Natural Language Processing homeworks

Andrea Gasparini – 1813486
Natural Language Processing – A.Y. 2020-2021
Sapienza University of Rome

Word-in-Context (WiC) disambiguation

Homework 1

WiC disambiguation is a binary classification task:
given a pair of sentences with the same **target word**, predict whether it has the same meaning in both or not.

- Use the **mouse** to click on the button

$\neq$

- The cat eats the **mouse**

- The cat eats the **mouse**

$=$

- The **mouse** escaped from the predator

Word-in-Context (WiC) disambiguation

Preprocessing

- *GloVe* [1] pre-trained vectors as static (context-free) embeddings to encode input tokens
- Replacement of separation characters (e.g. hyphens and underscores) with actual whitespaces
- Removal of special characters and punctuation
- OOV (out-of-vocabulary) words handled with a randomly initialized embedding vector

Word-in-Context (WiC) disambiguation

Approaches

Word-level model

- 2-layer MLP with ReLU activation
- Considers which words are contained in the sentences
- Does not keep track of the context

Sequence encoding model

- Exploits sequence-level semantics through Recurrent Neural Networks
- Built on top of the word-level model to classify the recurrent output

Word-in-Context (WiC) disambiguation

Word-level model: experiments

- Concatenation of the two sentences’ encodings, separated by a special token
- Weighted (on the target words) average of the encodings to reduce dimensionality
- Concatenation of the two encodings’ individual averages
- Stop words removal

Word-in-Context (WiC) disambiguation

Sequence encoding model: experiments

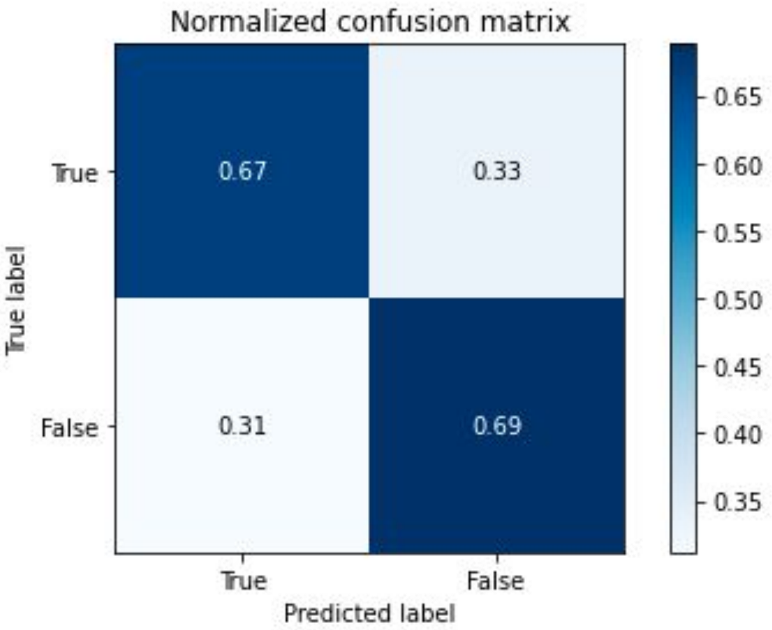
- LSTM with the concatenation of the sentences' encodings as input
- (Baseline LSTM) Individually feeding the LSTM with the two encodings and then concatenate the last hidden states
- Dropout regularization to prevent overfitting
- Bidirectional LSTM to improve the contextual representation and then make use of the target words' hidden states

Word-in-Context (WiC) disambiguation

Word-level model: results

Hyperparameter	Tested values
Adam learning rate	{0.01, 0.005, **0.001**, 0.0001}
SGD learning rate	[0.01 – 0.5]
SGD momentum	[0.0 – 0.5]
Hidden size	{100, 200, **n_features // 2**}

Model	Accuracy	F1-score
Embeddings average	0.6200	0.6175
+ stop words removal	0.6450	0.6447
Individual avg. concat.	0.6580	0.6537
+ stop words removal	**0.6770**	**0.6770**

