The Alan Turing Institute

WARWICK
THE UNIVERSITY OF WARWICK

Queen Mary
University of London

Imperial College London

GiBERT - Introducing Linguistic Information into BERT through a Lightweight Gated Injection Method

Nicole Peinelt, Marek Rei, Maria Liakata

EMNLP Findings, 7-11 November 2021

Motivation

Pretrained Transformers

BERT (Devlin et al, 2019)
RoBERTa (Liu et al., 2019)
GPT2 (Radford et a., 2019)
ALBERT (Lan et al., 2020)

Motivation

Pretrained embeddings

Collocations (Mikolov et al., 2013, Pennington et al., 2014)
Dependencies (Levy and Goldberg, 2014)
Subword information (Bojanowski et al., 2017)
Semantic lexicons (Faruqui et al., 2015)

Pretrained Transformers

BERT (Devlin et al, 2019)
RoBERTa (Liu et al., 2019)
GPT2 (Radford et a., 2019)
ALBERT (Lan et al., 2020)

Motivation

Pretrained embeddings

Collocations (Mikolov et al., 2013, Pennington et al., 2014)
Dependencies (Levy and Goldberg, 2014)
Subword information (Bojanowski et al., 2017)
Semantic lexicons (Faruqui et al., 2015)

Many resources & useful for
Semantic Similarity Detection

Pretrained Transformers

BERT (Devlin et al, 2019)
RoBERTa (Liu et al., 2019)
GPT2 (Radford et a., 2019)
ALBERT (Lan et al., 2020)

Motivation

Pretrained embeddings

Collocations (Mikolov et al., 2013, Pennington et al., 2014)
Dependencies (Levy and Goldberg, 2014)
Subword information (Bojanowski et al., 2017)
Semantic lexicons (Faruqui et al., 2015)

Many resources & useful for Semantic Similarity Detection

Pretrained Transformers

BERT (Devlin et al, 2019)
RoBERTa (Liu et al., 2019)
GPT2 (Radford et a., 2019)
ALBERT (Lan et al., 2020)

Enriched Transformers

Knowledge bases (Peters et al., 2019)
Multi-modal information (Lu et al., 2019)
Topic models (Peinelt et al., 2020)

Successful enrichment with external information

Motivation

Pretrained embeddings

Collocations (Mikolov et al., 2013, Pennington et al., 2014)
Dependencies (Levy and Goldberg, 2014)
Subword information (Bojanowski et al., 2017)
Semantic lexicons (Faruqui et al., 2015)

Pretrained Transformers

BERT (Devlin et al, 2019)
RoBERTa (Liu et al., 2019)
GPT2 (Radford et a., 2019)
ALBERT (Lan et al., 2020)

Enriched Transformers

Knowledge bases (Peters et al., 2019)
Multi-modal information (Lu et al., 2019)
Topic models (Peinelt et al., 2020)

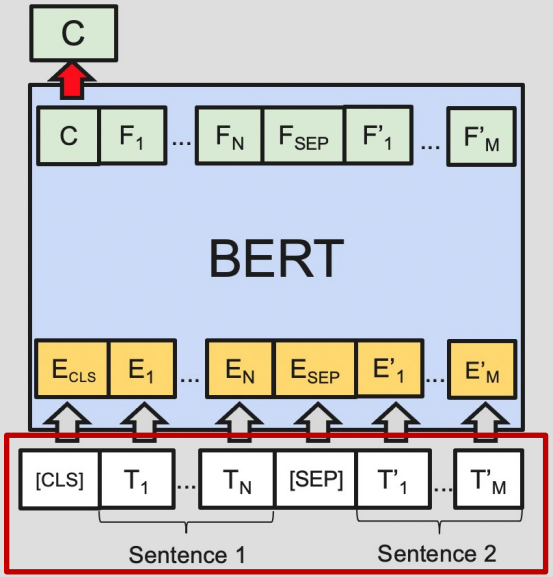
Many resources & useful for Semantic Similarity Detection

