Selection of ACL 2020 Papers

- Joint Modelling of Emotion and Abusive Language Detection
- Intermediate-Task Transfer Learning with Pre-trained Language Models: When and Why Does It Work?
- How does BERT's attention change when you fine-tune? An analysis methodology and a case study in negation scope
- SKEP: Sentiment Knowledge Enhanced Pre-training for Sentiment Analysis
- CamemBERT: a Tasty French Language Model

Joint Modelling of Emotion and Abusive Language Detection

- Paper: https://www.aclweb.org/anthology/2020.acl-main.394.pdf
- Video: https://slideslive.com/38929125/joint-modelling-of-emotion-and-abusive-language-detection

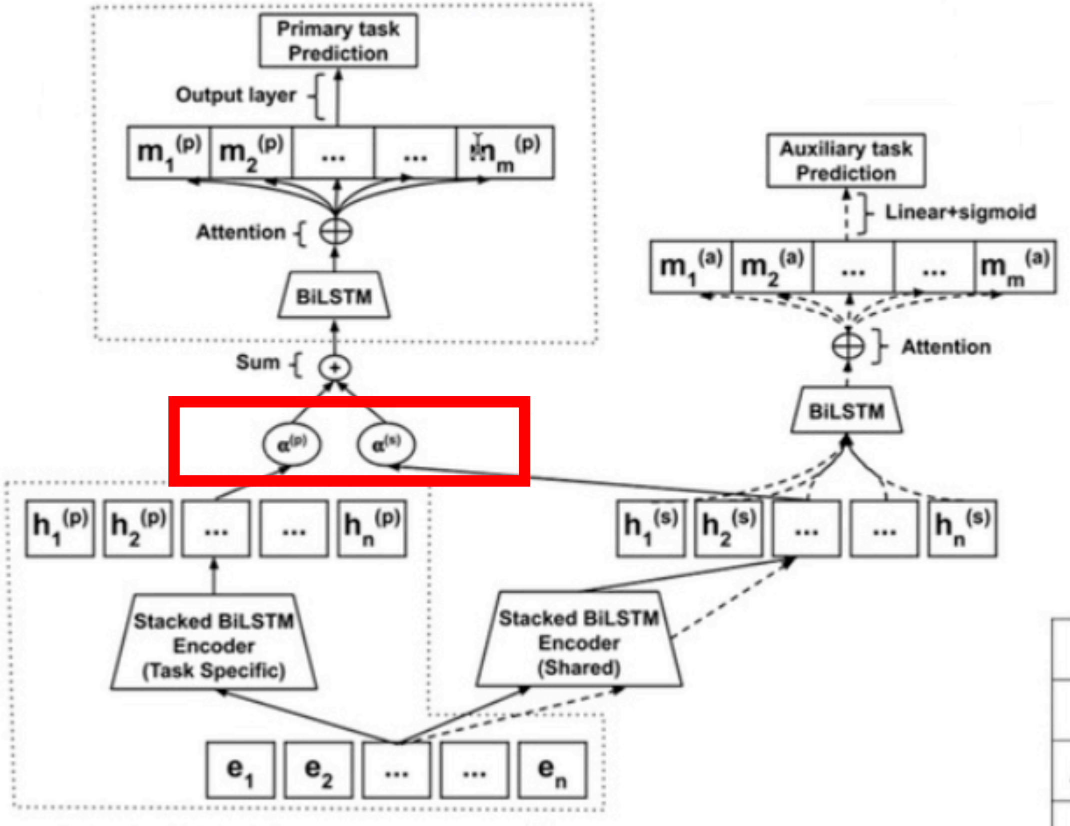
Joint Modelling of Emotion and Abusive Language Detection

- **Aim:** Investigate whether Emotion Detection helps improve the performance of Abuse Detection by leveraging the MTL framework for joint learning and sharing parameters.
- **Primary task:** Abuse Detection; **Auxiliary Task:** Emotion Detection

Joint Modelling of Emotion and Abusive Language Detection

Models

MTL$_{GatedDEncoder}$ - Gated Double Encoder Model



14

e_i	embedding layer output
h_i	encoder output
$m_1 ... m_m$	intermediate representation
α	gate (learnable parameter)

Joint Modelling of Emotion and Abusive Language Detection

Training MTL

Abuse Detection + Emotion Detection

- In a MTL framework, the shared parameters are updated alternately by each task.
- The two task objectives are weighted by a hyperparameter β that controls the importance of each task.
  - With more importance to the primary task
  - (1 – β) for abuse detection and β for emotion detection with β=0.1

