What is best for spoken langage understanding: small but task-dependent embeddings or huge but out-of-domain embeddings?

Sahar Ghannay[1], Antoine Neuraz[1,2], Sophie Rosset[1]

[1] Université Paris-Saclay, CNRS, LIMSI, Orsay, France

[2] Department of Biomedical Informatics, Hôpital Necker-Enfants Malades, APHP INSERM UMRS 1138, Team 22, Paris Descartes, Université Sorbonne Paris Cit, Paris, France

ICASSP2020 Barcelona

Goal

- Focus on semantic evaluation of common word embeddings approaches for spoken language understanding task
  - with the aim of building a fast, robust, efficient and simple SLU system, to be integrated in a dialogue system.
- Investigate the use of two different data sets to train the embeddings: small and task-dependent corpus or huge and out of domain corpus.
- Evaluate different benchmark corpora ATIS, SNIPS, M2M, and MEDIA.

Natural/Spoken language understanding task

- Produce a semantic analysis and a formalization of the user's utterance.
- SLU is often divided into 3 sub-tasks: domain classification, intent classification, and **slot-filling (concept detection).**

Example:

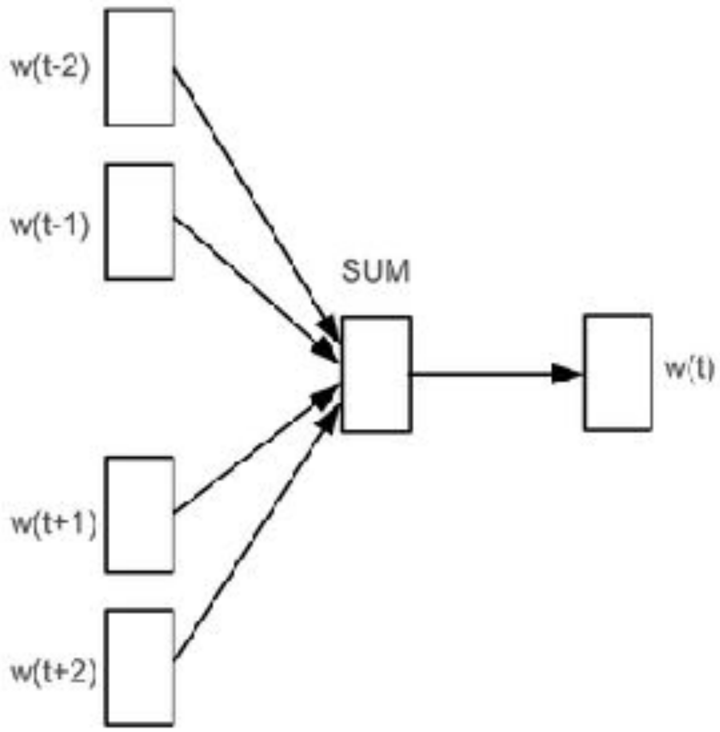
Hyp	wednesday	and	the	theater	Is	amc	Cupertino	square	16
Concept	B-date	O	O	O	O	B-theatre_name	I-theatre_name	I-theatre_name	I-theatre_name

Word Embeddings

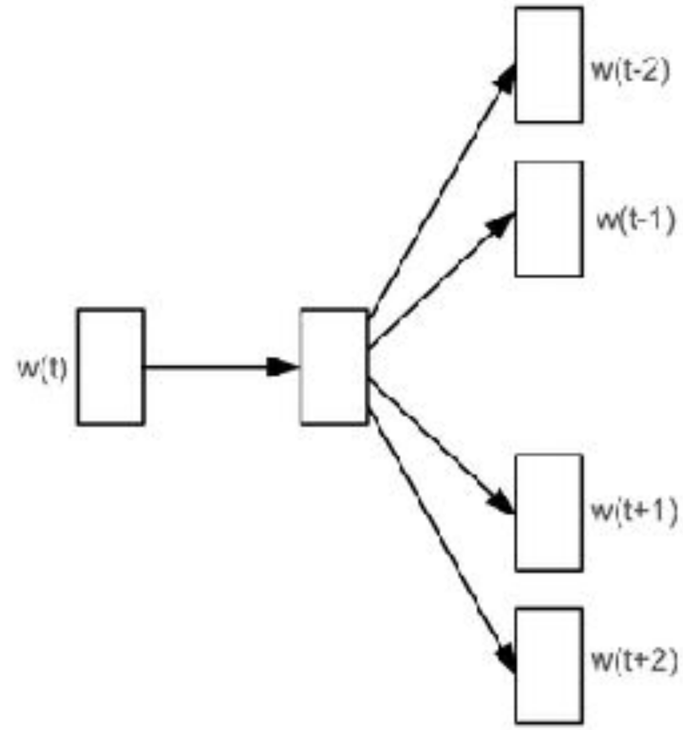
- Context independent embeddings:
  - Skip-gram, CBOW, GloVe, FastText.
- Contextual embeddings
  - ELMO.

