92586 Computational Linguistics

Lesson 20. Beyond

Alberto Barrón-Cedeño

Alma Mater Studiorum-Università di Bologna
a.barron@unibo.it @_albarron_

11/05/2022

Ad: TinfFoil Seminar; Friday 13th, 10 to 12, dit.lab, I16

Lavinia Aparaschivei · 2nd-year PhD student
Are Crescia and Piadina the same? Towards Identifying Synonymy or non-Synonymy between Italian words to Enable Crowdsourcing from Language Learners
CLIC-it 2022, Milan, Italy

Katerina Korre · 1st-year PhD student
Enriching Grammatical Error Correction Resources for Modern Greek
LREC 2022, Marseille, France

Arianna Muti · 1st-year PhD student
A checkpoint on multilingual misogyny identification
ACL Student Workshop 2022, Dublin, Ireland
LeaningTower@LT-EDI 2022: When hate and hope collide
LT-EDI@ACL 2022, Dublin, Ireland

Francisco Jañez · Erasmus+ visiting PhD student, Universidad de León (Spain)
On spotting propaganda in spam email

Alberto Barrón
Informal talk: attending conferences (time allowing)

Better text generation

Adding Extra Stuff

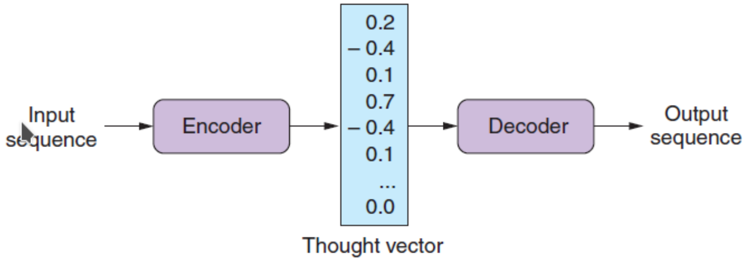
- Expand the quantity and quality of the corpus
- Expand the complexity of the model (units/layers/LSTMs)
- Better pre-processing:
  - Better case folding
  - Break into sentences
- Post-processing
  - Add filters on grammar, spelling, and tone
  - Generate many more examples than actually shown to users
- Select better seeds (e.g., context, topic)

Most of these strategies apply to any problem you can think about!

(Lane et al., 2019, 307)

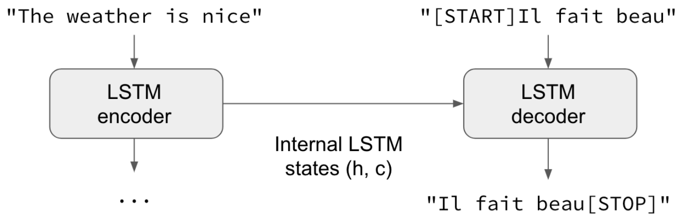
Sequence-to-Sequence Models

Encoder-Decoder architecture



(Lane et al., 2019, 315)

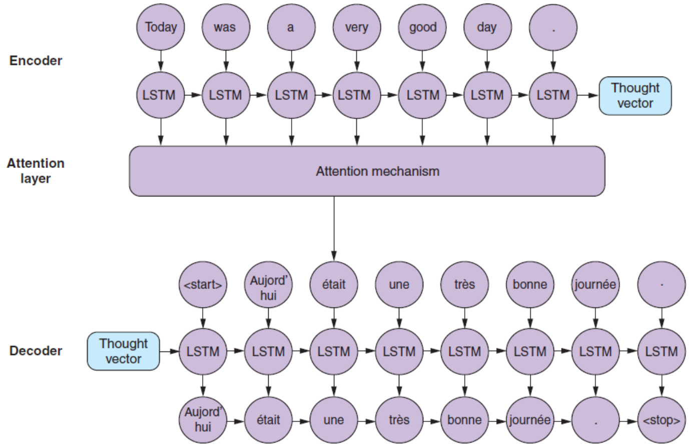
seq2seq models



https://blog.keras.io/a-ten-minute-introduction-to-sequence-to-sequence-learning-in-keras.html

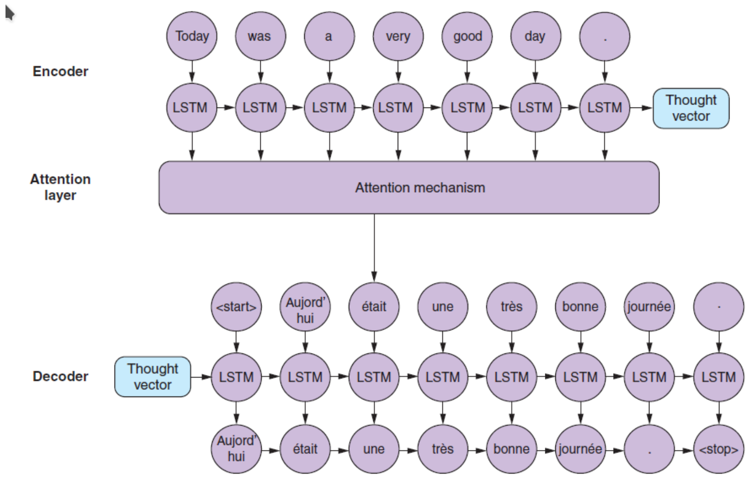
Sequence Labelling

- Part-of-speech tagging
- Dependency parsing
- Named entity recognition



(Lane et al., 2019, 334)