Joint Modelling of Emotion and Abusive Language Detection

Model	Precision	Recall	F1
$STL_{BiLSTM+attn}$	77.40	73.27	74.40
MTL_{Hard}	77.21	73.30	74.51
$MTL_{DEncoder}$	**77.47**	73.82	74.97
$MTL_{GatedDEncoder}$	77.46	**75.27**	**76.03***

(a) Twitter - OffensEval results

Model	Precision	Recall	F1
$STL_{maxpool+MLP}$	79.39	78.20	78.33
MTL_{Hard}	79.34	77.61	77.90
$MTL_{DEncoder}$	**80.77**	78.18	79.02
$MTL_{GatedDEncoder}$	80.12	**79.60**	**79.55***

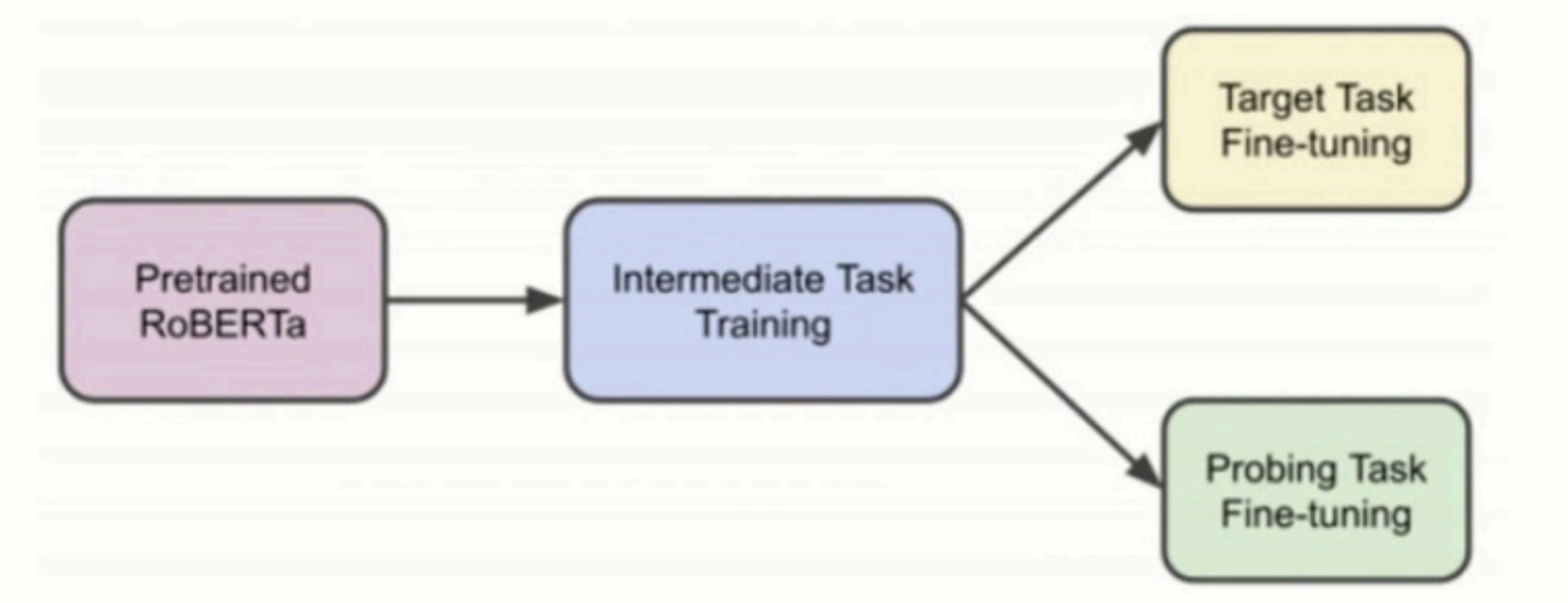
(b) Twitter - Waseem&Hovy results

* indicates statistically significant improvement over STL

Intermediate-Task Transfer Learning with Pre-trained Language Models: When and Why Does It Work?

- Paper: https://www.aclweb.org/anthology/2020.acl-main.467.pdf
- Video: https://slideslive.com/38929152/intermediatetask-transfer-learning-with-pretrained-models-for-natural-language-understanding-when-and-why-does-it-work

Intermediate-Task Transfer Learning with Pre-trained Language Models: When and Why Does It Work?



Intermediate-Task Transfer Learning with Pre-trained Language Models: When and Why Does It Work?

