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

Sahar Ghannay, Antoine Neuraz, Sophie Rosset

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

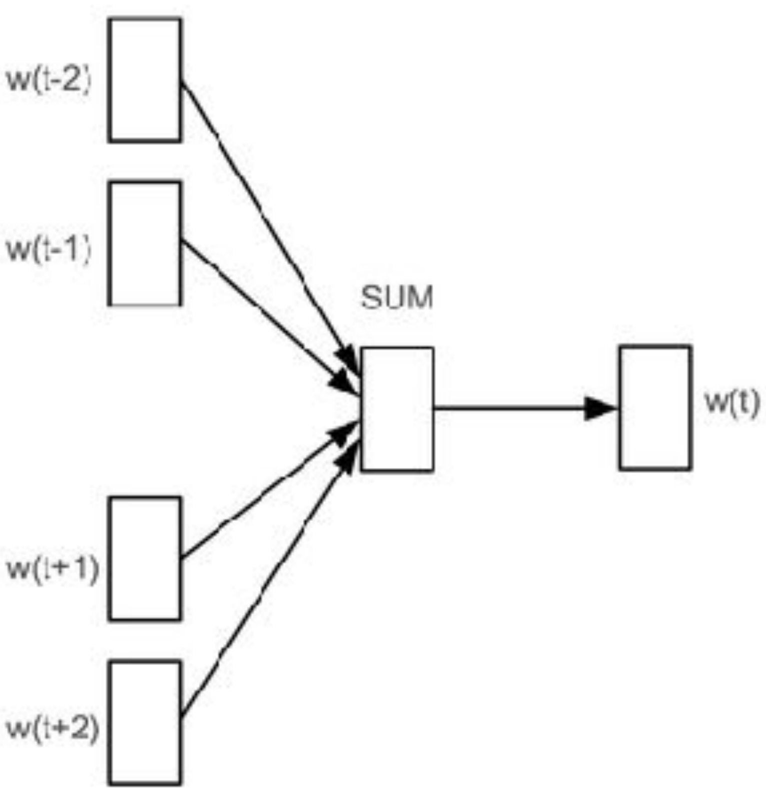
- Example

Hyp	je	veux	réserver	une	chambre
Concept	commande			nombre	objet
Label	commande-B	commande-I	commande-I	nombre-B	objet-B
Valeur	réservation			1	chambre

