POLBERT FOR WSD

Darek Kłeczek

ABOUT ME

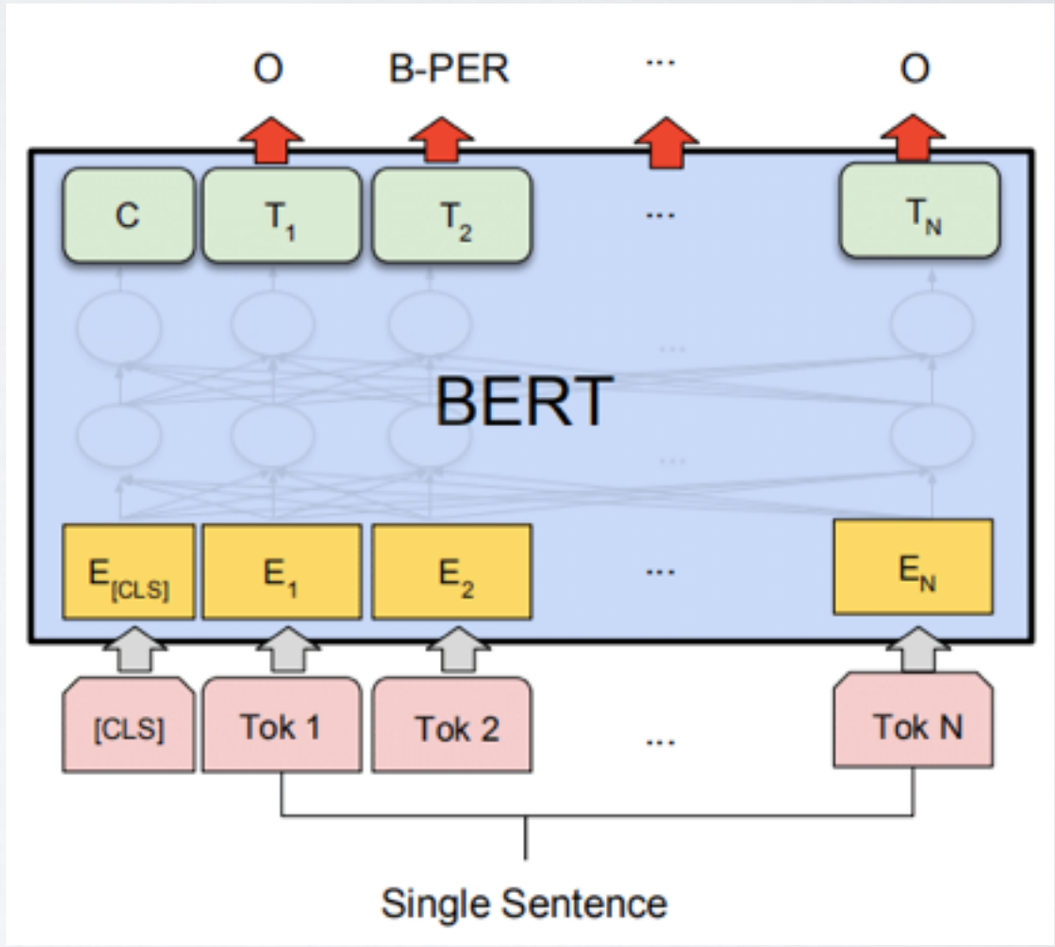
- Intelligent Automation at P&G
- Polish NLP
  - Polbert
  - Meetup Group
- Personal website: skok.ai
- Twitter: @dk21

FRAMING THE WSD TASK

- Token classification given the sentence context
- Number of classes = number of WordNet senses
- Different subset of classes relevant for each token (all that are associated with the token's lemma)

TOKEN CLASSIFICATION

- Token embeddings capture meaning in context (via attention mechanism)
- Number of classes is huge!
- Need labeled data for supervision!
- Idea for later: multilingual transformer trained on SemCore



GLOSSBERT

Use sentence pair classification to determine the correct word sense

- Token-CLS
- Sent-CLS
- Sent-CLS-WS (Weak Supervision)