Attention is all you need



(Vaswani et al., 2017); Figure from (Lane et al., 2019, 334)

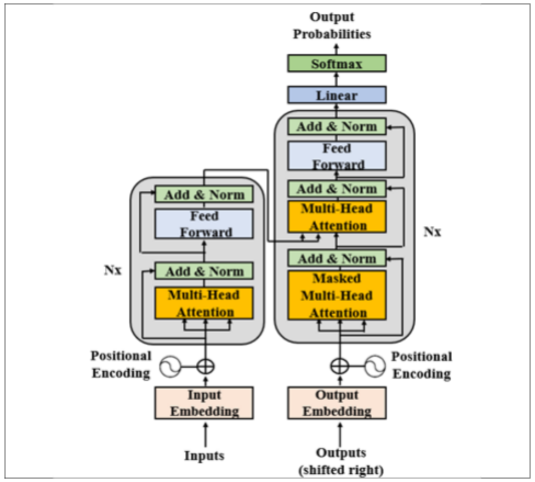
Transformers

A Transformer [. . .] helps in transforming one sequence of input into another depending on the problem statement. Examples:

- Translation from one language to another
- Paraphrasing
- Question answering

https://medium.com/data-science-in-your-pocket/attention-is-all-you-need-understanding-with-example-c8d074c37767

Transformers

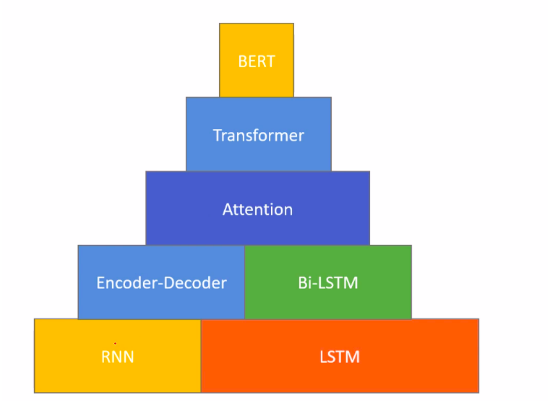


Vaswani et al. (2017)

Transformers: BERT

Bi-directional **e**ncoder **r**epresentations from **t**ransformers

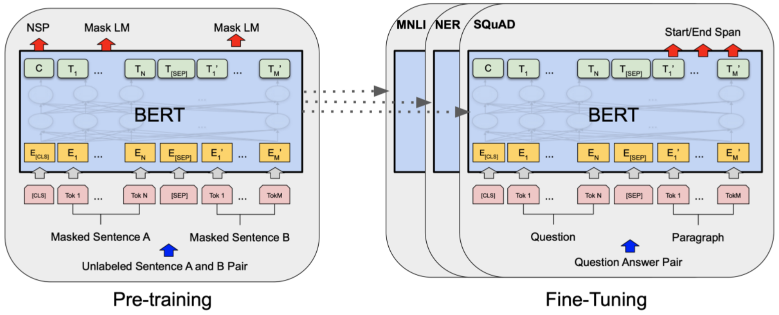
Learning pyramid:



Picture from https://iq.opengenus.org/introduction-to-bert/

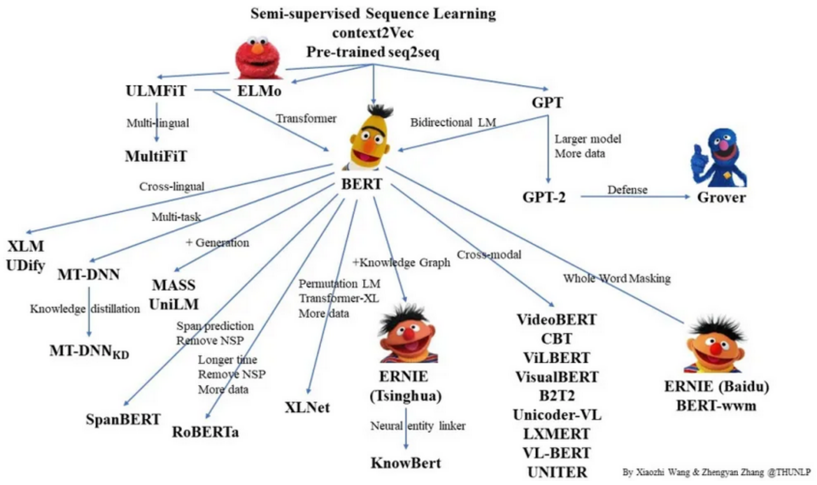
Transformers: BERT

Bi-directional **e**ncoder **r**epresentations from **t**ransformers



(Devlin et al., 2019)
Picture from https://towardsdatascience.com/bert-why-its-been-revolutionizing-nlp-5d1bcae76a13

