Hicham El Boukkouri Thesis summary

Subject : Domain adaptation of word embeddings thought the exploitation of in-domain corpora and knowledge bases

Supervisors : Oliver Ferret, Thomas Lavergne, Pierre Zweigenbaum

- Context :
  - Static embeddings VS. Contextualized embeddings -> similar words, similar vectors
  - need a lot of text : the result is general but not optimal for a specific domain
  - Need to improve it for specialized domain

  The idea is to train a general-domain representation

1. Entity recognition - embeddings / corpus experiences

i2b2/VA 2010 Clinical Concept Detection

- different concept to extract and recognize : problem, treatments and test
- comparison of embeddings : static (word2vec) / contextualized (Elmo)
- general domain (Wikipedia/ Gigawords) VS small in-domain corpus (MIMICIII and PubMed)
- A deep-learning pipeline : Bi-LSTM and CRF model for prediction

Results :

- Wiki > Gigaword
- in domain > general domain
- large in domain > small in domain
- MIMICIII > PubMed

See it here : https://aclanthology.org/P19-2041.pdf

2. Improving General domain with small in-domain corpus

BioBERT / SciBERT / BlueBERT are result of a general pretrained BERT fine tuned. But is a specialized vocabulary step useful for increasing performance of a model ?

training BERT from scratch by integrating a different tokenization phase called the "worpieces" for out of domain integration

Still experiments about embeddings types / domain corpus

-> comparison of BERT training : vocabulary type | initial corpus | specialization corpus

-> test of combinations of corpus for training steps

3 tasks : i2b2/VA (concept detection) - textual implication (contradiction/entailment/neutral) - relation extractions

Results :

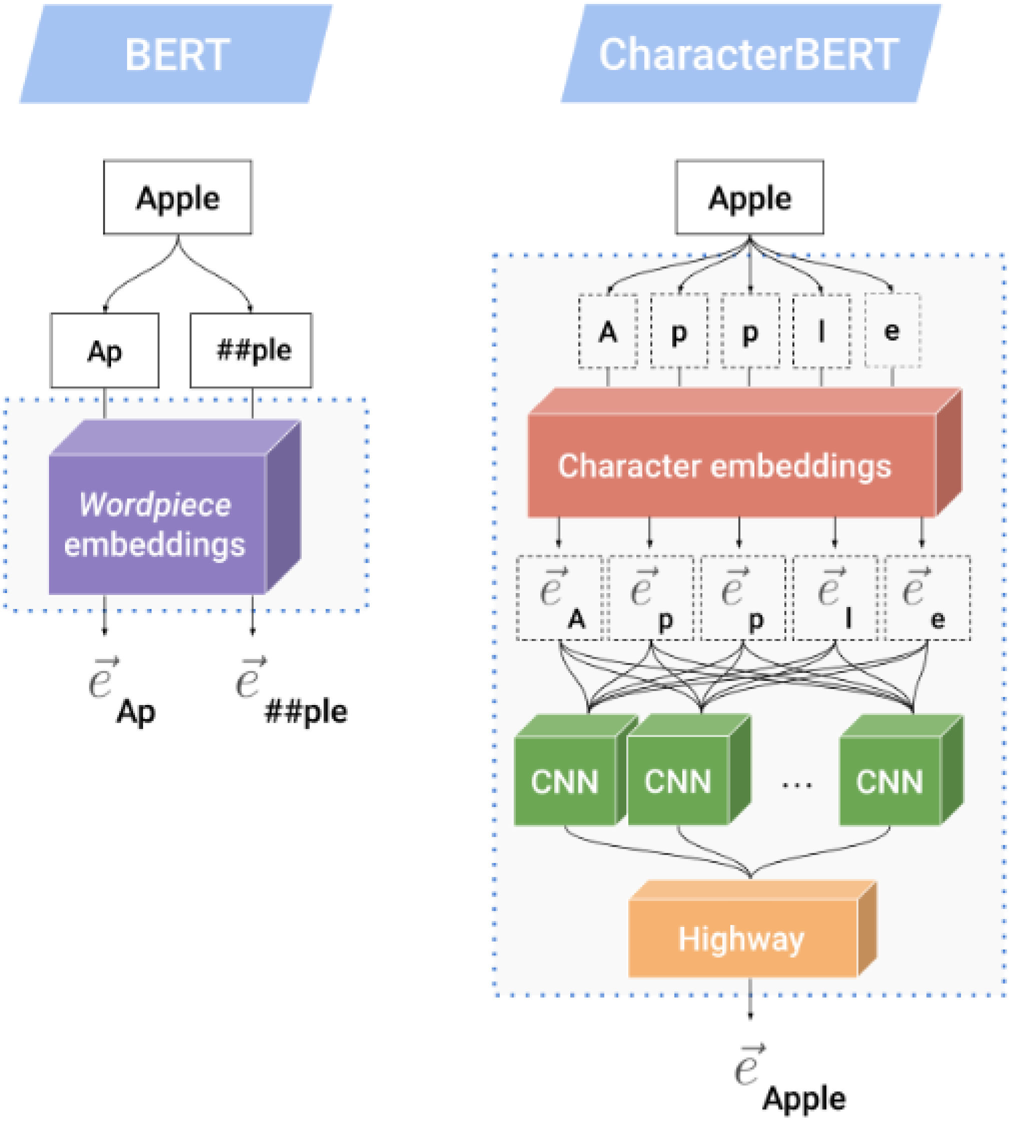
- a initial domain vocabulary always increase the perf
- combination of corpus better than general but not better than medical domain
- the second training step led to sensitively the same results
- better than BlueBERT if domain vocab and same training pipeline

