Learning Semantic Representations for Novel Words: Leveraging Both Form and Context

Timo Schick
Sulzer GmbH, Munich, Germany
timo.schick@sulzer.de

Hinrich Schütze
CIS, LMU Munich, Germany
inquiries@cislmu.org

Motivation

Why explicitly learn representations for novel words?

- Distributed word representations are a foundational aspect of many natural language processing systems
- Current approaches require many observations of a word for its embedding to become reliable; as a consequence, they struggle with small corpora and infrequent words
- As models are typically trained with a fixed vocabulary, they lack the ability to assign vectors to novel, out-of-vocabulary (OOV) words once training is complete

Motivation

Why use both form and context?

We should write no one off as being **unemployable**.

Motivation

Why use both form and context?

We should write no one off as being **unemployable**.

A **cardigan** is a knitted jacket or sweater with buttons up the front.

Motivation

Why use both form and context?

We should write no one off as being **unemployable**.

A **cardigan** is a knitted jacket or sweater with buttons up the front.

Unlike the grapefruit, the **pomelo** has very little importance in the marketplace.

The Form-Context Model

Unlike the grapefruit, the **pomelo** has very little importance in the marketplace.

$\mathbf{w} = \text{pomelo}$

The Form-Context Model

Unlike the grapefruit, the **pomelo** has very little importance in the marketplace.

$\mathbf{w} = \text{pomelo}$

$$\begin{aligned}\mathcal{S}_{\mathbf{w}} &= \{\langle s\rangle\text{p}, \text{po}, \text{om}, \text{me}, \text{el}, \text{lo}, \text{o}\langle e\rangle, \langle s\rangle\text{po}, \text{pom}, \text{ome}, \text{mel}, \text{elo}, \text{lo}\langle e\rangle\} \\ &= \{s_1, \dots, s_n\}\end{aligned}$$

The Form-Context Model

Unlike the grapefruit, the **pomelo** has very little importance in the marketplace.

$$\mathbf{w} = \text{pomelo}$$

$$\begin{aligned}\mathcal{S}_{\mathbf{w}} &= \{\langle s\rangle\text{p}, \text{po}, \text{om}, \text{me}, \text{el}, \text{lo}, \text{o}\langle e\rangle, \langle s\rangle\text{po}, \text{pom}, \text{ome}, \text{mel}, \text{elo}, \text{lo}\langle e\rangle\} \\ &= \{s_1, \dots, s_n\}\end{aligned}$$

$$\begin{aligned}\mathcal{C} &= \{\text{unlike}, \text{the}, \text{grapefruit}, \text{the}, \text{has}, \text{very}, \text{little}, \dots, \text{marketplace}\} \\ &= \{c_1, \dots, c_m\}\end{aligned}$$

The Form-Context Model

$$\mathcal{S}_{\mathbf{w}} = \{s_1, \ldots, s_n\}$$

$$\mathcal{C} = \{c_1, \ldots, c_m\}$$

The Form-Context Model

$\mathcal{S}_{\mathbf{w}} = \{s_1, \ldots, s_n\}$

$\mathcal{C} = \{c_1, \ldots, c_m\}$

$e_{\text{ngram}}(s_1) \quad \cdots \quad e_{\text{ngram}}(s_n)$

The Form-Context Model

$$\mathcal{S}_{\mathbf{w}} = \{s_1, \ldots, s_n\}$$

$$\mathcal{C} = \{c_1, \ldots, c_m\}$$

$$e_{\text{ngram}}(s_1) \quad \cdots \quad e_{\text{ngram}}(s_n)$$

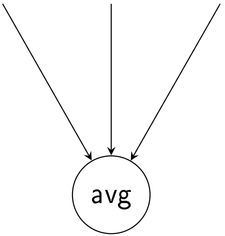
avg

$$v^{\text{form}}_{(\mathbf{w},\mathcal{C})}$$

The Form-Context Model

$\mathcal{S}_{\mathbf{w}} = \{s_1, \ldots, s_n\}$ $\mathcal{C} = \{c_1, \ldots, c_m\}$

$e_{\text{ngram}}(s_1)$ $\cdots$ $e_{\text{ngram}}(s_n)$ $e(c_1)$ $\cdots$ $e(c_m)$



