Context-Dependent Sense Embedding

Lin Qiu[1], Kewei Tu[2], Yong Yu[1]

[1] Shanghai Jiao Tong University

[2] ShanghaiTech University

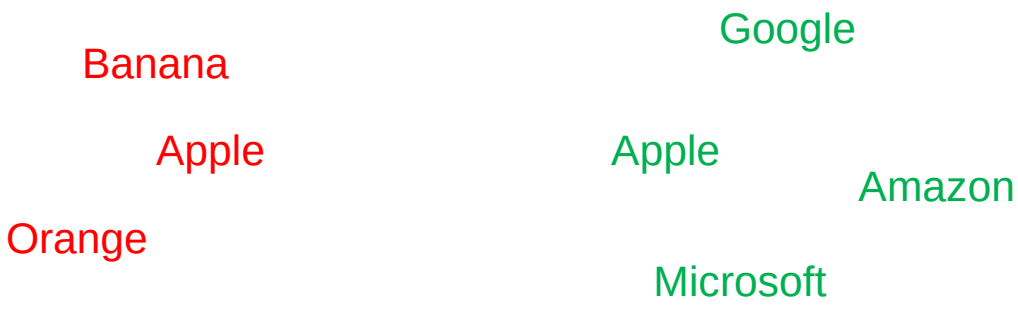
上海交通大學 SHANGHAI JIAO TONG UNIVERSITY

Content

- Motivation
- Related Work
- Our Model
- Results & Discussion
- Summary

Word Embedding

- Represent a word with a continuous vector
- Useful in many NLP tasks
  - Overcome the curse of dimensionality
  - Improve the performance of other models
- Problems
  - A word may have multiple senses (meanings)
  - Each sense should have a different embedding



上海交通大學 SHANGHAI JIAO TONG UNIVERSITY

Sense Embedding

- Represent each sense of a word with a different vector
- Challenges
  - Determine the sense of each word: Word Sense Disambiguation (WSD)
  - Train sense embedding for each sense of a word
  - How to organize these two processes in one model?

上海交通大學 SHANGHAI JIAO TONG UNIVERSITY

Content

- Motivation
- Related Work
  - Heuristic methods
  - Probabilistic methods
- Our Model
- Results & Discussion

Related Work - Heuristic methods

- Do WSD by clustering the context words
- Utilizing external knowledge from dictionary like WordNet
  - Number of word senses
  - Gloss of word senses
- Train sense embedding directly with word embedding models

- Drawbacks
  - Heuristic in nature
  - Rely on external knowledge

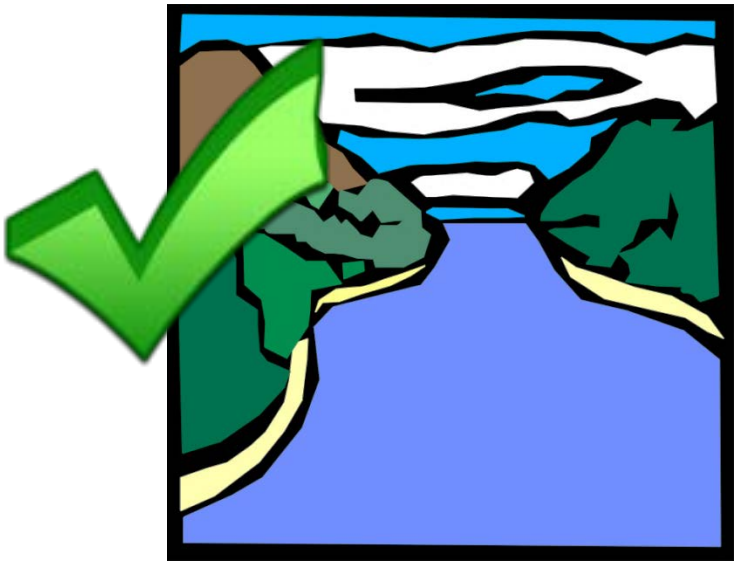
Related Work - Probabilistic methods

- Build a probabilistic model for sense embedding
- Learning with EM algorithm
  - E step: determine word senses
  - M step: train embedding of inferred word senses with word embedding models
- Drawbacks
  - They determine word senses dependent on the word embeddings of its context words
  - They do not model *dependency between sense choices*

Related Work - Probabilistic methods

- Why do we model dependency between sense choices?