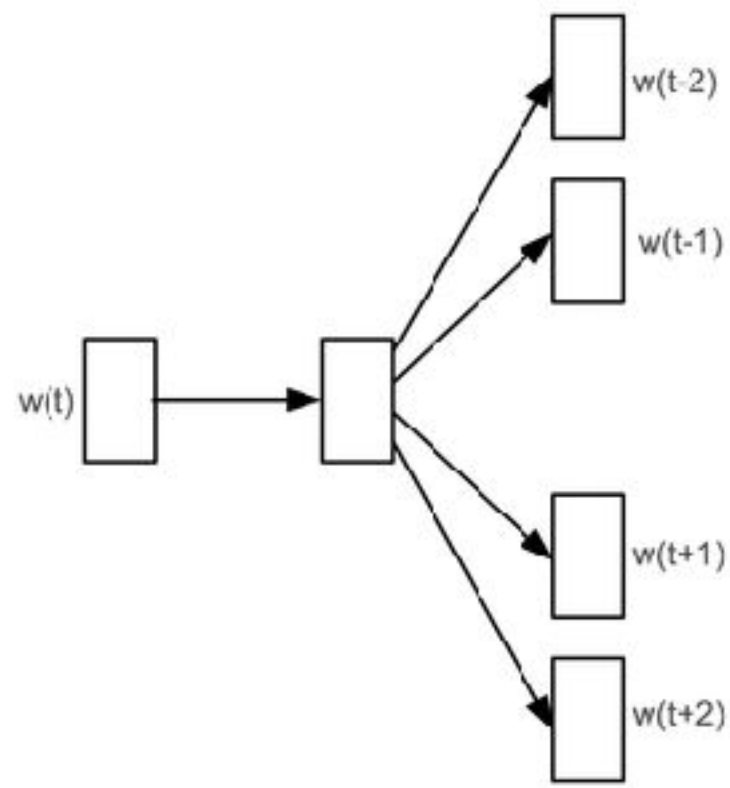
Word Embeddings

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

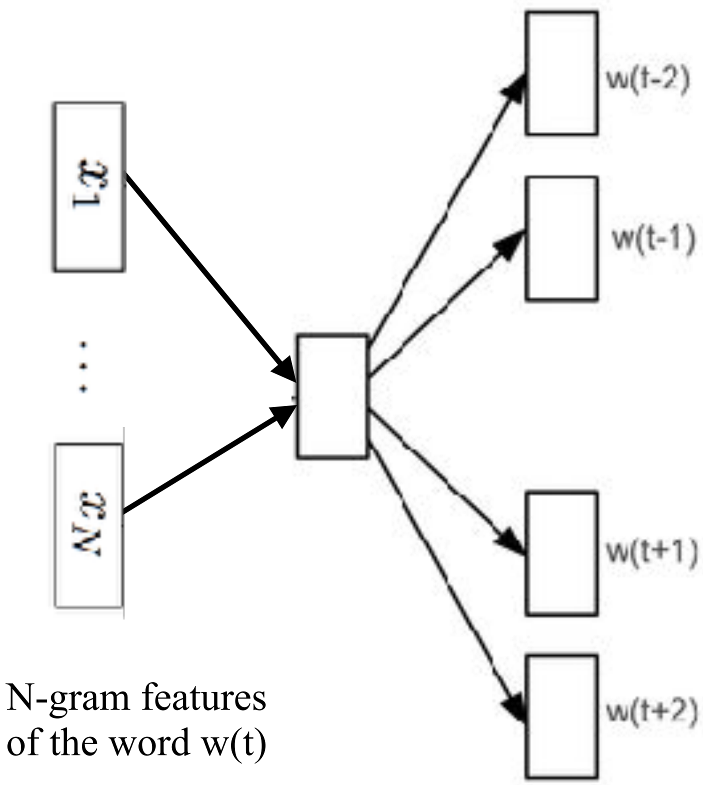
Word Embeddings
Context independent



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



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



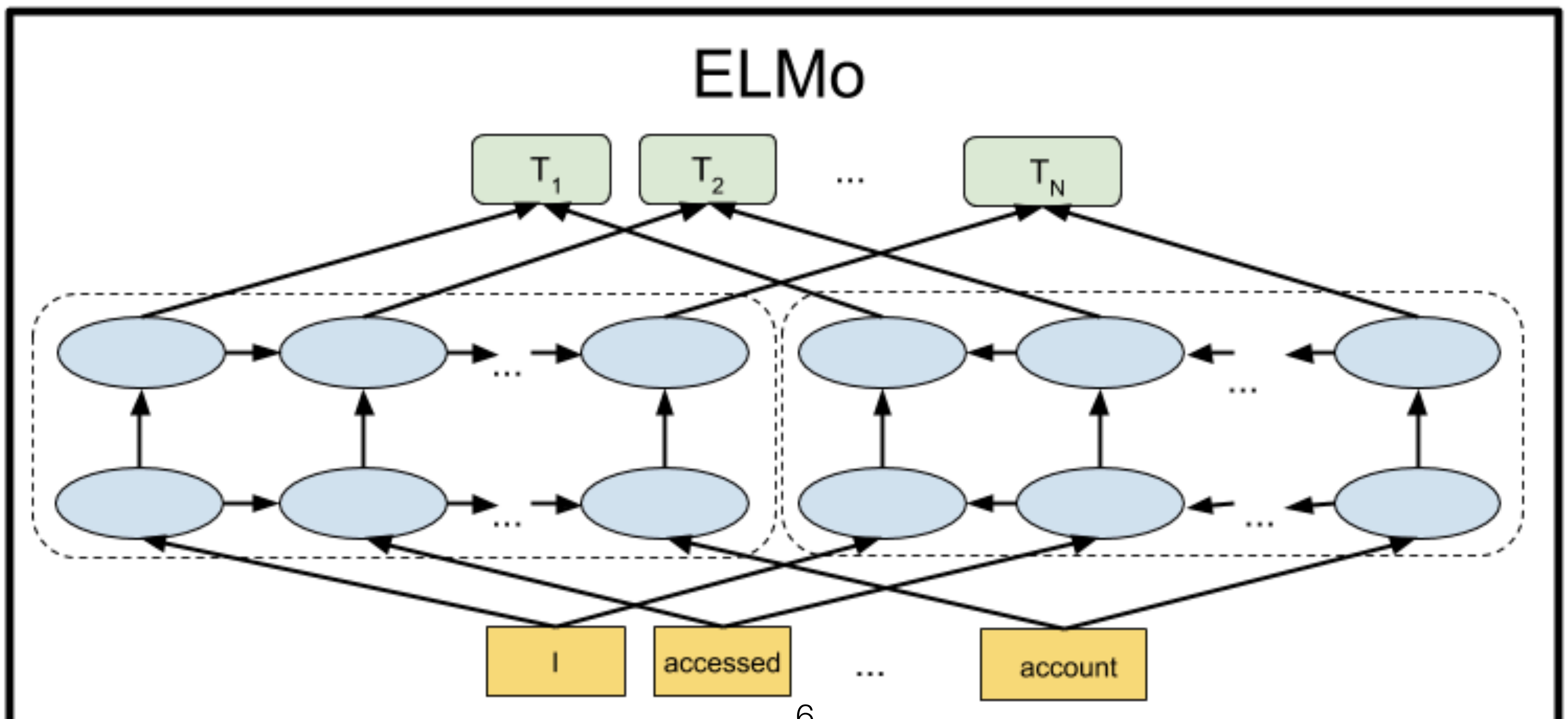
FastText
[P. Bojanowski *et al.* 2017]

GloVe [J. Pennington *et al.* 2014]

- Calcul d'une matrice de co-occurence X
- Factorisation de X pour obtenir les word embeddings

Contextual Word Embeddings

- Embeddings from Language Models: **ELMo**
  - 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

Experiments : Data and Results

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	2445	14354	4751	900
#tags	84	70	39	39	12
train size	4978	12908	13784	2100	8148
test size	893	3005	700	700	4800

Experiments : Data and Results

Word embeddings training:

- Studying the impact of the corpora used to train the embeddings:
  - small and task-dependent corpus
  - huge and out-of-domain corpus.
    - ELMo: using pre-trained models

Experiments : Data and Results

SLU model

- b-LSTM
  - Composed of 2 hidden layers
  - Fed with only word embeddings

Experiments : Data and Results

Quantitative evaluation:

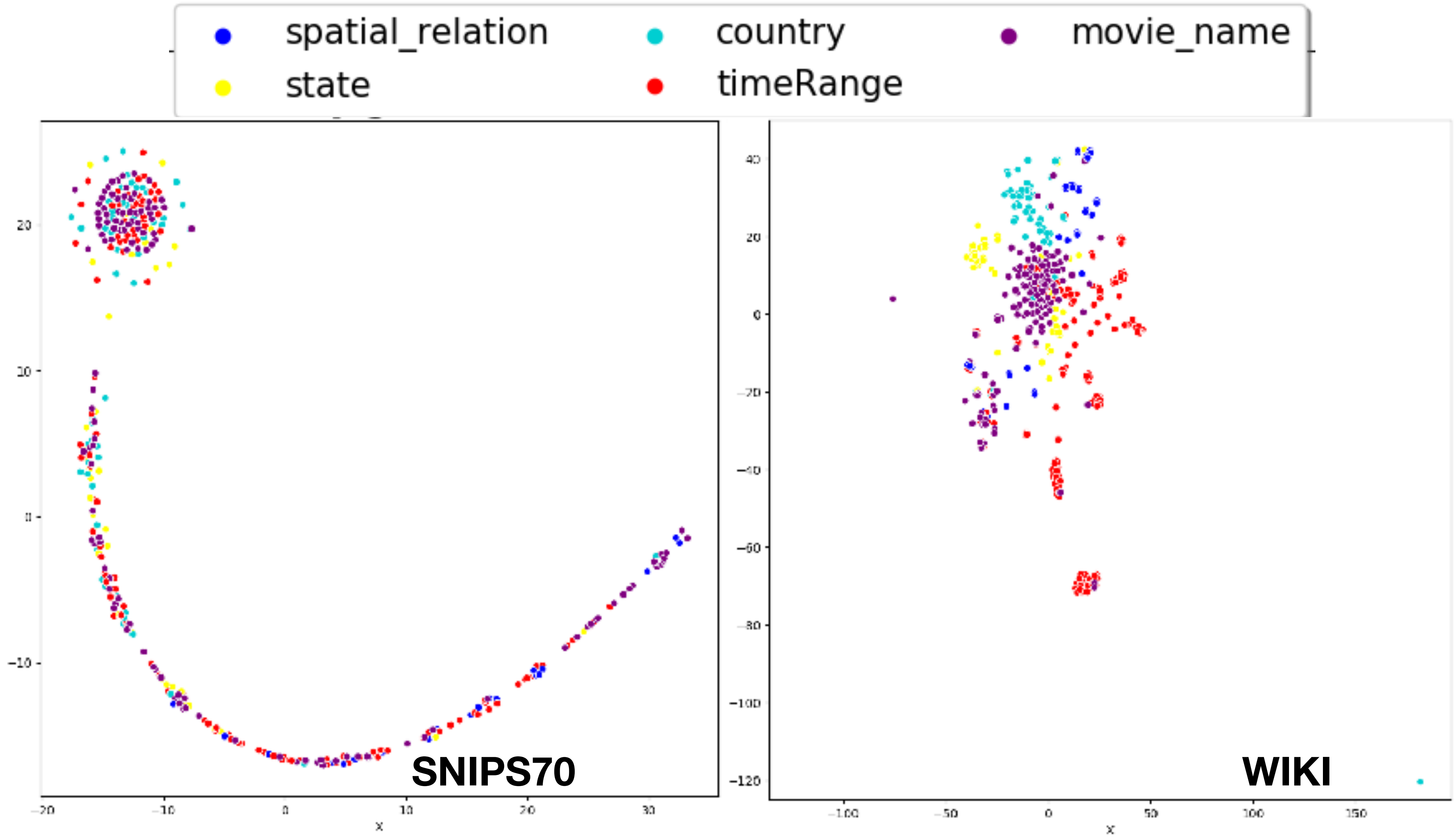
	task-dependent					Out-of-domain				
Bench.	**ELMo**	**FastText**	**GloVe**	**Skip-gram**	**CBOW**	**ELMo**	**FastText**	**GloVe**	**Skip-gram**	**CBOW**
M2M	88.89	72.13	**92.54**	88.87	89.39	91.14	93.01	91.77	**93.19**	92.13
ATIS	**94.38**	85.72	92.95	90.84	91.87	94.93	95.52	95.35	95.62	**95.77**
SNIPS	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