Intermediate tasks:

SocialIQA (QA), QA-SRL (QA), QAMR (QA), Cosmos QA (QA), CommonsenseQA (QA), SciTail (NLI), MNLI (NLI), CCG (tagging), QQP (paraphrase detection), HellaSWAG (sentence completion), SST-2 (sentiment analysis)

Target tasks:

COPA (QA), MultiRC (QA), CommonsenseQA (QA), Cosmos QA (QA), BoolQ (QA), ReCoRD (QA), CommitmentBank (NLI), WSC (coreference resolution), RTE (NLI), WiC (word sense disambiguation)

Probing Tasks

Edge probing tasks (Tenney et al., 2019)

Acceptability Judgment Tasks (Kim et al., 2019, Warstadt et al., 2019)

SentEval probing suite (Conneau et al., 2018)

Intermediate-Task Transfer Learning with Pre-trained Language Models: When and Why Does It Work?

Target	QAMR	CSenseQA	SciTail	CosmosQA	SocialIQA	CCG	HellaSwag	QA-SRL	SST-2	QQP	MNLI	Baseline Performance
CB	-4.0	-0.4	-6.2	-0.4	-21.7	-12.2	-3.1	-7.2	-1.2	-31.0	-0.4	99.1
COPA	-4.0	8.7	4.3	6.0	-3.7	-20.7	6.7	-3.7	-2.0	0.7	-0.7	86.0
WSC	-0.3	0.0	1.3	2.9	-4.8	-3.2	3.6	4.8	2.6	-3.8	0.3	67.3
RTE	0.6	3.4	3.4	5.1	-4.3	-18.2	4.8	1.1	2.6	-2.4	3.1	83.5
MultiRC	2.4	7.9	2.6	10.1	-10.6	-8.1	6.8	2.6	1.1	-4.2	6.5	47.4
WiC	-1.3	0.1	2.5	1.7	-2.0	-1.1	0.1	2.1	-6.4	1.4	0.9	70.5
BoolQ	-0.1	0.9	0.1	1.1	-2.8	-10.6	0.7	0.0	0.9	-4.2	1.4	86.6
CSenseQA	-4.7	-1.6	-2.6	0.1	-7.8	-12.0	0.4	-5.1	-0.9	-7.6	-2.6	74.0
CosmosQA	-2.5	-0.1	-2.1	-0.4	-9.1	-6.9	-0.0	-3.0	-0.0	-8.4	-0.5	81.9
ReCoRD	-4.0	-0.0	-1.5	-0.1	-12.4	-6.1	0.2	-4.7	-0.5	-11.9	-1.6	86.0
Avg. Target	-1.8	1.9	0.2	2.6	-7.9	-9.9	2.0	-1.3	-0.4	-7.1	0.7	78.2

Intermediate-Task Transfer Learning with Pre-trained Language Models: When and Why Does It Work?

Probing

	QAMR	CSenseQA	SciTail	CosmosQA	SocialIQA	CCG	HellaSwag	QA-SRL	SST-2	QQP	MNLI	Baseline Performance
EP-POS	0.0	0.0	-0.0	-0.1	-0.1	-0.0	0.0	-0.0	0.1	-97.4	0.0	98.1
EP-NER	-0.1	0.0	-0.1	-0.1	-21.5	-0.2	0.0	-0.2	0.0	-64.9	-0.3	97.0
EP-SRL	12.2	0.1	30.7	12.4	-61.7	31.2	30.9	31.1	31.9	-61.9	31.3	61.9
EP-Coref	0.0	0.0	0.0	0.1	-0.6	-0.3	0.1	0.0	-0.1	-13.4	0.1	97.1
EP-Const	-0.0	-0.1	-0.1	0.0	-0.0	-0.2	-0.1	0.0	-0.9	-0.2	-0.1	88.8
EP-SPR1	-0.2	0.1	0.1	0.2	-1.7	-0.4	0.2	0.1	0.3	-21.9	0.2	87.2
EP-SPR2	-0.2	-0.0	-0.1	0.1	-3.9	-0.4	-0.1	-0.3	-0.1	-8.2	-0.1	83.8
EP-DPR	7.5	7.9	7.3	8.6	-15.6	3.5	8.3	8.2	7.9	-14.7	6.6	81.4
EP-Rel	0.1	-25.0	0.4	0.1	-55.1	0.2	0.4	-28.8	0.8	-85.4	0.1	85.4
EP-UD	-0.2	0.0	0.0	0.1	-62.0	-0.2	0.0	-0.1	0.1	-89.7	-0.0	95.8
SE-SentLen	-0.0	-0.2	-0.1	-0.3	-0.4	0.5	-0.1	0.1	0.1	-0.9	-0.2	46.4
SE-WC	-0.1	-0.0	-0.0	-0.0	-33.3	-0.0	0.0	-0.0	-0.0	-33.8	-0.0	99.8
SE-TreeDepth	0.1	-0.1	-0.1	-0.1	-1.1	0.3	-0.5	-0.1	-0.1	-1.4	-0.6	76.1
SE-TopConst	-0.2	-0.3	-0.3	-0.1	-0.4	-0.2	-0.2	-0.2	-0.2	-0.4	-0.3	93.5
SE-BShift	-0.1	0.2	0.1	0.0	-0.4	-0.2	0.2	0.0	0.1	-0.1	0.1	97.7
SE-Tense	-1.1	-0.4	-0.5	-0.0	-0.3	-1.3	0.0	-0.8	-0.2	-1.5	-1.2	91.1
SE-SubjNum	0.3	0.5	0.4	0.9	-0.1	0.8	0.8	0.2	0.5	-0.1	0.4	93.3
SE-ObjNum	-0.6	-0.1	-0.1	0.0	-0.5	0.2	-0.3	0.2	-0.4	0.2	-0.1	95.7
SE-SOMO	-2.2	0.4	-1.1	0.1	-4.1	-3.6	0.2	-1.8	-1.0	-2.5	-1.2	77.2
SE-CoordInv	-0.7	-0.1	-0.4	-0.2	-1.3	-1.0	-0.0	-0.3	-0.2	-3.0	-0.1	88.3
AJ-CoLA	-2.6	-0.7	-1.9	-1.6	-10.3	-6.9	-0.7	-3.7	-0.6	-5.5	-1.1	68.1
AJ-Wh	13.4	26.8	3.4	14.5	14.2	26.8	14.5	28.4	28.4	3.8	11.8	69.9
AJ-Def	23.1	46.0	11.1	0.0	18.0	46.4	32.4	22.5	14.0	11.1	23.7	47.2
AJ-Coord	25.2	17.7	11.1	20.2	22.3	32.6	11.1	22.2	17.4	11.1	11.1	47.2
AJ-EOS	11.9	13.2	13.9	13.2	-21.3	8.5	5.0	11.8	-4.5	-13.9	6.0	84.7

