ETNLP: a visual-aided systematic approach to select pre-trained embeddings for a downstream task

Xuan-Son Vu[1], Thanh Vu[2], Son N. Tran[3], Lili Jiang[1]

[1]Department of Computing Science, Umeå University, Sweden
[2]The Australian E-Health Research Centre, CSIRO, Australia
[3]The University of Tasmania, Australia
{sonvx; lili.jiang}@cs.umu.se; thanh.vu@csiro.au; sn.tran@utas.edu.au

UMEÅ UNIVERSITY

Database & Data Mining Research Group

INTRODUCTION

Given many recent advanced embedding models, selecting pre-trained word embedding (a.k.a., word representation) models best fit for a specific downstream task is non-trivial. In this paper, we propose a systematic approach, called ETNLP, for extracting, evaluating, and visualizing multiple sets of pre-trained word embeddings to determine which embeddings should be used in a downstream task.

- Codes and Data: https://github.com/vietnlp/etnlp
- Demo Video: https://vimeo.com/317599106

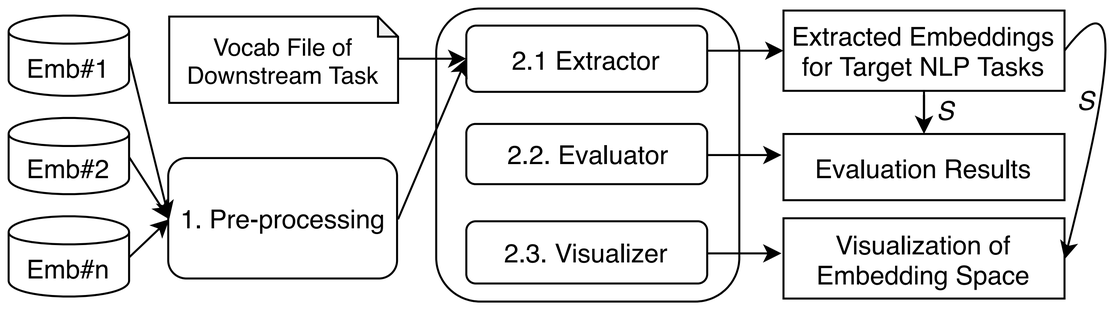

Figure 1 – General process of the ETNLP pipeline where $\mathcal{S}$ is the set of extracted embeddings for Evaluation and Visualization of multiple embeddings on a downstream NLP task.

PIPELINE DEMONSTRATION

We demonstrate the effectiveness of the proposed approach on our pre-trained word embedding models in Vietnamese to select which models are suitable for a named entity recognition (NER) task.

1. Extractor: extract embeddings for a downstream task

```
$python3 etnlp_api.py  -input      "<emb_in#1>;<emb_in#2>"
                       -input_c2v <emb_in#3>
                       -vocab <file>
                       -output <out_file.gz>
                       -args extract;solveoov:1
```

Figure 2 – Run *extractor* to export single or multiple embeddings for NLP tasks.

2. Evaluator: evaluate embeddings for a downstream task

Vietnamese	English
ông nội I bà ngoại I ông I bà	grandfather I grandmother I grandpa I grandma
ông nội I bà ngoại I vua I nữ_hoàng	grandfather I grandmother I king I queen

```
$python3 etnlp_api.py  -input "<emb_in#1>;<emb_in#2>"
                       -analoglist <file>
                       -output <eval_results> -args eval
```

Figure 3 – Run *evaluator* on multiple word embeddings on the word analogy task.

3. Visualizer: visualize embeddings for final decision

```
$python3 etnlp_api.py  -input      "<emb_in#1>;<emb_in#2>"
                       -args visualizer
```

Figure 4 – Run *visualizer* to explore given pre-trained embedding models.

Remarks: For visualization, the ETNLP framework offers two different ways to visualize multiple pre-trained embeddings called (1) zoom-out (the side-by-side visualization) and (2) zoom-in (the interactive visualization). For more information, please see the github repository.

EVALUATION RESULTS

Table 1 – Performance of the NER task using different embedding models. The $MULTI_{WC\_F\_E\_B}$ is the concatenation of four embeddings: W2V_C2V, fastText, ELMO, and Bert_Base. "wemb dim" is the dimension of the embedding model. VnCoreNLP* means we retrain the VnCoreNLP with our pre-trained embeddings.

	F1	wemb dim	cemb dim	drpt	lstm-s	lrate
BiLC3 (Ma and Hovy, 2016)	88.28	300	-	-	-	-
VNER (Dong and Nguyen, 2018)	89.58	300	300	0.6	-	0.001
VnCoreNLP (Vu et al., 2018a)	88.55	300	-	-	-	-
VnCoreNLP (*)	**91.30**	1024	-	-	-	-
BiLC3 + W2V	*89.01*	300	50	0.5	100	0.0005
BiLC3 + BERT-Base	*88.26*	768	500	0.3	100	0.0005
BiLC3 + W2V_C2V	*89.46*	300	100	0.5	500	0.0005
BiLC3 + fastText	89.65	300	500	0.3	100	0.001
BiLC3 + ELMO	89.67	1024	100	0.7	500	0.0005
BiLC3 + $MULTI_{WC\_F\_E\_B}$	**91.09**	2392	100	0.7	100	0.001

CONCLUSION

We have presented a new systematic pipeline, ETNLP, for extracting, evaluating and visualizing multiple pre-trained embeddings on a specific downstream task.

- The pipeline is easy to apply on any language processing task.
- It allows users to easily construct and explore data over heterogeneous data sources.
- Be able to handle unknown vocabulary in real-world data (using C2V).
- Evaluated on (1) Vietnamese NER task and (2) privacy- guaranteed embedding selection task showed its effectiveness.

REFERENCES

- **Xuan-Son Vu**, Son N. Tran, Lili Jiang, "dpUGC: Learn Differentially Private Representationfor User Generated Contents", *In: Proceedings of the 20th International Conference on Computational Linguistics and Intelligent Text Processing, April, 2019, to appear,* ***3rd place for best paper awards***.
- Thanh Vu, Dat Quoc Nguyen, **Xuan-Son Vu**, Dai Quoc Nguyen, Michael Catt, Michael Trenell, "NIHRIO at SemEval-2018 Task 3: A Simple and Accurate Neural Network Model for Irony Detection in Twitter", *In: Proceedings of NAACL-HTL'18, at the 12nd International Workshop on Semantic Evaluation (SemEval-2018), New Orleans, USA*.

This research is funded by Umeå University in Sweden on federated database research. The experiments were partially run on resources provided by SNIC Science Cloud at HPC2N center.