Word Embeddings

Context independent

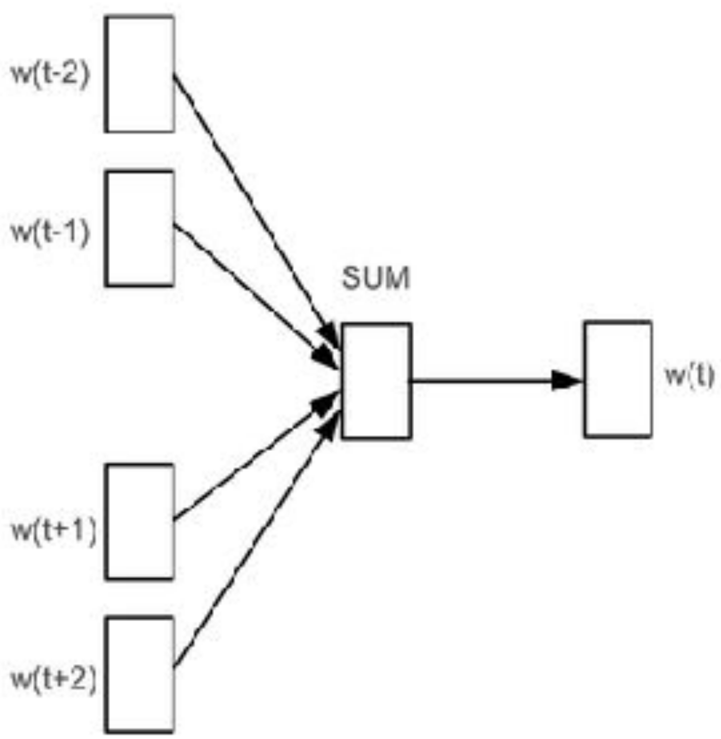


CBOW
[T. Mikolov *et al.* 2013]

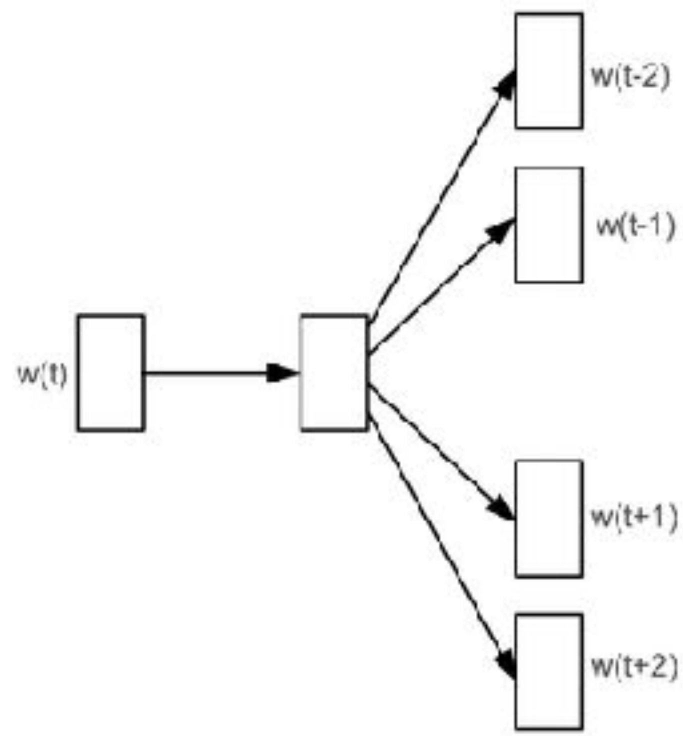


Skip-gram
[T. Mikolov *et al.* 2013]