$v^{\text{form}}_{(\mathbf{w},\mathcal{C})}$

The Form-Context Model

$\mathcal{S}_{\mathbf{w}} = \{s_1, \ldots, s_n\}$

$e_{\text{ngram}}(s_1) \quad \cdots \quad e_{\text{ngram}}(s_n)$

avg

$v^{\text{form}}_{(\mathbf{w},\mathcal{C})}$

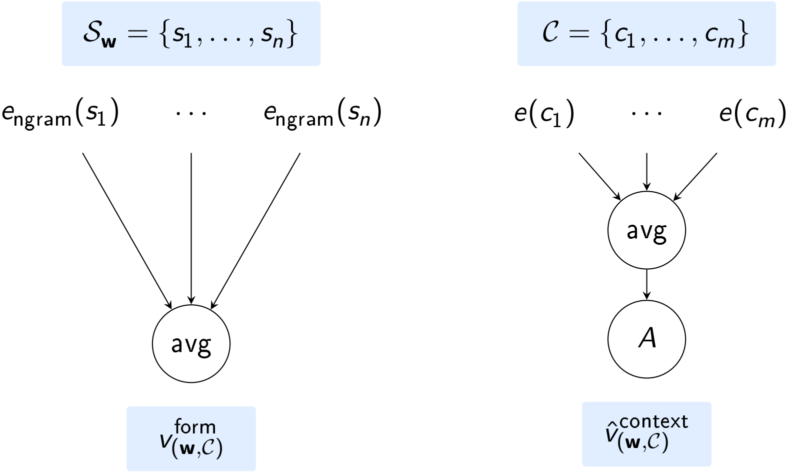
$\mathcal{C} = \{c_1, \ldots, c_m\}$

$e(c_1) \quad \cdots \quad e(c_m)$

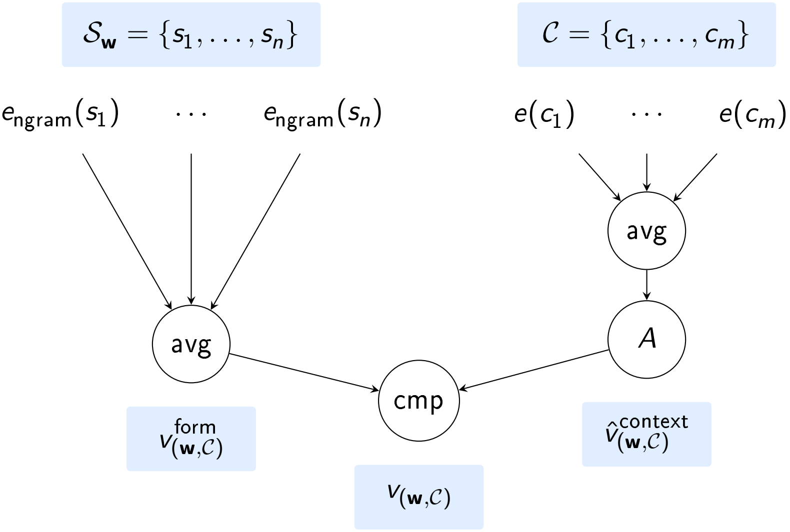
avg

$v^{\text{context}}_{(\mathbf{w},\mathcal{C})}$

The Form-Context Model



The Form-Context Model



Composition Functions

(i) single-parameter

$$v_{(\mathbf{w},\mathcal{C})} = \alpha \cdot \hat{v}^{\text{context}}_{(\mathbf{w},\mathcal{C})} + (1-\alpha) \cdot v^{\text{form}}_{(\mathbf{w},\mathcal{C})}.$$

with $\alpha \in [0,1]$ being a learnable parameter.

Composition Functions

(i) single-parameter

$$v_{(\mathbf{w},\mathcal{C})} = \alpha \cdot \hat{v}^{\text{context}}_{(\mathbf{w},\mathcal{C})} + (1-\alpha) \cdot v^{\text{form}}_{(\mathbf{w},\mathcal{C})}.$$

with $\alpha \in [0,1]$ being a learnable parameter.

(ii) gated

As above, except:

$$\alpha = \sigma(w^\top [v^{\text{context}}_{(\mathbf{w},\mathcal{C})} \circ v^{\text{form}}_{(\mathbf{w},\mathcal{C})}] + b)$$

with $w \in \mathbb{R}^{2k}$, $b \in \mathbb{R}$ being learnable parameters.