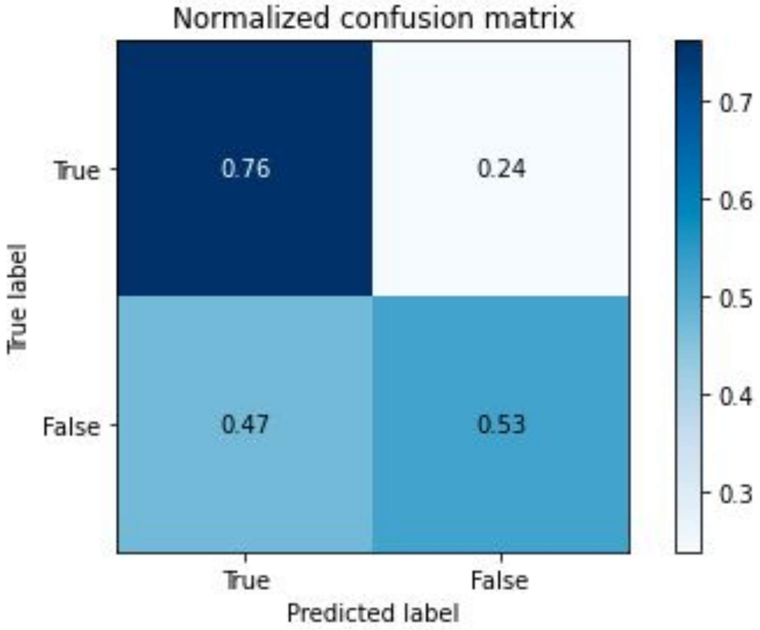


Word-in-Context (WiC) disambiguation

Sequence encoding model: results

Hyperparameter	Tested values
Adam learning rate	{0.01, 0.005, 0.001, **0.0001**}
SGD learning rate	[0.01 – 0.5]
SGD momentum	[0.0 – 0.5]
Embedding dropout	[0.0 – **0.5**]
Fully conn. dropout	[0.0 – **0.5**]
LSTM dropout	[0.0 – **0.5**]
LSTM hidden size	{**100**, 200, 250, 300}

Model	Accuracy	F1-score
Baseline LSTM	0.5810	0.5801
+ dropout	0.6280	0.6276
++ stop words removal	**0.6440**	**0.6391**
Bidirectional LSTM	0.6040	0.5984
+ dropout	0.6140	0.6084



Word-in-Context (WiC) disambiguation

Conclusions

- Simpler word-level approach overcomes the sequence encoding one
- Harder predictions of “False” samples for the sequence encoding approach
- Difficulty in exploiting sequence-level semantics by only employing LSTMs
- Contextualized word embeddings and Transformer-based models may improve the results [2]

Aspect-Based Sentiment Analysis (ABSA)

Homework 2

ABSA is a pipeline composed of 4 sub-tasks:

(A) Aspect term identification
(B) Aspect term polarity classification
(C) Aspect category identification
(D) Aspect category polarity classification

e.g. “The sangria was pretty tasty and good”

(A) { **sangria** }
(B) { (sangria, **positive**) }
(C) { **food** }
(D) { (food, **positive**) }

Address ABSA by jointly solving A+B and C+D with different models

Aspect-Based Sentiment Analysis (ABSA)

Task A+B

- Sequence labelling task (similar to NER)
- Tagging with IOB scheme (Inside, Outside, Beginning) + polarity (*positive*, *negative*, *neutral*, *conflict*)

Tokenized sentence	The	hard	drive	crashed	as	well	so	I	bought	a	new	power	cord
Polarity		negative										negative	
IOB + Polarity	O	B-negative	I-negative	O	O	O	O	O	O	O	O	B-negative	I-negative

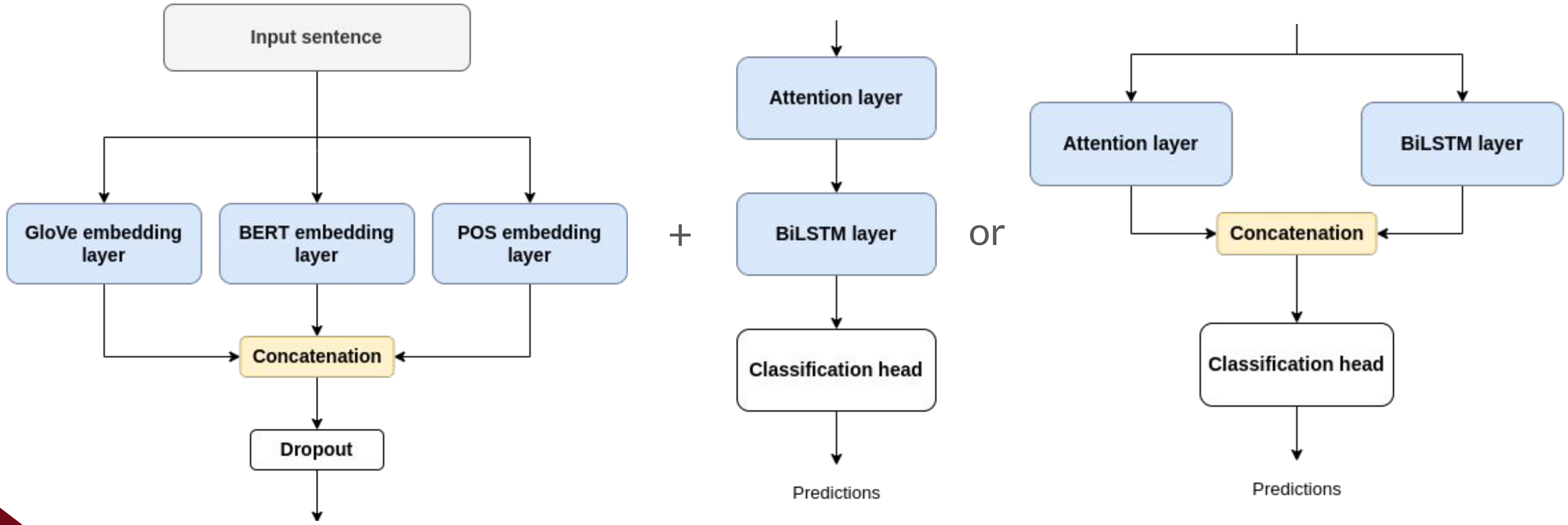
Aspect-Based Sentiment Analysis (ABSA)