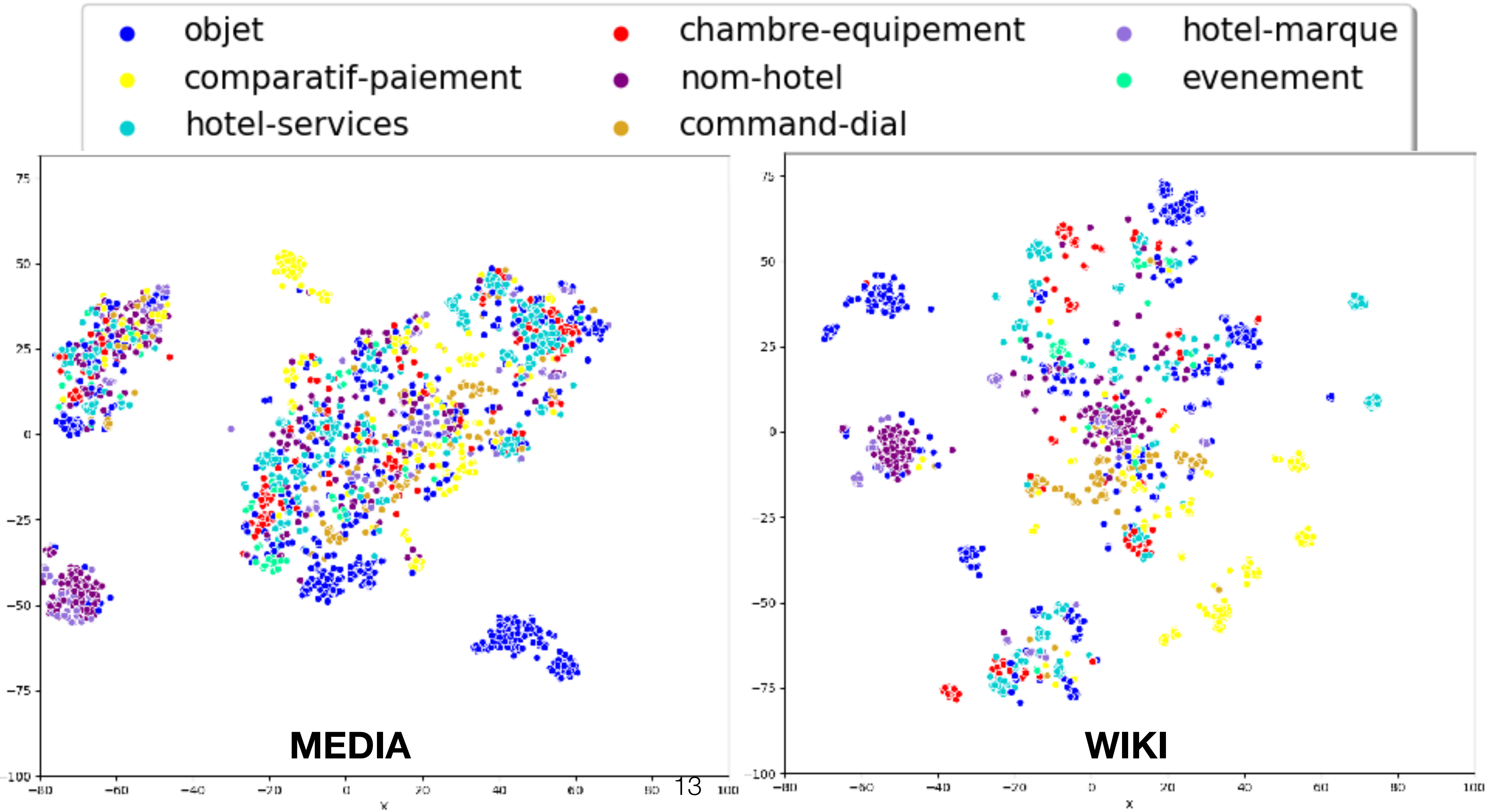
Experiments : Data and Results

Qualitative evaluation:
Skip-gram



Experiments : Data and Results

Qualitative evaluation:
ELMo



Experiments : Data and 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 achieves the best results followed by CBOW 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 on SLU task
- 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
- 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 and test time, and for downstream tasks (*e.g.* dialog system), it is preferable to use the fastest embedding model that achieves good performance.

Thank you !