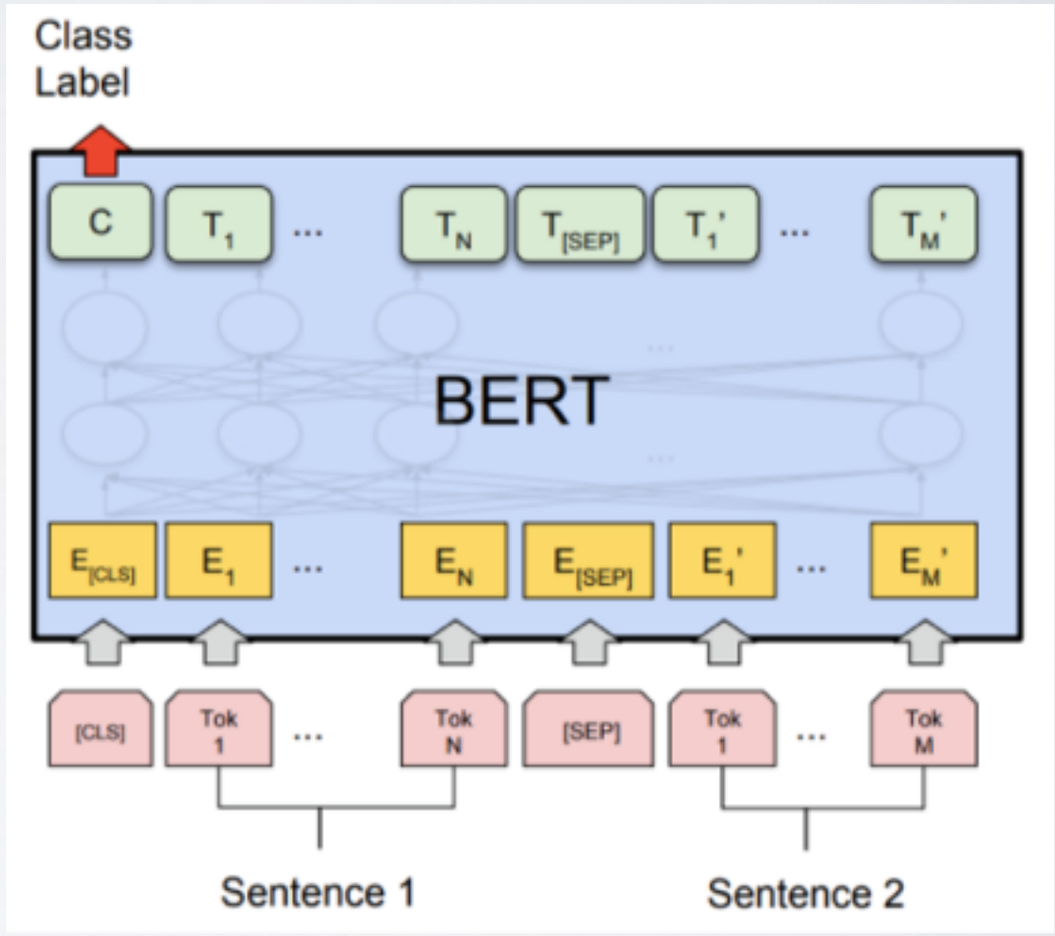
Sentence with four targets:
Your research stopped when a convenient assertion could be made.

Context-Gloss Pairs of the target word [research]	Label	Sense Key
[CLS] Your research ... [SEP] systematic investigation to ... [SEP]	Yes	research%1:04:00::
[CLS] Your research ... [SEP] a search for knowledge [SEP]	No	research%1:09:00::
[CLS] Your research ... [SEP] inquire into [SEP]	No	research%2:31:00::
[CLS] Your research ... [SEP] attempt to find out in a ... [SEP]	No	research%2:32:00::

Context-Gloss Pairs with weak supervision of the target word [research]	Label	Sense Key
[CLS] Your "research" ... [SEP] research: systematic investigation to ... [SEP]	Yes	research%1:04:00::
[CLS] Your "research" ... [SEP] research: a search for knowledge [SEP]	No	research%1:09:00::
[CLS] Your "research" ... [SEP] research: inquire into [SEP]	No	research%2:31:00::
[CLS] Your "research" ... [SEP] research: attempt to find out in a ... [SEP]	No	research%2:32:00::

SENTENCE PAIR CLASSIFICATION

- [CLS] token embedding represents relationship between sentences
- Each token learns representation in the context of both sentences



POLBERT FOR WSD

Same approach as GlossBERT Sent-CLS-WS (Weak Supervision)

- Using both glosses and examples from WordNet
- Mirror *weak supervision* for sentence 1 and 2
- Pretrained encoder: Polbert uncased, fine-tuned on plWordNet examples only (443316 pairs)

Example	Label
[CLS] pokój: Wszystkie trzy **pokoje** mają okna od strony parku [SEP] pokój: pomieszczenie mieszkalne, w którym się przebywa, także w hotelu	TRUE
[CLS] pokój: Wszystkie trzy **pokoje** mają okna od strony parku [SEP] pokój: Dwa **pokoje** z kuchnią w zupełności mi wystarczą, nie potrzebuję większego mieszkania.	TRUE
[CLS] pokój: Wszystkie trzy **pokoje** mają okna od strony parku [SEP] pokój: stan, gdy nie ma wojny	FALSE
[CLS] pokój: Wszystkie trzy **pokoje** mają okna od strony parku [SEP] pokój: Za szczególny wkład w promowanie **pokoju** i praw człowieka na świecie przyznawana jest Pokojowa Nagroda Nobla.	FALSE

DISAMBIGUATION BASELINE

- Split text into sentences
- Identify lemma for each word in a sentence
- List all synsets associated with the lemma
- For each synset, list all glosses and examples from WordNet
- Create sentence pairs with the original lemma and sentence as sentence 1, and the listed glosses and examples as sentence 2
- Run the model and average results across all examples per synset
- Highest score is selected as the disambiguated sense for the target word

BASELINE + MWE + RELS (V2)

- Multi-Word Expressions
  - for each bigram/trigram in a sentence, check if there is associated WordNet lemma
  - if yes, use that lemma for disambiguation
- Relationships
  - some synsets associated with a lemma do not have glosses or examples
  - for those synsets, look up synsets connected via hypernymy/hiponymy relationships and associated examples
  - use these examples to score the synsets that didn't have examples

RESULTS AND FUTURE WORK

Submission	Affiliation	Precision KPWr	Recall KPWr	Precision Sherlock	Recall Sherlock
Polbert for WSD (v2)	**skok.ai**	**0.599296**	**0.588727**	**0.592263**	**0.576850**
Polbert for WSD	skok.ai	0.564432	0.550860	0.564384	0.542966
PolevalWSDv1		0.318547	0.231085	0.291732	0.200867

- More experiments!
- Improve sentence-pair classification model (ablation study - encoder, finetuning protocol etc.)
- Improve disambiguation algorithm (e.g. max score vs average score)
- Use annotated data (e.g. SemCore with multilingual models)