Task C+D

- Multi-label classification task
- 5 possible categories (*anecdotes/miscellaneous*, *price*, *food*, *ambience*, *service*)
- 4 possible polarities (*positive*, *negative*, *neutral*, *conflict*) for each category

Aspect-Based Sentiment Analysis (ABSA)

Modular architecture



Aspect-Based Sentiment Analysis (ABSA)

BERT pooling strategies

- Experimentation on BERT [3] layer pooling strategies
- WordPiece pooling by averaging

	Task A+B		Task C+D	
Pooling strategy	F1 Ident.	F1 Class.	F1 Ident.	F1 Class.
Last	80,75	45,54	82,60	49,87
Second-to-Last	**81,70**	**47,41**	83,99	52,20
Concat Last Four	80,73	44,37	82,82	49,56
Sum Last Four	79,77	44,19	82,09	48,04
Average Last Four	81,67	46,44	**84,36**	**52,61**

Aspect-Based Sentiment Analysis (ABSA)

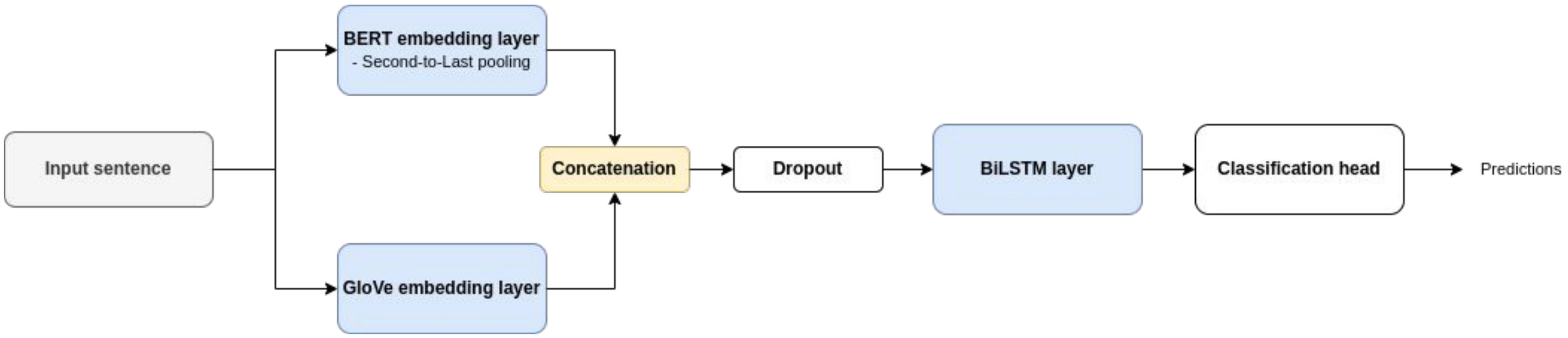
Results

- POS does not bring improvements
- Contextualized embeddings (BERT) and Bidirectional LSTMs are key aspects
- Attention brings a small improvement only for task C+D