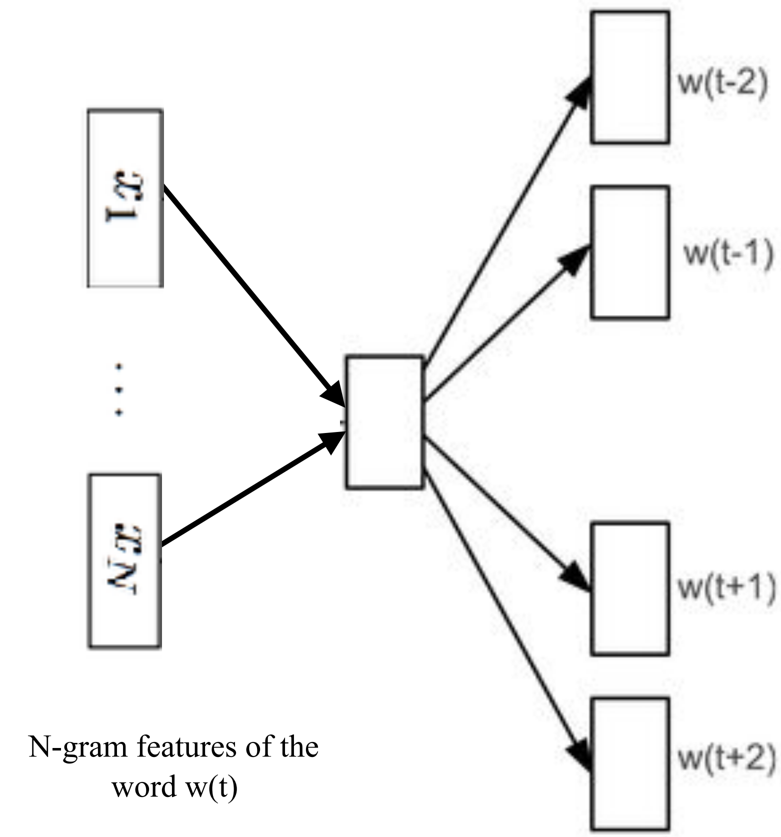
Word Embeddings
Context independent



CBOW
[T. Mikolov *et al.* 2013]

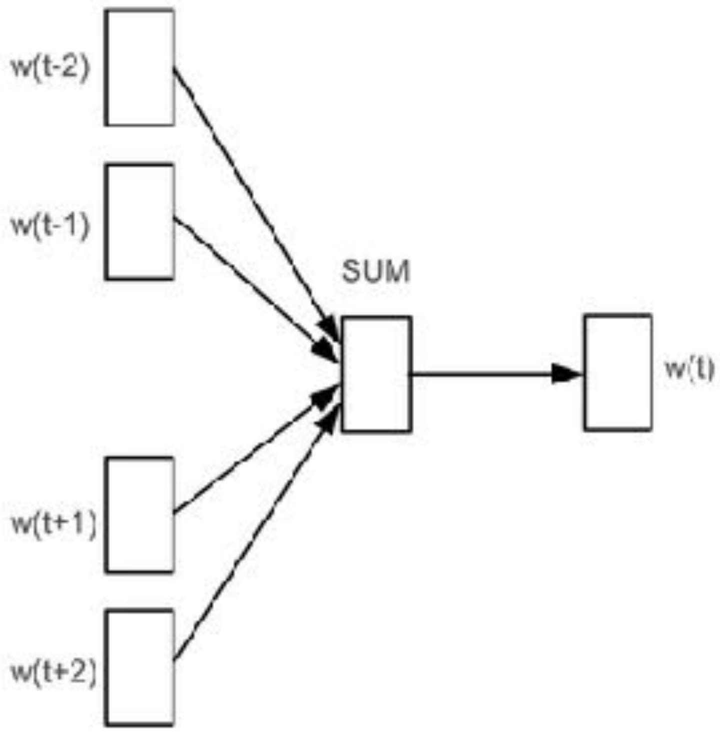


Skip-gram
[T. Mikolov *et al.* 2013]

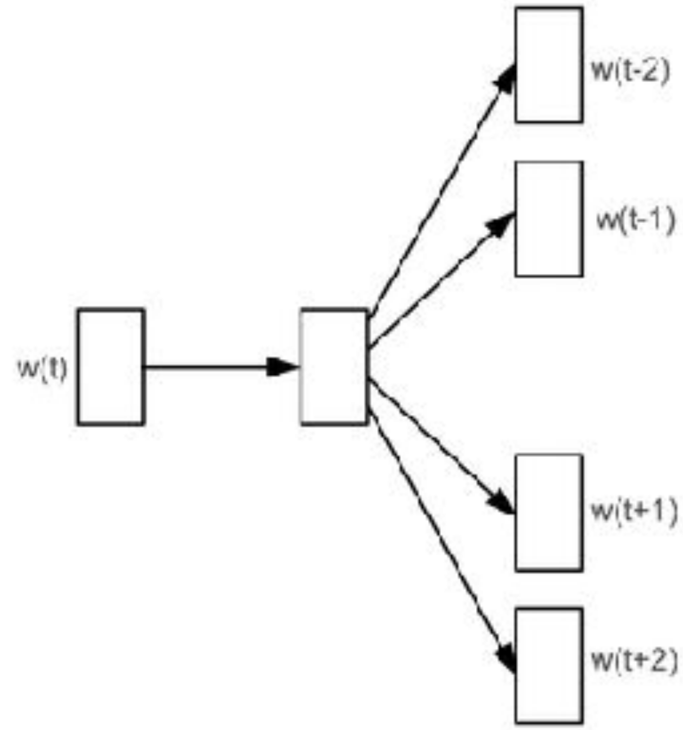


FastText
[P. Bojanowski et al. 2017]