BERT Family



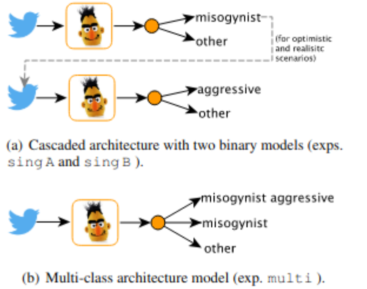
Picture from https://neptune.ai/blog/bert-and-the-transformer-architecture-reshaping-the-ai-landscape
(actually from https://github.com/thunlp/)

BERT in other Languages

For instance:

- Spanish (Cañete et al., 2020)
- Italian (AlBERTo) (Polignano et al., 2019)

Use example: misogyny identification in Italian



(a) Cascaded architecture with two binary models (exps. singA and singB).

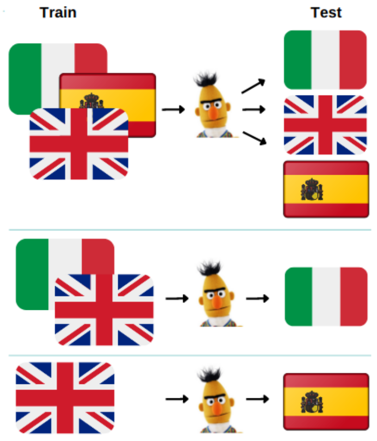
(b) Multi-class architecture model (exp. multi).

Figure 1: The two alternative system architectures for misogyny and aggressiveness identification.

(?)

Multilingual models

What makes multilingual BERT multilingual? (Liu et al., 2020)
Use example: multilingual misogyny identification



(Muti and Barrón-Cedeño, 2022 to appear)

(Other) Reference Libraries

- **Spacy**
  Industrial-Strength Natural Language Processing
  `https://spacy.io/`
- **Stanza**
  A Python NLP Package for Many Human Languages
  `https://stanfordnlp.github.io/stanza/`
- **Hugging Face**
  The AI community building the future
  `https://huggingface.co/`

Conferences (non-exhaustive)

NLP-ish	IR-ish	MT-ish
Top		
ACL	SIGIR	WMT
EMNLP	CIKM	EAMT
NAACL	WSDOM	
EACL	ECIR	
Nice		
SemEval	CLEF	
CICLing	TREC	
LREC		
National		
CLIC-it	IIR	
Evalita		

Recap

Recap: The path

1. Baby steps into computing
2. What is NLP? From rule-based to statistical
3. Pre-processing text: tokens, stemming, stopwording. . .
4. From words to vectors: the vector space model
5. A few supervised models
6. Training and evaluating in machine learning
7. From words to meaning: topic modeling
8. Using one neuron: perceptrons
9. Fully-connected neural networks
10. From words to semantics: word embeddings
11. Taking snapshots of text: CNNs
12. Texts as sequences: (Bi)RNNs
13. Using a better memory: LSTM
14. LSTM to produce text

Recap: The future path

- We covered Parts 1 and 2 of Lane et al. (2019) (up to Section 9)
- That's 9 out of 13 chapters of Natural Language Processing in Action

You are ready to go on your own now and become a star

Now go and celebrate the end of the course

. . . and worry about your project from Monday!

- I'm available until mid-July for 1-to-1 discussion on your project **upon request!**

References

Cañete, J., G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, and J. Pérez
2020. Spanish pre-trained bert model and evaluation data. In *PML4DC at ICLR 2020*.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova
2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Lane, H., C. Howard, and H. Hapkem
2019. *Natural Language Processing in Action*. Shelter Island, NY: Manning Publication Co.

Liu, C.-L., T.-Y. Hsu, Y.-S. Chuang, and H. yi Lee
2020. What makes multilingual bert multilingual? *arXiv*.

Muti, A. and A. Barrón-Cedeño
2020. UniBO@AMI: A Multi-Class Approach to Misogyny and Aggressiveness Identification on Twitter Posts Using AlBERTo. In *Proceedings of the 7th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2020)*.

Polignano, M., P. Basile, M. de Gemmis, G. Semeraro, and