Successful enrichment with external information

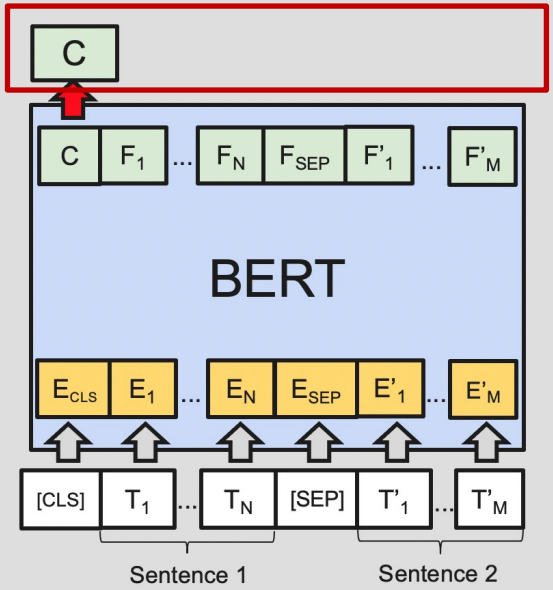
→Idea: Combining linguistically enriched embeddings with BERT

Combining external information with BERT

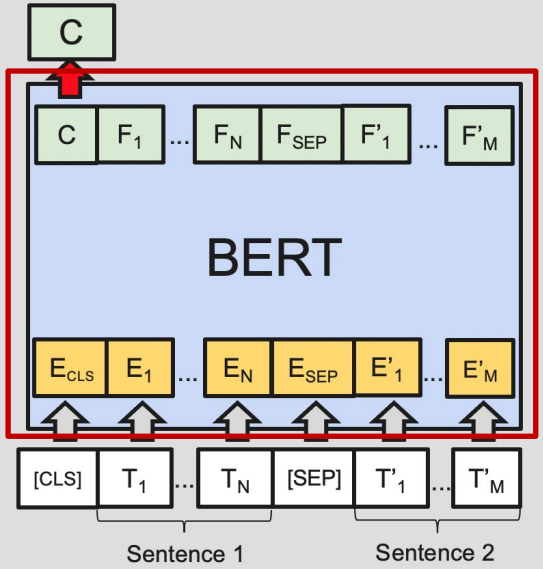
Input modifications



Output modifications

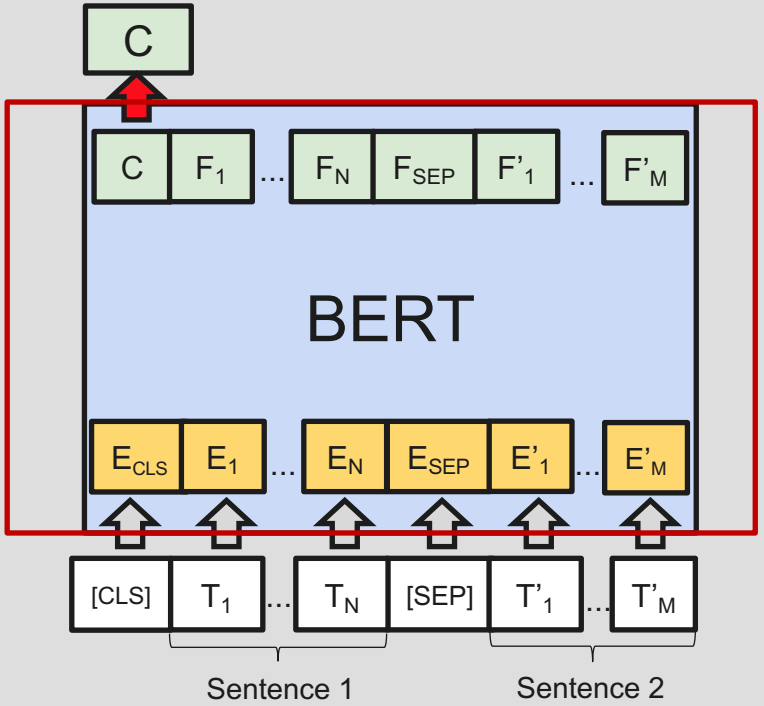


Internal modifications



Combining external information with BERT

– Internal modifications:
  - Changing BERT's internal architecture
  - Examples:
    - VilBERT (Lu et al., 2019)
    - KnowBERT (Peters et al., 2019)
    - InterBERT (Lin et al., 2020)