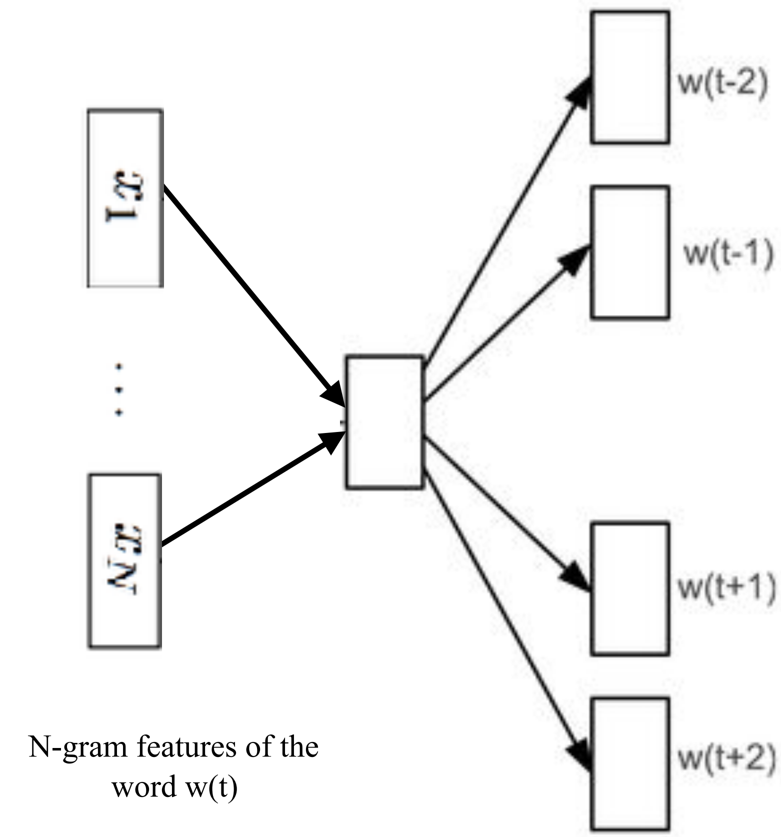
Word Embeddings
Context independent



CBOW
[T. Mikolov *et al.* 2013]

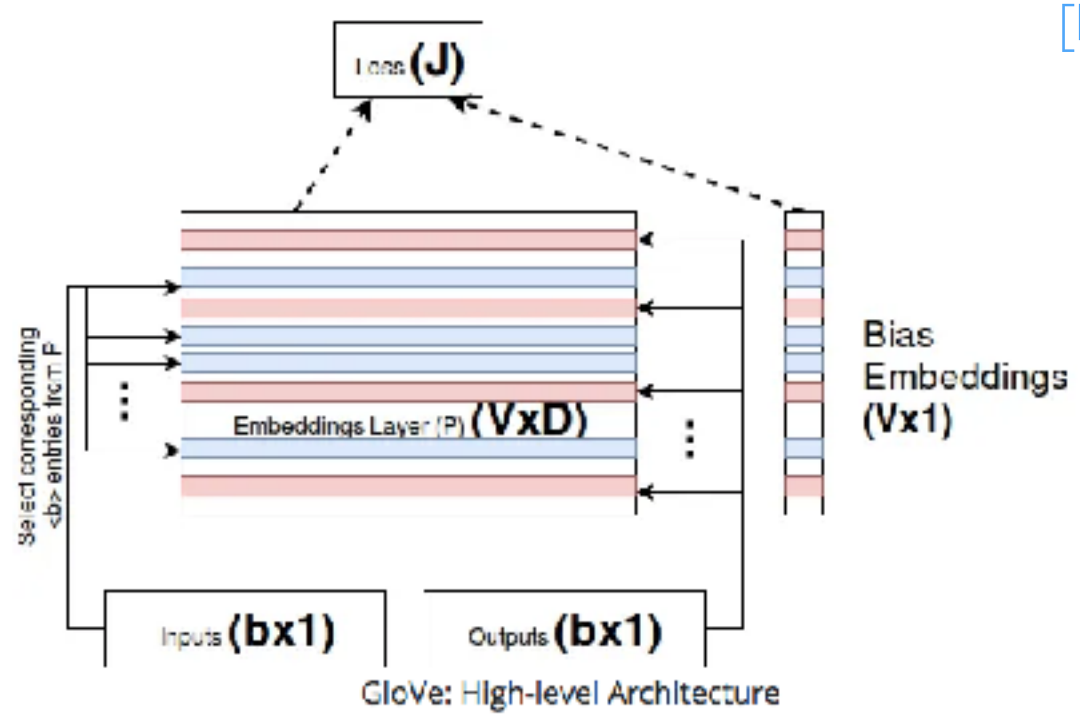


Skip-gram
[T. Mikolov *et al.* 2013]



