Document Embeddings with Context Sampling

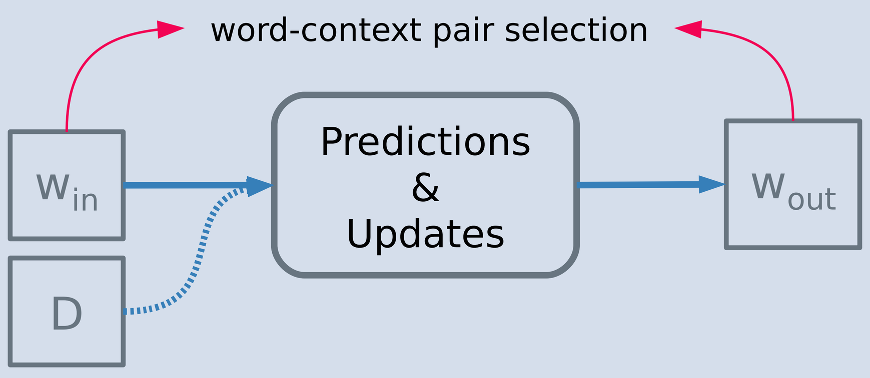
Stefanos Angelidis, Victor Lavrenko, Mirella Lapata
Institute for Language, Cognition and Computation, School of Informatics, University of Edinburgh

THE UNIVERSITY of EDINBURGH informatics

Why Document Embeddings?

- Embedding models are used widely for learning word representations from vast amounts of unlabeled text.
- Represent meaning of longer pieces of text → embedding composition.
- Averaging / Syntax-aided composition:
  - Can work for phrases or short sentences.
  - Severe loss of semantic information as sequence length increases.
- CNNs / Recurrent NNs / Hierarchical NNs:
  - State-of-the-art in many supervised tasks.
  - Computationally demanding (need GPUs).
- Middle ground → Paragraph Vector (Le and Mikolov, 2014).

Paragraph Vector



- Extension of Skip-gram (Mikolov et al., 2013).
- Word and document embeddings are learned jointly w/o supervision.
- Words are paired with their window-based contexts.
- Document embeddings are also used to predict each word they contain.
- **Issue**: Are selected word-context pairs representative of content?

PV

The Ministry of industry is expected to give Elf Aquitaine, the oil group controlled by the French state, total clearance to take over the UK petrol-station.

Window-based context

- Disregards word importance.
- Implicitly forces doc. embeddings towards frequent words.

This work: Context Sampling Framework

- We introduce arbitrary contexts via Context Sampling.
- Different sampling policies will result in different embedding spaces.

DE idf

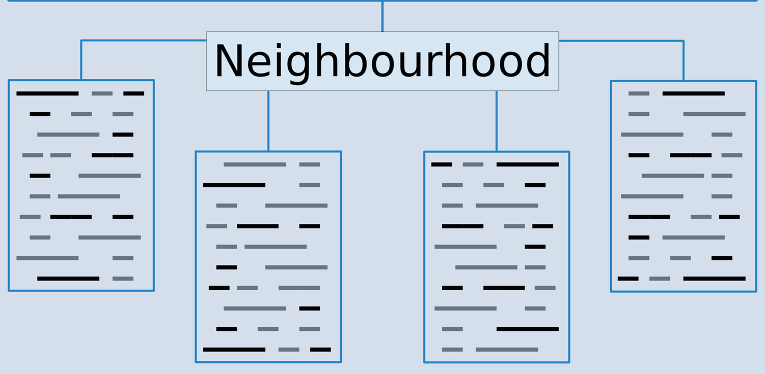
The Ministry of industry is expected to give Elf Aquitaine, the oil group controlled by the French state, total clearance to take over the UK petrol-station.

IDF Sampling

- Context words sampled from document-wide *tf.idf* distribution.
- Doc. embeddings are positioned closer to content-heavy words.

DE nn

The Ministry of industry is expected to give Elf Aquitaine, the oil group controlled by the French state, total clearance to take over the UK petrol-station.