Model architecture	Task A+B F1 Ident.	Task A+B F1 Class.	Task C+D F1 Ident.	Task C+D F1 Class.
LSTM + GloVe	73,53	32,15	76,47	37,26
+ POS	68,55	30,20	70,32	35,71
+ BERT_{frozen}	79,96	40,53	82,09	44,95
+ BERT_{frozen} + POS	74,45	35,23	77,37	41,02
BiLSTM + GloVe	75,26	38,82	78,04	44,12
+ POS	75,74	39,23	78,51	44,78
+ BERT_{frozen}	**81,70**	**47,41**	**84,36**	52,61
+ BERT_{frozen} + POS	80,85	45,86	83,57	50,42
+ $\text{BERT}_{finetuned}$	79,78	40,75	82,86	47,58
BiLSTM + GloVe + Attention	74,17	36,40	78,37	39,59
+ Concat outputs	76,63	41,11	-	-
+ BERT_{frozen}	71,02	32,94	83,06	**53,59**
+ BERT_{frozen} + Concat outputs	80,56	45,89	-	-
BiLSTM + GloVe + Transformer encoder	76,54	38,90	79,59	37,95
+ Concat outputs	77,41	41,66	-	-
+ BERT_{frozen}	65,27	30,27	81,33	48,94
+ BERT_{frozen} + Concat outputs	80,97	46,00	-	-

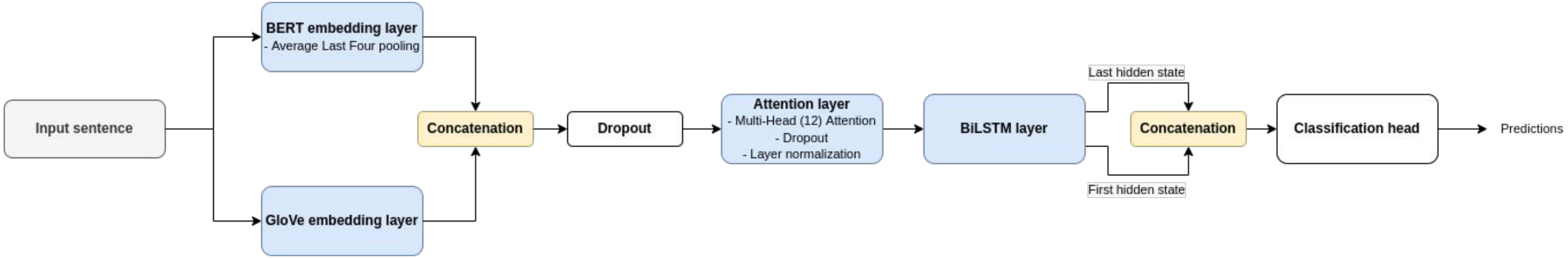
Aspect-Based Sentiment Analysis (ABSA)

Task A+B architecture



Aspect-Based Sentiment Analysis (ABSA)

Task C+D architecture



Aspect-Based Sentiment Analysis (ABSA)

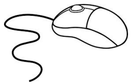
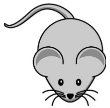
Conclusions

- Extensive experimentation with different models
- Already satisfactory results by combining static and contextual embeddings on top of a BiLSTM
- POS “manual” tagging may not be 100% accurate
- BERT fine-tuning and Attention mechanisms [4] improvements may be further investigated

Word Sense Disambiguation (WSD) of WiC data

Homework 3

WSD aim is to identify the meaning of ambiguous words by assigning sense identifiers from a pre-defined inventory, e.g. WordNet [5]

- Use the **mouse** to click on the button → **mouse%1:06:00::** 
- The cat eats the **mouse** → **mouse%1:05:00::** 

A prediction for WiC disambiguation is obtained “for free” by comparing the two sense ids.