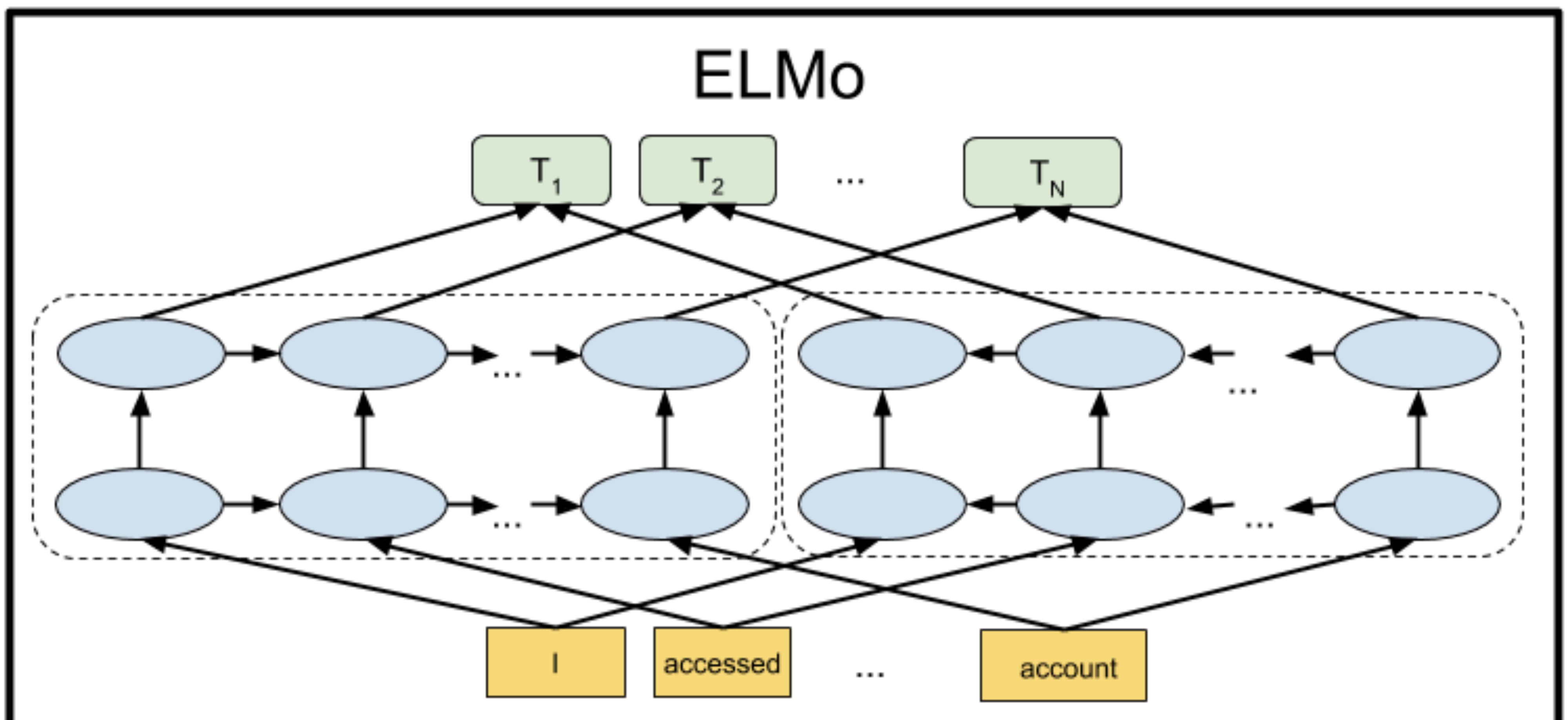
FastText
[P. Bojanowski et al. 2017]

GloVe : Count-based approach
[J. Pennington *et al.* 2014]



Contextual Word Embeddings

- Embeddings from Language Models: **ELMo** [Matthew E Peters *et al.* 2019]
  - Learn word embeddings through building *bidirectional language models* (biLMs).
    - biLMs consist of forward and backward LMs.



Contextual Word Embeddings

- ELMo can models:
  - Complex characteristics of word use (e.g., syntax and semantics).
  - How these uses vary across linguistic contexts (i.e., to model polysemy).
- ELMo differ from previous word embeddings approaches:
  - Each token is assigned a representation.