See it here : https://hal.archives-ouvertes.fr/hal-02786184/document

3. How to improve the wordpiecing ?

When BERT met unknown word it tokenize it into vector of sub words. In the medical domain we can tokenize it into more meaningful units

The idea : CharacterBERT a CNN module producing single representation for group of subwords



Tasks : Medical entity recognition / natural language inference / relation classification / sentence similarity

Evaluation :

- use of Statistical significance : *Almost Stochastic Order tests (ASO)*
- robustness to noise : removing/adding/replacing of characters depending to a level of error

Results : CharacterBERT better than BERT, more robust to misspellings and having a better open-vocab representation. This is important because BERT is sensitive to misspelling.

But : 26 hours of training for BERT : 55 for CharacterBERT

Perspectives : comparing it with recent *tokenization-free models* | exploring the cross lingual potential of CharacterBERT

See it here : https://arxiv.org/pdf/2010.10392.pdf

4. How to integrate external knowledge base to contextualized embeddings ?

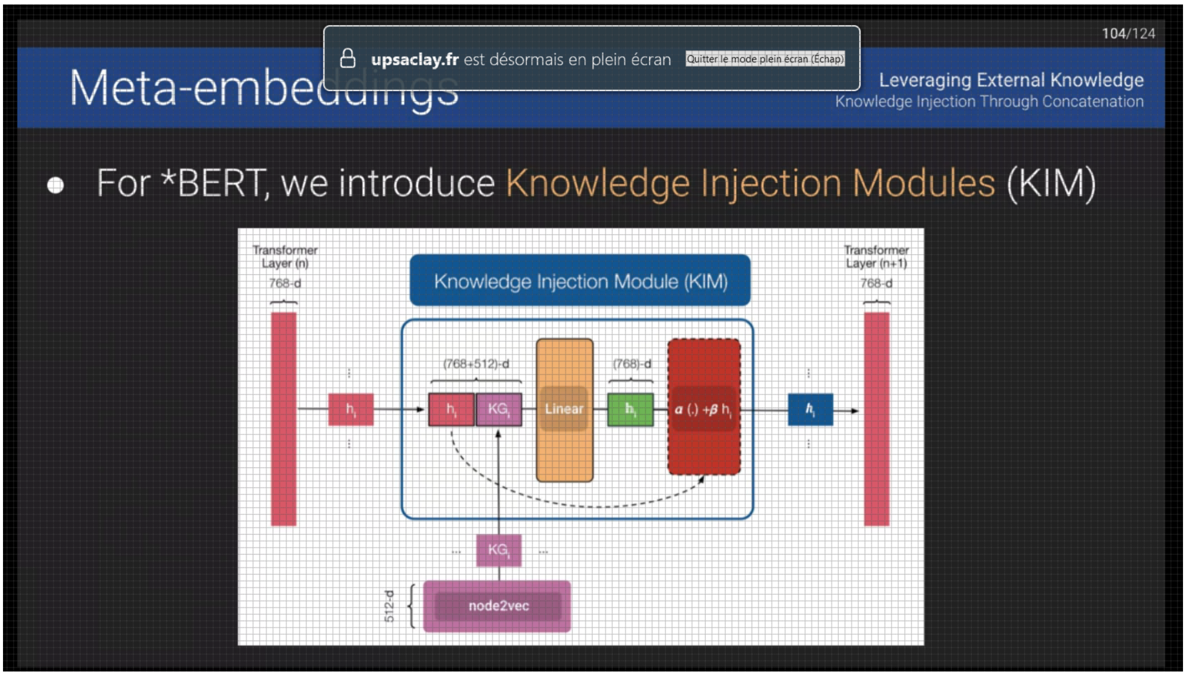
> *not published as an article for the moment*

Knowledge bases focused on medical concepts representation : UMLS Metathesaurus *graph*, the SNOWMED and the MESH thesaurus

Use of Node2Vec

How to map text into a graph node : string matching

Approach : meta-embeddings : concatenation of embeddings via Knowledge injection modules (KIM)



20 models combinations for the experiments

Results : improve performance but need to be refined

Perspectives of refining : terminologies, entity linking, use of more advanced knowledge embeddings

General perspectives

- Can be applied to other specific domain
- Need a real methodology for measuring domain similarity of a corpus

Going further ...

All models are available here : https://github.com/helboukkouri