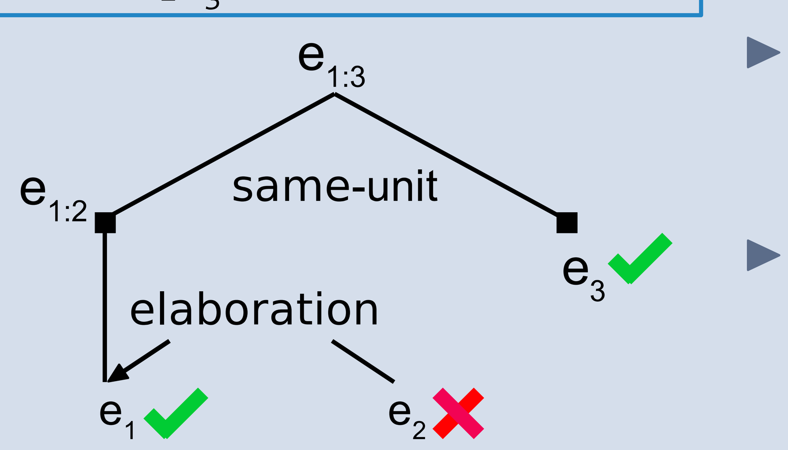


Neighborhood Sampling

- Incorporates clustering hypothesis to context selection.
- Context words sampled from fixed-size neighborhood of similar documents.
- Words that do not appear in current document may be used as well.

DE disc

[The Ministry of industry is expected to give Elf Aquitaine,]e_1 [the oil group controlled by the French state,]e_2 [total clearance to take over the UK petrol-station.]e_3



Discourse-based Sampling

- Attempts to *inject* discourse-level linguistic information.
- Not all parts of a document are equally important.
- We parse documents using an RST-style discourse parser (Feng and Hirst, 2012).
- Potentially insignificant *elementary discourse unit* (EDU) types are filtered-out before context selection.

Evaluation

- Compare the quality of document embeddings learned using PV's window-based contexts and our context sampling policies (DE).
- We chose 2 document-centric tasks:
  - Ad-Hoc Search
  - Document Classification

Results: Ad-hoc Search

- Used 2 established TREC collection for Information Retrieval.
- Learned query and document embeddings (no supervision).
- Documents ranked by their cosine distance to a query in embedding space.

Methods	ROBUST MAP	ROBUST %-change vs. PV	ROBUST %-change vs. DE_{idf}	AP88-89 MAP	AP88-89 %-change vs. PV	AP88-89 %-change vs. DE_{idf}
PV	0.1179	—	—	0.0938	—	—
DE_{idf}	0.1328	12.6*	—	0.1154	23.0*	—
$\mathrm{DE}_{q.nn}$	0.1693	43.6*	27.5*	0.1442	53.7*	24.9*
$\mathrm{DE}_{f.nn}$	**0.1823**	54.6*	37.3*	**0.1631**	73.8*	41.3*

* indicates significant improvement based on a two-tailed t-test with $p < 0.01$.

- Can provide complementary signal to term-based IR methods.

Results: Document Classification

- Sentiment Analysis (IMDB) & Topic Classification (RCV1).
- Learned document embeddings (no supervision).
- Trained logit classifier using document embeddings as features.

Classification Performance

Methods	Accuracy (%) IMDB	Accuracy (%) RCV1
N-gram	86.52	85.12
RNN-LM	86.61	85.08
PV	88.93	86.95
DE_{idf}	89.29^a	87.71^a
DE_{nn}	$\mathbf{89.34}^a$	87.75^a
$\mathrm{DE}_{disc.pv}$	88.37	87.05^a
$\mathrm{DE}_{disc.idf}$	88.82	87.97^{abc}
$\mathrm{DE}_{disc.nn}$	88.87	$\mathbf{88.01}^{abc}$

Markers *a*, *b* and *c* denote significant improvements over PV, DE_{idf} and DE_{nn} resp. (one-tail t-test with $p < 0.01$).

- IDF and Neighborhood sampling outperform PV on both datasets.
  - Great performance – complexity trade-off.
- Discourse-based sampling helps slightly on RCV1, not on IMDB.
  - Parsing quality.
  - Idiosyncrasies of sentiment analysis.
  - Better EDU filtering strategies?

Qualitative Evaluation

	PV	DE_{idf}	$\mathrm{DE}_{disc.idf}$
health	*care*	*medical*	*medical*
	medical	*physician*	*hospital*
	education	*hospital*	*nhs*
	benefits	*therapy*	*nursing*
politics	*political*	*political*	*political*
	politicians	*party*	*party*
	candidature	*polls*	*election*
	dirty	*election*	*leader*
weather	*cooler*	*temperatures*	*temperatures*
	warm	*meteorologist*	*dry*
	dry	*warm*	*meteorologist*
	warmer	*rain*	*precipitation*

Based on cosine similarity

- Ranked words against common RCV1 topics.
- PV produces embeddings that reflect co-occurrence patterns.
- Document-wide context sampling highlights topical similarities.

Conclusions

- Argued that the window-based contexts of the Paragraph Vector model may have detrimental effect on the learned document embeddings.
- Proposed a Context Sampling Framework that allows for the instantiation of context policies of varying complexity.
- Achieved significant improvements over PV on multiple tasks & datasets.

Contact: s.angelidis@sms.ed.ac.uk