Experimental setup

Data:

- ATIS: concerns flight information.
- MEDIA: hotel reservation and information.
- M2M: restaurant and movie ticket booking.
- SNIPS: multi-domain dialogue corpus collected by the SNIPS company: 7 in-house tasks such as Weather information, restaurant booking, managing playlist, etc.
- SNIPS70: sub-part of the SNIPS corpus, in which the training set is limited to 70 queries per intent randomly chosen.

Corpus	ATIS	MEDIA	SNIPS	SNIPS70	M2M
vocab.	1117	2463	14354	4751	900
#tags	84	70	39	39	12
train size	4978	12908	13784	2100	8148
test size	893	3518	700	700	4800

Experimental setup

Word embeddings training: data

- Studying the impact of the corpora used to train the embeddings:
  - **small but task-dependent (in domain) corpus.**
  - **huge but out-of-domain corpus (wiki data):**
    - French wiki data composed of 573 million of words.
    - English wiki data composed of 2 billion of words.
      - words occurring less than 5 times have been discarded:
        - resulting in a vocabularies sizes of 923k words for French and for 2 million words for English.

Experimental setup

Word embeddings training: hyper-parameters

- Skip-gram, CBOW, Glove and Fasttext :
  - window size = 5, negative sampling = 5, dimension = 300.
- ELMO : weighted average of all biLM layers
  - **trained on small but task dependent corpus:**
    - Default parameters : dimension=1024.
  - **trained on huge but out-of-domain corpus:**
    - using pre-trained models form ELMoForManyLangs lib [Che, Wanxiang *et al.* 2018], dimension=1024.
    - trained on 20-million-words data randomly sampled from the raw text released by the CoNLL 2018 shared task.

Experimental setup

SLU model

- Bi-LSTM
  - Composed of 2 hidden layers.
  - hyper-parameters tuning :
    - the size of the BiLSTM hidden layers $n \in \{128, 256, 512\}$.
    - the batch size $b \in \{16, 32, 64\}$.
  - Fed with only word embeddings of size $d \in \{1024, 300\}$.
    - Embeddings are are not tuned during training.

Experimental setup

Evaluation metrics

- The results are evaluated using the standard evaluation metrics:
  - F-measure F1 computed by conlleval evaluation script that consider a segment correct if both boundaries and class are correct.
- We used Wilcoxon signed-rank test [5] to evaluate the significance of the results. The result is significant if the P-value [6] is lower than 0.05.

Experimental results

Quantitative evaluation:

	task-dependent					Out-of-domain				
Bench.	**ELMo**	**FastText**	**GloVe**	**Skip-gram**	**CBOW**	**ELMo**	**FastText**	**GloVe**	**Skip-gram**	**CBOW**
M2M	88.89	72.13	**92.54**	88.87	89.39	91.14	93.01	91.77	**93.19**	92.13
ATIS	**94.38**	85.72	92.95	90.84	91.87	94.93	95.52	95.35	95.62	**95.77**
SNIPS	78.68	76.35	**87.40**	82.10	83.94	90.29	**94.85**	93.90	94.43	94.05
SNIPS70	53.06	38.19	**63.65**	47.11	49.76	75.19	79.75	78.68	78.90	**80.13**
MEDIA	80.26	71.73	**82.66**	80.01	79.57	**86.42**	85.30	85.11	85.95	86.06

Tagging performance of different word embeddings trained on task-dependent corpus (ATIS, MEDIA, M2M, SNIPS or SNIPS70) and on huge and out of domain corpus (WIKI English or French) on all benchmark corpora in terms of F1 using conlleval scoring script (in %)

- ✓ The embeddings trained on huge and out-of-domain corpus yields to better results than the ones trained on small and task-dependent corpus
- ✓ Context independent approaches outperform significantly the contextual embeddings when they are trained on out-of-domain corpus except for MEDIA