He cashed a check at the bank



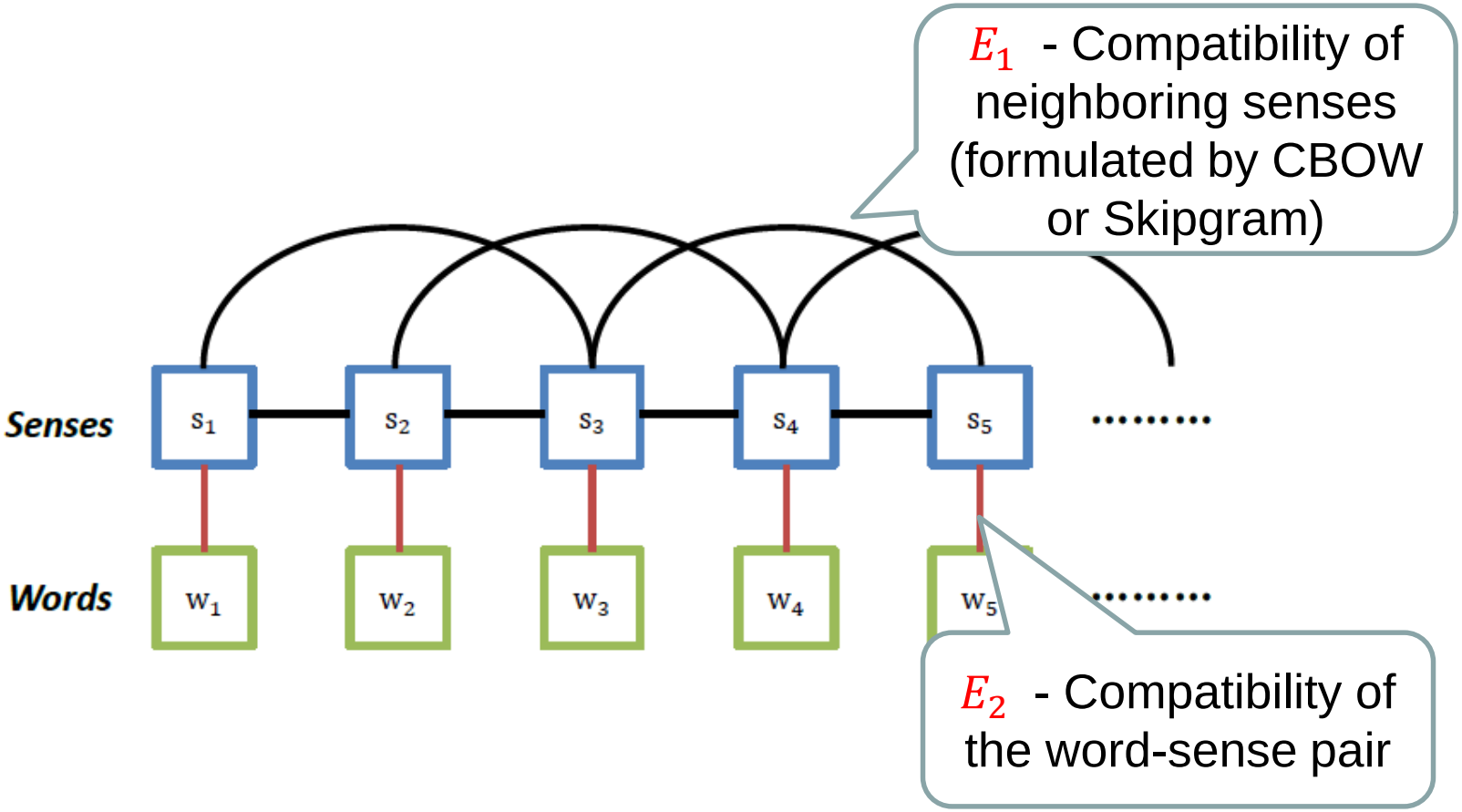
上海交通大學 SHANGHAI JIAO TONG UNIVERSITY

Content

- Motivation
- Related Work
- Our Model
- Results & Discussion
- Summary

Our Model

- Word2vec + high-order hidden Markov model



上海交通大學 SHANGHAI JIAO TONG UNIVERSITY

Our Model

- Energy function
  - $E(\boldsymbol{w}, \boldsymbol{s}) = \sum_i [E_1(s_{i-k}, \dots, s_{i+k}) + E_2(w_i, s_i)]$
  - E_1 can be formulated with CBOW or Skip-gram
    - For CBOW:

$$E_1(s_{i-k}, \dots, s_{i+k}) = -\sigma(\sum_{i-k \le j \le i+k, j \ne i} V^T(s_j) V'(s_i))$$

  - $E_2(w_i, s_i) = \begin{cases} 0 & s_i \in S(w_i) \\ +\infty & otherwise \end{cases}$