Word Sense Disambiguation (WSD) of WiC data

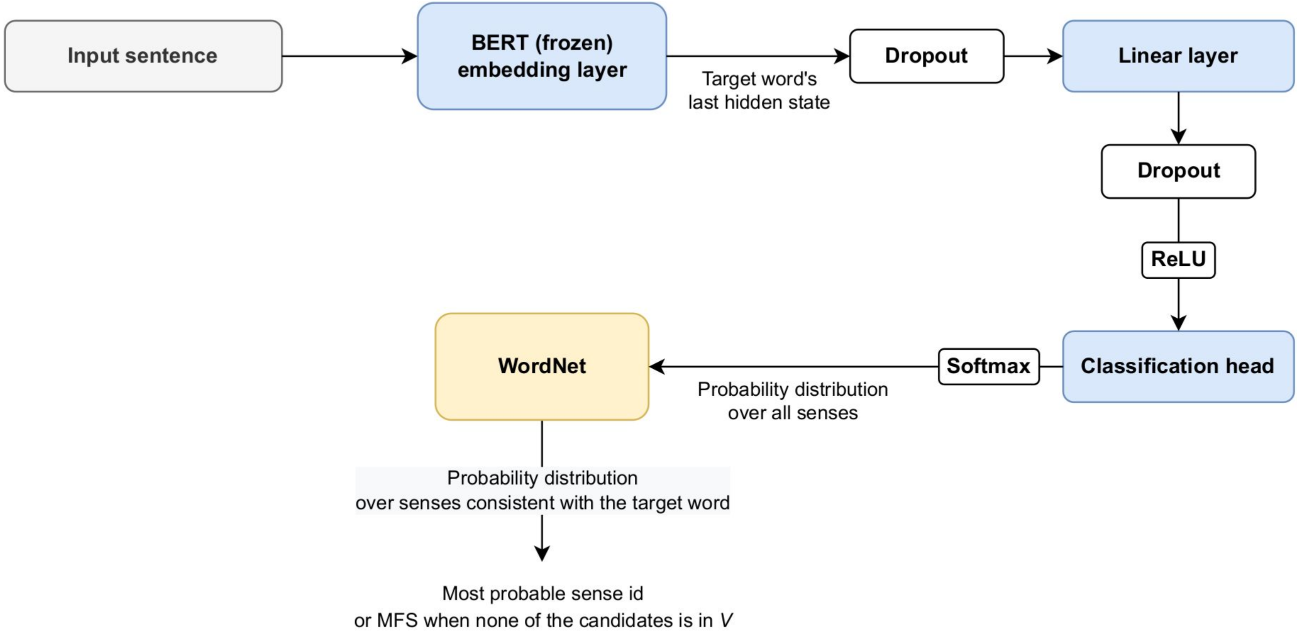
Datasets

WSD Unified Evaluation Framework [6]

- Training: SemCor (sampling a smaller fraction due to limited resources)
- Validation: SemEval-2007 (for hyperparameters tuning and early stopping)
- Testing: WiC data (for testing both WSD and WiC performance)

Word Sense Disambiguation (WSD) of WiC data

BERT (frozen) + WordNet



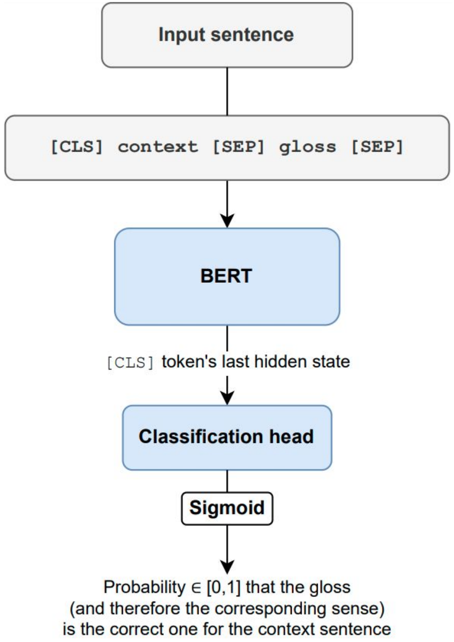
Word Sense Disambiguation (WSD) of WiC data

BERT fine-tuned with context-gloss pairs

Sentence with 2 target words to be disambiguated:
We have <u>made</u> no such <u>statement</u>.

Context-gloss pairs of the target word "statement" (with weak supervision on the gloss)	Label	Sense id
[CLS] We have made no such statement [SEP] statement: a message that is stated or ... [SEP]	No	statement%1:10:00::
[CLS] We have made no such statement [SEP] statement: a fact or assertion offered as ... [SEP]	No	statement%1:10:02::
[CLS] We have made no such statement [SEP] statement: (music) the presentation of a ... [SEP]	No	statement%1:10:04::
[CLS] We have made no such statement [SEP] statement: the act of affirming or asserting or ... [SEP]	Yes	statement%1:10:06::
...		

Context-gloss pairs of the target word "made" (with weak supervision on the gloss)	Label	Sense id
[CLS] We have made no such statement [SEP] make: engage in [SEP]	No	make%2:41:00::
[CLS] We have made no such statement [SEP] make: give certain properties to something [SEP]	No	make%2:30:00::
[CLS] We have made no such statement [SEP] make: create or manufacture a man-made ... [SEP]	No	make%2:36:01::
[CLS] We have made no such statement [SEP] make: perform or carry out [SEP]	Yes	make%2:36:12::
...		