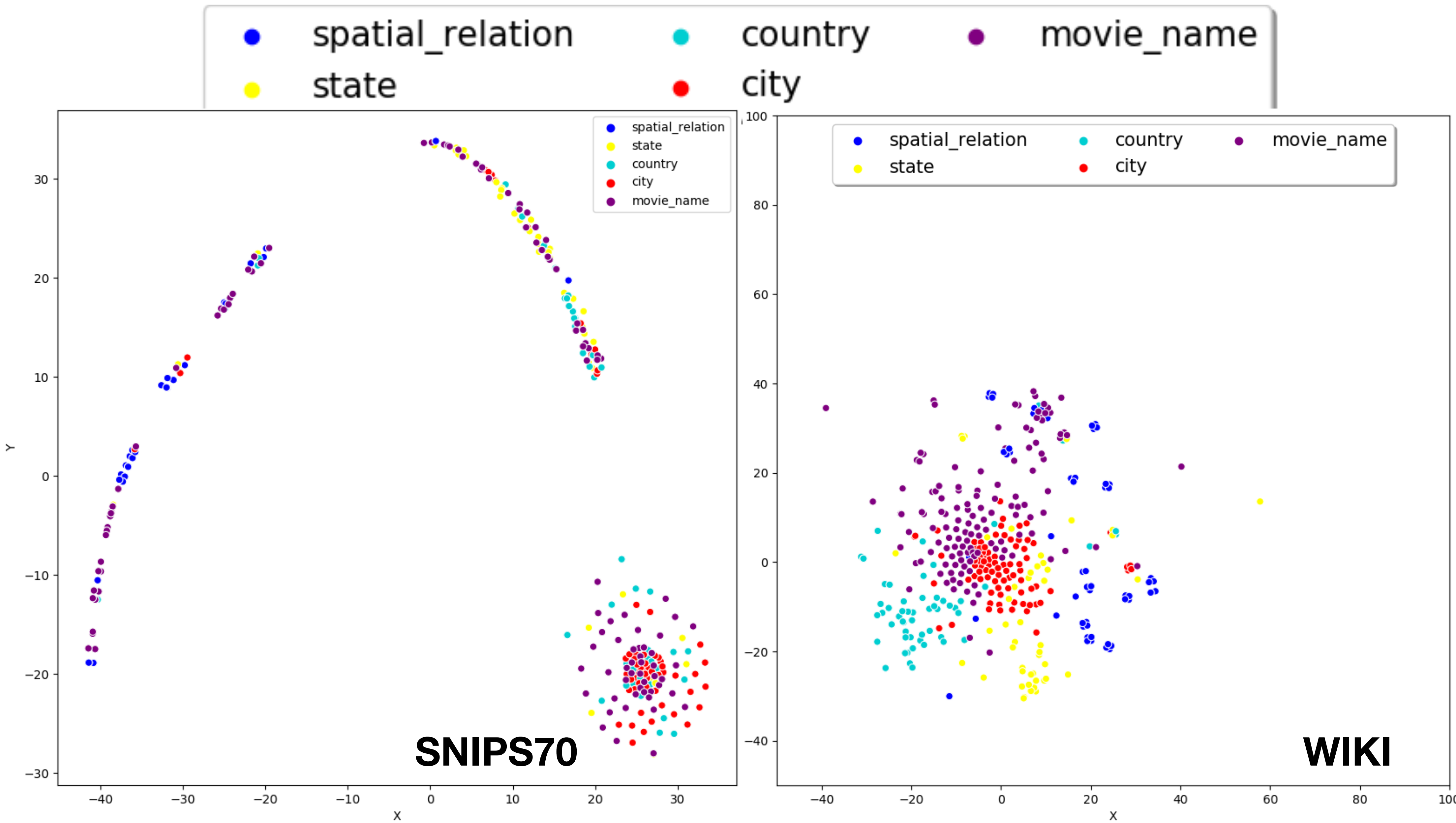
Experimental results

Qualitative evaluation:

- Perform a visual evaluation of the word representations.
- For a given method and task, we compared the t-SNE obtained using embeddings learned on a small but in-domain corpus versus a large but out-of-domain corpus (WIKI).
- This visual evaluation concerns the words that carry out frequent semantic tags that have an F1 score lower than the median.