Training

$$\begin{aligned}\mathcal{B} &= \{(\mathbf{w}_1, \mathcal{C}_1), (\mathbf{w}_2, \mathcal{C}_2), \ldots, (\mathbf{w}_k, \mathcal{C}_k)\} \\ &= \{(\text{pomelo}, \{\text{unlike}, \text{the}, \text{grapefruit}, \ldots\}), (\mathbf{w}_2, \mathcal{C}_2), \ldots, (\mathbf{w}_k, \mathcal{C}_k)\}\end{aligned}$$

$$L_{\mathcal{B}} = \frac{1}{|\mathcal{B}|} \sum_{(\mathbf{w}, \mathcal{C}) \in \mathcal{B}} \| v_{(\mathbf{w}, \mathcal{C})} - e(\mathbf{w}) \|^2$$

Evaluation

We train the form-context model using skipgram embeddings trained on Wikipedia. To construct our training set, we

- consider all words **w** that occur at least 100 times;
- create $\mathcal{C}$ by randomly sampling 20 sentences from Wikipedia in which **w** occurs;
- create $\mathcal{S}_{\mathbf{w}}$ from all 3-, 4- and 5-grams of **w**, considering only *n*-grams that occur in at least 3 different words.

We evaluate the model on two tasks: the **Definitional Nonce Task** and the **Contextual Rare Words Task**.

The Definitional Nonce Task

spies most commonly refers to people who engage in spying, espionage or clandestine operations

The Definitional Nonce Task

spies most commonly refers to people who engage in spying, espionage or clandestine operations

	form	context	frm-ctx
neighbours	pies, cakes, spied, sandwiches	espionage, clandestine, covert, spying	espionage, spying, clandestine, covert
rank	668	8	6

The Definitional Nonce Task

hygiene which comes from the name of the greek goddess of health hygieia is a set of practices performed for the preservation of health

The Definitional Nonce Task

hygiene which comes from the name of the greek goddess of health hygieia is a set of practices performed for the preservation of health

	form	**context**	**frm-ctx**
neighbours	hygienic, hygiene, cleansers, hypoallergenic	hygieia, goddess, eileithyia, asklepios	hygienic, hygieia, health, hygiene
rank	2	465	4

The Definitional Nonce Task

perception (from the latin percipio) is the organization, identification and interpretation of sensory information in order to represent and understand the environment

The Definitional Nonce Task

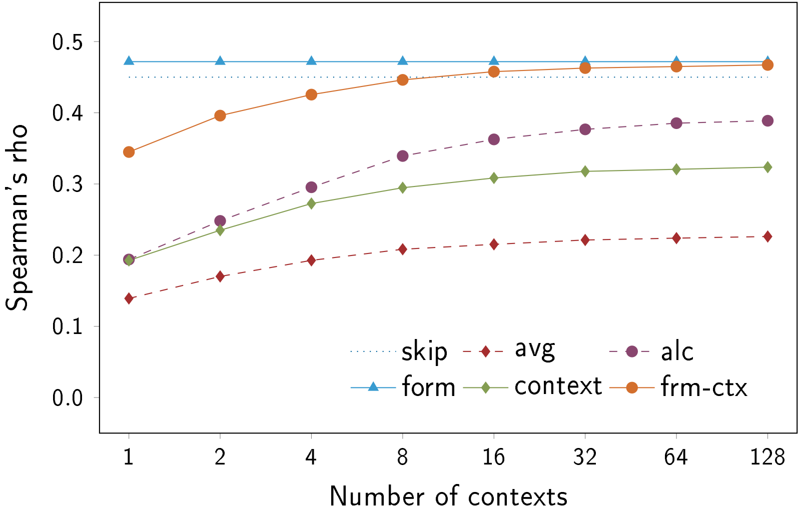
perception (from the latin percipio) is the organization, identification and interpretation of sensory information in order to represent and understand the environment

	form	context	frm-ctx
neighbours	interception, interceptions, fumble, touchdowns	sensory, perceptual, auditory, contextual	sensory, perceptual, perception, auditory
rank	115	51	3

The Definitional Nonce Task

Model	Type	Median Rank	MRR
Mimick	form	85573	0.00006
Skipgram	context	111012	0.00007
Additive	context	3381	0.00945
Nonce2Vec	context	623	0.04907
A La Carte	context	165.5	0.07058
surface-form	form	404.5	0.12982
context	context	184	0.06560
single-parameter	both	55	0.16200
gated	both	**49**	**0.17537**