Injection methods

Attention injection

$$\mathbf{H^{i'}} = \mathbf{H^i} + \text{MultiHeadAtt}(\mathbf{H^i}, \mathbf{I}, \mathbf{I})$$

$$\text{MultiheadAtt}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = [\text{head}_1; \ldots; \text{head}_h]\mathbf{W^O}$$

$$\text{head}_j = \text{Attention}(\mathbf{QW_j^Q}, \mathbf{KW_j^K}, \mathbf{VW_j^V})$$

Gated injection

$$\mathbf{P} = \text{FeedForward}(\mathbf{I})$$

$$\mathbf{H^{i'}} = \mathbf{H^i} + \mathbf{g} \odot \mathbf{P}$$

where

$\mathbf{H_i}$ = BERT's hidden representation after layer i

$\mathbf{I}$ = aligned injection sequence

$\mathbf{H_i'}$ = BERT's updated representation after layer i

Experimental Setup

- Embeddings:
  - Dependency-based (Levy and Goldberg 2014)
  - Counter-fitted (Mrkšic´ et al. 2016)
- Baselines:
  - KeLP (Filice et al., 2017)
  - ECNU (Wu et al., 2017)
  - Bunji (Koreeda et al., 2017)
  - **BERT** (Devlin et al. 2019)
  - SemBERT (Zhang et al. 2020)
  - **AiBERT (**attention injection)

1. Can the injection of linguistically enriched embeddings improve BERT's performance?

	F1					
	MSRP	Quora	SemEval			avg
			A	B	C	
Previous systems						
KeLP◇	-	-	-	.506	-	-
ECNU◇	-	-	.777	-	-	-
Bunji◇	-	-	-	-	.197	-
BERT⋆	.876	.902	.704	.473	.268	.645
SemBERT⋆	.876	.901	✗	✗	✗	-
Our implementation						
AiBERT$_{\text{dependency}}$	.863	.903	.738	.498	**.282**	.657
AiBERT$_{\text{counter-fitted}}$	.877	.904	.724	.496	.263	.653
GiBERT$_{\text{dependency}}$	.883	.904	.768	.474	.238	.653
GiBERT$_{\text{counter-fitted}}$	**.884**	**.907**	**.780**	**.511**	.256	**.668**

AiBERT and GiBERT both improve over BERT.

2. Which injection method works best?

	F1					
	MSRP	Quora	SemEval			avg
			A	B	C	
Previous systems						
KeLP◇	-	-	-	.506	-	-
ECNU◇	-	-	.777	-	-	-
Bunji◇	-	-	-	-	.197	-
BERT⋆	.876	.902	.704	.473	.268	.645
SemBERT⋆	.876	.901	✗	✗	✗	-
Our implementation						
AiBERT$_{\text{dependency}}$	.863	.903	.738	.498	**.282**	.657
AiBERT$_{\text{counter-fitted}}$	.877	.904	.724	.496	.263	.653
GiBERT$_{\text{dependency}}$	.883	.904	.768	.474	.238	.653
GiBERT$_{\text{counter-fitted}}$	**.884**	**.907**	**.780**	**.511**	.256	**.668**

Gated injection at least as good as attention injection while using fewer additional parameters (0.23M vs. 1.64M)

Conclusion

1. Linguistically enriched embeddings improve BERT’s performance

2. Gated injection method at least as effective as attention injection with fewer parameters

Thank you