- No problematic word embeddings

上海交通大學 SHANGHAI JIAO TONG UNIVERSITY

Inference

- Find the best sense sequence corresponding to the word sequence
- Formulation: $\boldsymbol{s}^* = \arg min_{\boldsymbol{s}} E(\boldsymbol{w}, \boldsymbol{s})$

- Solution: dynamic programming
  - Similar to the Viterbi Algorithm of Hidden Markov Model

Learning

- Mini Batch Hard-EM
- E step: predict sense sequence by dynamic programming
- M step: learning sense embedding with SGD
  - Using CBOW or Skip-gram with Negative sampling
  - Negative samples
    - Ambiguous words: other senses which is not selected
    - Unambiguous words: the same with Word2vec
- Mini Batch: sentence level
  - avoid trapping into local optima
  - faster convergence

Learning

- Determine the number of senses
  - Using the sense number of a sense inventory from SemEval 2007 Task 07
    - The only external knowledge we utilize in learning
  - Automatically learn the number of senses (future work)
    - Dirichlet process
    - Split-Merge Algorithm

上海交通大學 SHANGHAI JIAO TONG UNIVERSITY

Content

- Motivation
- Related Work
- Our Model
- Results & Discussion
- Summary

上海交通大學 SHANGHAI JIAO TONG UNIVERSITY

Experiment setup

- Training Corpus: Wikipedia corpus
  - 3.6 M articles
  - 1.3 B tokens
- Hyper parameters
  - Window size = 5
  - Vector size = 300
  - AdaGrad with initial learning rate = 0.025

Qualitative Evaluation

Word	Nearest Neigbors
bank_1	banking, lender, loan
bank_2	river, canal, basin
bank_3	slope, tilted, slant
apple_1	macintosh, imac, blackberry
apple_2	peach, cherry, pie
date_1	birthdate, birth, day
date_2	appointment, meet, dinner
fox_1	cbs, abc, nbc
fox_2	wolf, deer, rabbit

Table 1: The nearest neighbors of senses of polysemous words

Stanford's Contextual Word Similarities (SCWS) dataset

- Heuristic methods outperform probabilistic methods
  - Possible reason: heuristic methods use more external knowledge
- Our model outperforms other probabilistic models
- Caution
  - The use of word similarity tasks for evaluation of word vectors is not sustainable. [Faruqui et al. (2016)]

Model	Similarity Metrics	$\rho \times 100$
Huang	AvgSim	62.8
Huang	AvgSimC	65.7
Chen	AvgSim	66.2
Chen	AvgSimC	68.9
Neelakantan	AvgSim	67.2
Neelakantan	AvgSimC	69.2
Li		**69.7**
Tian	Model_M	63.6
Tian	Model_W	65.4
Bartunov	AvgSimC	61.2
Ours + CBOW	HardSim	64.3
Ours + CBOW	SoftSim	65.6
Ours + Skip-gram	HardSim	64.9
Ours + Skip-gram	SoftSim	**66.1**

Table 2: Spearman's rank correlation results on the SCWS dataset

Word Sense Induction

- Two fuzzy metrics:
  - Fuzzy B-Cubed (FBC)
  - Fuzzy Normalized Mutual Information (FNMI)

Model	FBC(%)	FNMI(%)	HM
AI-KU	35.1	4.5	8.0
MSSG	45.9	3.7	6.8
ICE-online	48.7	5.5	9.9
ICE-kmeans	51.1	5.9	10.6
Ours + CBOW	53.8	6.3	11.3
Ours + Skip-gram	56.9	6.7	**12.0**

← 13% gain

Table 3: Results of single-sense instances on task 13 of SemEval-2013

上海交通大學 SHANGHAI JIAO TONG UNIVERSITY

Content

- Motivation
- Related Work
- Our Model
- Results & Discussion
- Summary

上海交通大學 SHANGHAI JIAO TONG UNIVERSITY

Summary

- We propose a novel probabilistic model to learn sense embedding
  - Word2vec + HMM
- Advantage
  - We model the dependency between sense choices
  - We avoid using word embeddings of polysemous words
- Inference by dynamic programming
- Learning by mini batch hard EM algorithm
- We achieve competitive results on SCWS dataset and a 13% gain on a WSI task

上海交通大學 SHANGHAI JIAO TONG UNIVERSITY

Future work

- Soft EM
  - E step: assign a weight for each sense choice
  - Train the aggregated embedding
- Incorporating shared senses instead of lexemes
  - Different words can share the same sense, like WordNet
  - The weight of word-sense pair can be a real value
- Learning of sense numbers

上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

Q & A