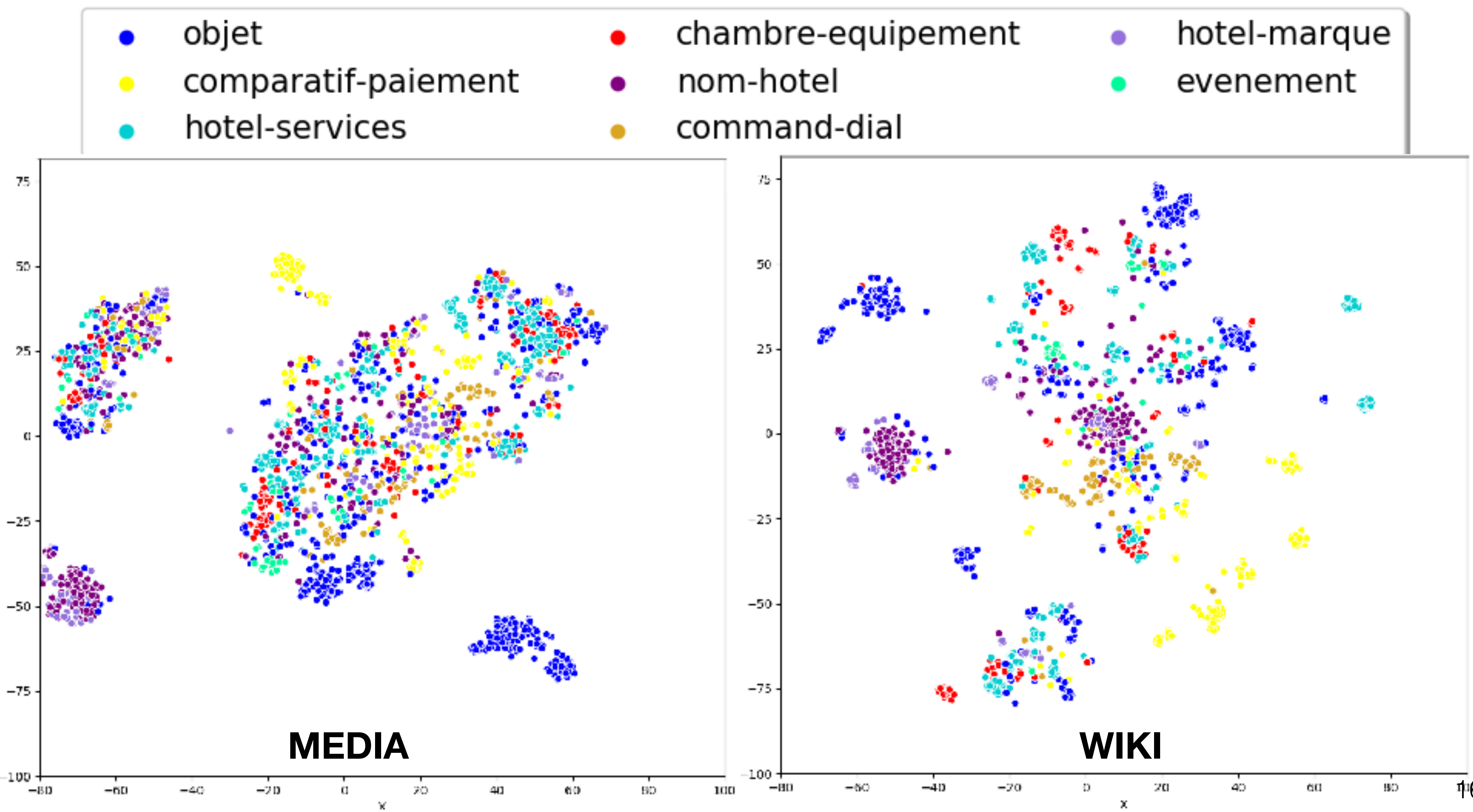
Experimental results

Qualitative evaluation: CBOW



Experimental results

Qualitative evaluation: ELMo



Experimental results

Computation time:

- For training and test time, we observe that ELMo is the slowest one
  - we can avoid training time by using pre-trained models.
- For MEDIA, ELMo (86.48) achieves the best results followed by CBOW (86.06) which is the fastest in terms of train and test time.
- As for dialog system the SLU model has to be simple, robust, efficient and fast, in this case CBOW is the adequate approach we can use.

Conclusions

- Evaluation of different word embeddings approaches (ELMo, FastText, GloVe, Skip-gram and CBOW) on SLU task.
  - small and task-dependent corpus VS huge and out-of-domain corpus.
  - 5 benchmark corpora: ATIS, SNIPS, SNIPS70, M2M, and MEDIA.
- Embeddings trained on huge and out-of-domain corpus yields to better results than the ones trained on small and task-dependent corpus.
- Count-based approaches like GloVe are not impacted by the lack of data.
  - CBOW, Skip-gram and especially FastText need more data for training to be efficient.
- Context independent approaches outperform the contextual embeddings (ELMo) when they are trained on out-of-domain corpus except for MEDIA.
- The obtained results are interesting, since the embeddings are not tuned during training and we are not using additional features, so those results can be easily improved.
- ELMo is the slowest one in terms of train and test time
  - for downstream tasks (e.g. dialog system), it is preferable to use the fastest embedding model that achieves good performance.

References

[T. Mikolov et al. 2013]:Tomas Mikolov, Kai Chen, Greg Corrado, and Jef- frey Dean. 2013a. Efficient Estimation of Word Representations in Vector Space. In ICLR Workshop Papers, 2013.

[J. Pennington et al. 2014]: Jeffrey Pennington, Richard Socher, and Christopher D Manning, "Glove: Global vectors for word representation.," in EMNLP, 2014, vol. 14.

[P. Bojanowski et al. 2017]: Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov, "Enriching word vectors with subword information," Transactions of the Association for Computational Linguistics, vol. 5, 2017.

[Matthew E Peters et al. 2019]: Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gard- ner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer, "Deep contextualized word representations," arXiv preprint arXiv:1802.05365, 2018.

[Che, Wanxiang et al. 2018]: Che, W., Liu, Y., Wang, Y., Zheng, B., & Liu, T. (2018). Towards better UD parsing: Deep contextualized word embeddings, ensemble, and treebank concatenation, Association for Computational Linguistics, 2018

Thank you !