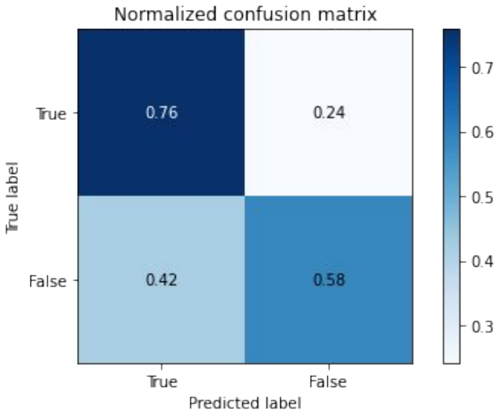


Word Sense Disambiguation (WSD) of WiC data

Results

SemEval-2007

Model architecture	WSD Accuracy	Epoch	Batch size
BERT_{frozen}	36,48	14	8
BERT_{frozen} + WordNet	**65,61**	14	8
Context-gloss $\text{BERT}_{finetuned}$ (15% SemCor)	60,66	2	8
Context-gloss $\text{DistilBERT}_{finetuned}$ (50% SemCor)	60,00	5	16



WiC data

Model architecture	WSD Accuracy	WiC Accuracy	Epoch	Batch size
Word-level MLP (HW1) (Gasparini, 2021)	-	67,70	7	32
BERT_{frozen} + WordNet	58,10	61,77	14	8
Context-gloss $\text{BERT}_{finetuned}$ (15% SemCor)	**60,10**	**68,04**	2	8
Context-gloss $\text{DistilBERT}_{finetuned}$ (50% SemCor)	58,38	68,04	5	16

Word Sense Disambiguation (WSD) of WiC data

Conclusions

- No relevant improvements for WiC in the frozen approach
- Fine-tuning improves performance after few epochs despite the training data reduction
  - WiC performance notably improves (w.r.t. WSD) and is comparable to the task-specific one
- DistilBERT [7] does not make the cut in further improving

Better results may be achieved on both approaches by:

- Re-training the fine-tuning approach on the whole SemCor corpus [8]
- Injecting relatedness knowledge from WordNet in the frozen approach also at training time [9]

Thank you for the attention!

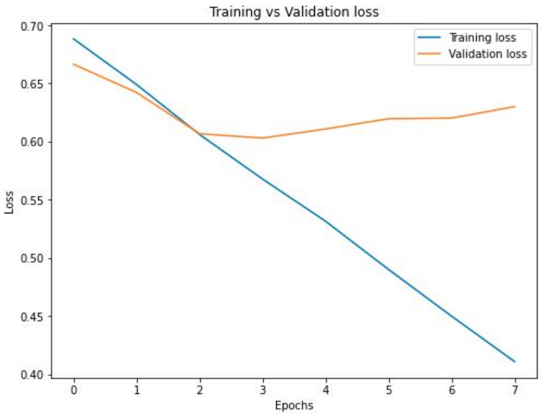
References

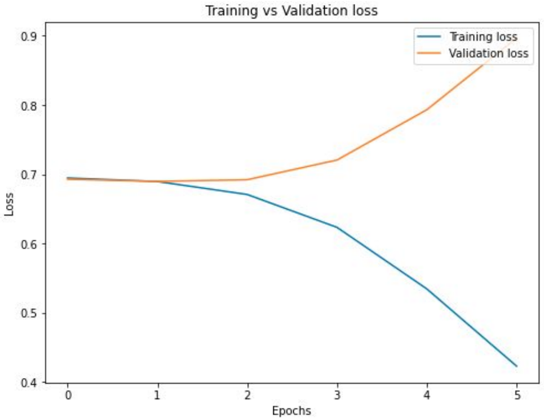
- **Papers:**
  - [1] GloVe: Global Vectors for Word Representation
  - [2] Transformer-based Multilingual and Cross-lingual Word-in-Context Disambiguation
  - [3] BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding
  - [4] Attention Is All You Need
  - [5] WordNet: A Lexical Database for English
  - [6] Word Sense Disambiguation: A Unified Evaluation Framework and Empirical Comparison
  - [7] DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter
  - [8] GlossBERT: BERT for Word Sense Disambiguation with Gloss Knowledge
  - [9] Breaking Through the 80% Glass Ceiling: Raising the State of the Art in Word Sense Disambiguation by Incorporating Knowledge Graph Information
- **Slides:**
  - https://github.com/pietro-nardelli/sapienza-ppt-template