The Contextual Rare Words Task



The Gated Model

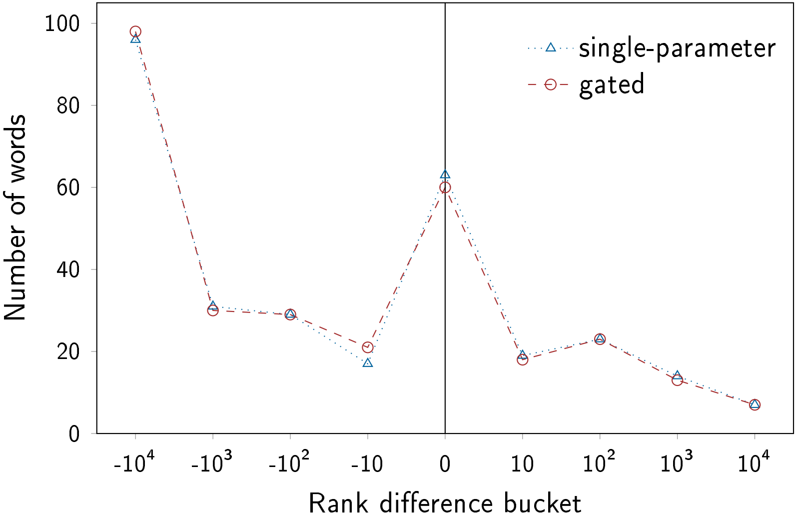
Words with high form weights:

cookstown, feltham, sydenham, wymondham, cleveland, banbury, highbury, shaftesbury

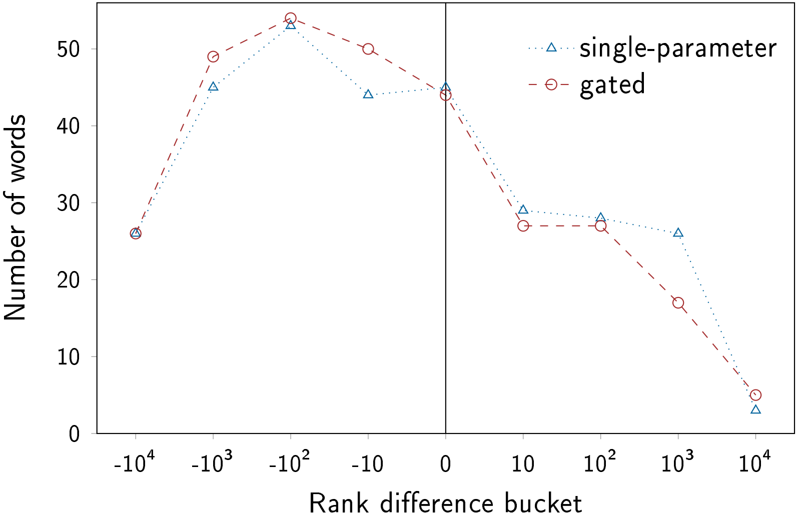
Words with high context weights:

poverty, hue, slang, flax, rca, bahia, atari, snooker, icq, bronze, esso

Adding Context Information



Adding Subword Information



Related Work

Bojanowski, P.; Grave, E.; Joulin, A.; and Mikolov, T. 2017. **Enriching word vectors with subword information**. *Transactions of the ACL*

Herbelot, A., and Braoni, M. 2017. **High-risk learning: acquiring new word vectors from tiny data**. In *Proceedings of the 2017 Conference on EMNLP*

Khodak, M.; Saunshi, N.; Liang, Y.; Ma, T.; Steward, B.; and Arora, S. 2018. **A la carte embedding: Cheap but effective induction of semantic feature vectors**. In *Proceedings of the 56th Annual Meeting of the ACL*

Pinter, Y.; Guthrie, R.; and Eisenstein, J. 2017. **Mimicking word embeddings using subword RNNs**. In *Proceedings of the 2017 Conference on EMNLP*

Conclusion and Future Work

The **form-context model** is capable of inferring high-quality representations for novel words by processing both the word's internal structure and words in its context.

Possible directions for future work include:

- investigating the model's performance for other languages;
- incorporating the number and informativeness of all available contexts into the composition function;
- using more complex ways than averaging to obtain surface-form and context embeddings.