [Wieting et al., 2015]
    - comparable to Retrofitting but includes an automatically generated repelling component
      - for each target word to specialize, selection of a repelling word, either randomly or according to their dissimilarity

# INTRINSIC EVALUATION

- **Experimental setup**
  - 1 billion lemmatized words randomly selected from the Annotated English Gigaword corpus [Napoles et al., 2012] at the level of sentences
  - word embeddings built with the best parameters from [Baroni et al., 2014]
  - focus on nouns

- **Word similarity evaluation**
  - Spearman's rank correlation between human judgments and similarity between vectors for 3 representative datasets of word pairs

| | SimLex-999 | MEN | Mturk 771 |
|---|---|---|---|
| INITIAL | 49.5 | 78.3 | 65.6 |
| Pseudofit | 51.2 | 79.9 | 68.0 |
| Retrofitting | 49.6 | 77.4 | 65.0 |
| Counter-fitting | 49.5 | 77.2 | 64.9 |

$\times$ 100

# SYNONYM EXTRACTION

- **Evaluation framework**
  - Gold Standard: WordNet's synonyms
    - 2.9 / word
  - evaluated words = 11,481 nouns
    - frequency > 20
  - for each evaluated noun, retrieval of its 100 nearest neighbors
    - neighbors ranked from most similar (Cosine) to less similar
  - Information Retrieval (IR) paradigm
    - evaluated word ≡ query; neighbors ≡ docs
    - IR measures: MAP, R-precision, precision@{1,2,5}

| | R-prec. | MAP | P@1 | P@2 | P@5 |
|---|---|---|---|---|---|
| INITIAL | 13.0 | 15.2 | 18.3 | 13.1 | 7.7 |
| Pseudofit | +2.5 | +3.3 | +3.0 | +2.5 | +1.8 |

$\times$ 100

# SENTENCE SIMILARITY

- **Evaluation task**
  - Semantic Textual Similarity: STS Benchmark dataset [Cer et al., 2017]
  - Pearson rank correlation between human judgments and similarity between sentences for a set of reference sentence pairs
- **Computation of sentence similarity**
  - strong baseline approach based on word embeddings
  - sentence representation: elementwise addition of the embeddings of the plain words of the sentence
    - use of Pseudofit$_{\text{[max,fus-max-pooling]}}$ embeddings, defined for nouns, verbs and adjectives
  - sentence similarity: *Cosine* between sentence representations

| | **$\rho \times 100$** |
|---|---|
| INITIAL | 63.2 |
| Pseudofit$_{\text{[max,fus-max-pooling]}}$ | 66.0 |
| Best baseline (Cer et al., 2017) | 56.5 |

# CONCLUSIONS AND PERSPECTIVES

- **To sum up**
  - Pseudofit: method for improving word embeddings towards semantic similarity without external semantic relations
  - method based on the convergence of several representations built from the same corpus → more general representation
  - successful intrinsic and extrinsic evaluations for word similarity, synonym extraction and sentence similarity

- **Research directions**
  - transposition of Pseudofit with several corpora → link with researches about meta-embeddings and ensembles of word embeddings

Commissariat à l’énergie atomique et aux énergies alternatives
Institut List | CEA SACLAY NANO-INNOV | BAT. 861 – PC142
91191 Gif-sur-Yvette Cedex - FRANCE
www-list.cea.fr

Établissement public à caractère industriel et commercial | RCS Paris B 775 685 019