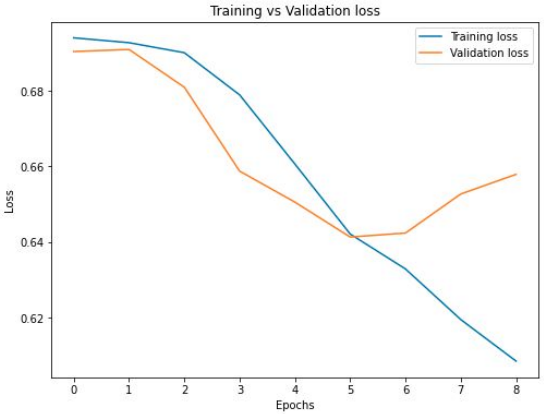
Appendix

Word-in-Context (WiC) disambiguation

Loss histories



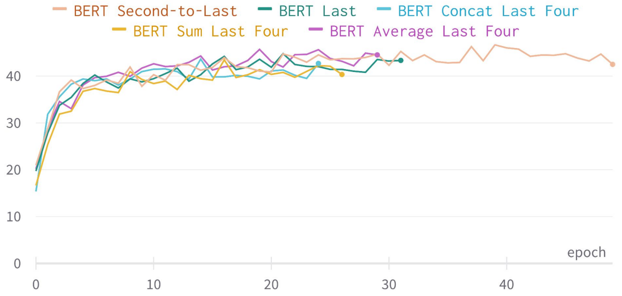




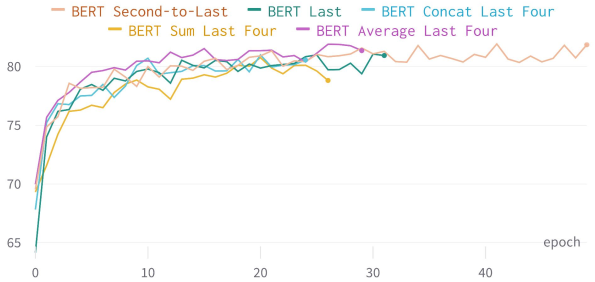
Aspect-Based Sentiment Analysis (ABSA)

BERT pooling strategies experimentation histories

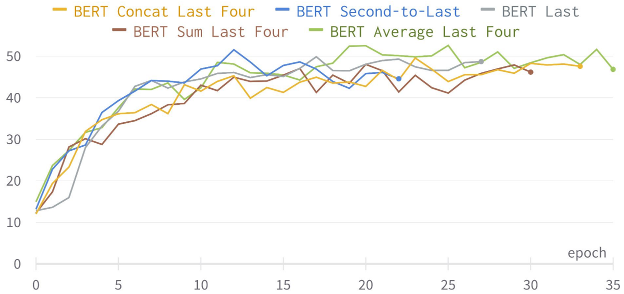
Classification F1-score (task A+B)



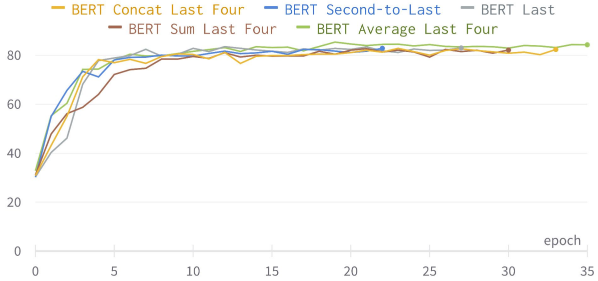
Identification F1-score (task A+B)



Classification F1-score (task C+D)



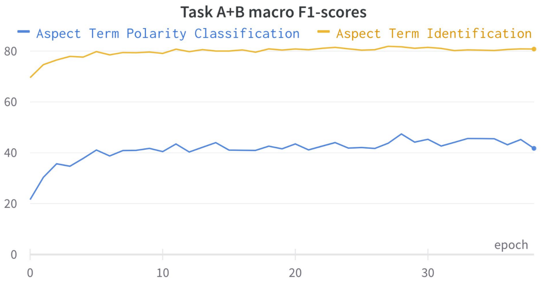
Identification F1-score (task C+D)

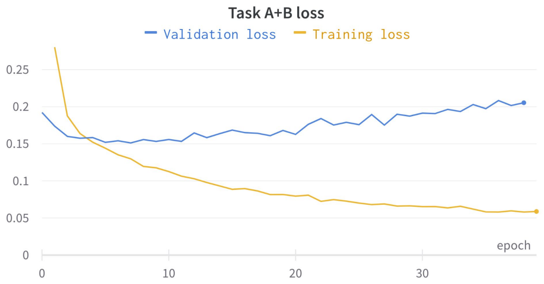


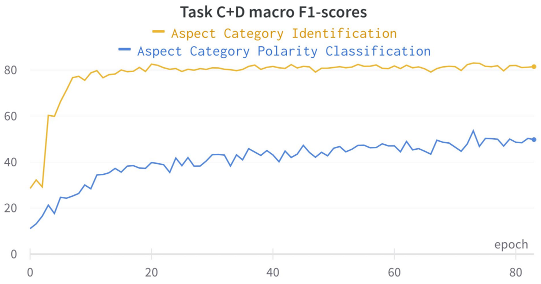
Aspect-Based Sentiment Analysis (ABSA)

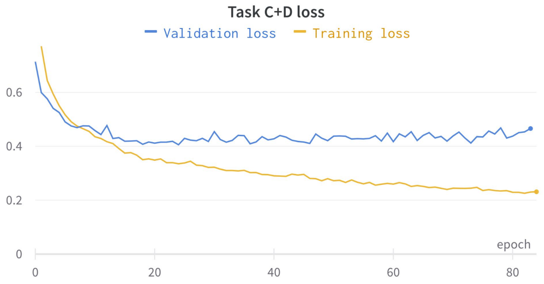
Hyperparameters and resulting histories

Hyperparameter	Task A+B	Task C+D
Epochs	28	73
Random seed	42	42
Optimizer	Adam	Adam
Learning rate	1e-3	1e-3
Loss function	Cross Entropy	Cross Entropy
Batch size	8	8
Static embeddings	GloVe	GloVe
Static embeddings size	300	300
POS embeddings	False	False
POS embeddings size	-	-
LSTM layers	2	2
LSTM bidirectional	True	True
LSTM hidden size	128	128
LSTM input packing	True	True
Dropout	0,5	0,5
BERT model	bert-base-cased	bert-base-cased
BERT finetuning	False	False
BERT layer pooling strategy	second_to_last	mean
BERT pooled layers	-	[-1, -2, -3, -4]
BERT WordPiece pooling strategy	mean	mean
Attention	False	True
Attention heads	-	12
Attention dropout	-	0,2
Concat Attention out to LSTM out	-	False





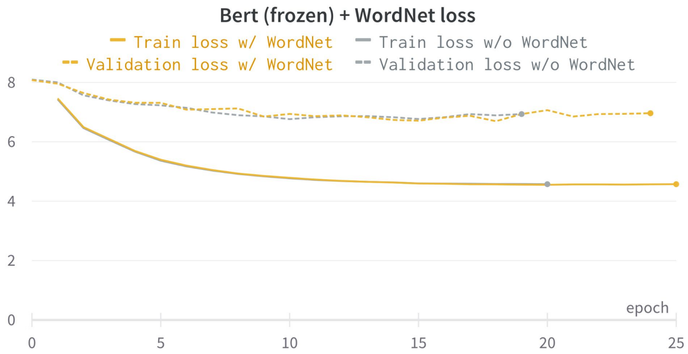


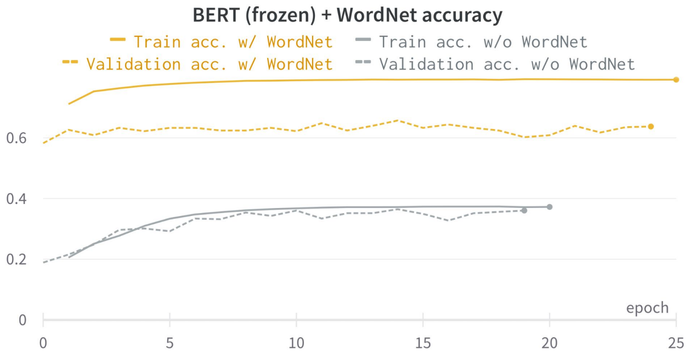


Word Sense Disambiguation (WSD) of WiC data

Hyperparameters and resulting histories

Hyperparameter	Value
Max epoch	100
Early stopping patience	5
Random seed	42
Input size	768
Hidden size	100
Vocabulary size (num. classes)	34.074
Dropout probability	0,2
Learning rate	1e-3
Optimizer	Adam
Loss function	cross-entropy
BERT model	bert-base-cased
BERT layer pooling	last
BERT WordPiece pooling	mean
BERT fine-tuning	false





Word Sense Disambiguation (WSD) of WiC data

Hyperparameters and resulting histories

Hyperparameter	Value
Max epoch	10
Early stopping patience	5
Random seed	42
Learning rate	2e-5
Optimizer	Adam
Loss function	binary cross-entropy
BERT model	bert-base-cased distilbert-base-cased
BERT layer pooling	last
BERT WordPiece pooling	mean
BERT fine-tuning	true