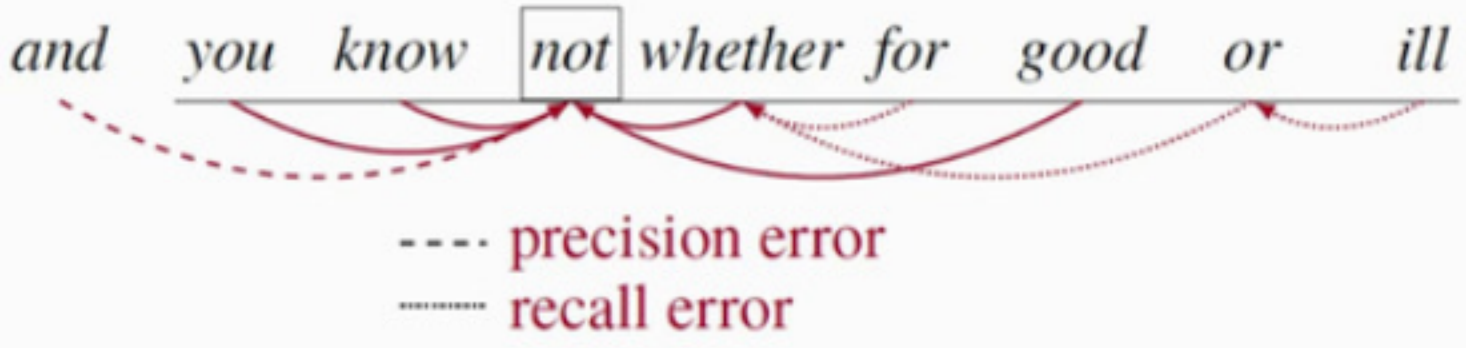
- We find that HellaSWAG, Cosmos QA, and CommonsenseQA transfer most positively to downstream tasks.
- Our proposed approach of using probing tasks to identify skills learned from intermediate task training wasn't as informative as we'd hoped.
- Higher level semantic abilities tend to have a higher correlation with the target-task performance.
- We find catastrophic forgetting to be a limitation of our experiments.
- Preliminary results show that further training on MLM yields mixed results.
    - MTL with MLM during Intermediate Task Training phase
    - Mitigates negative transfer, but also decreases positive transfer

How does BERT's attention change when you fine-tune? An analysis methodology and a case study in negation scope

- Paper: https://www.aclweb.org/anthology/2020.acl-main.429.pdf
- Video: https://slideslive.com/38928830/how-does-berts-attention-change-when-you-finetune-an-analysis-methodology-and-a-case-study-in-negation-scope

How does BERT's attention change when you fine-tune? An analysis methodology and a case study in negation scope

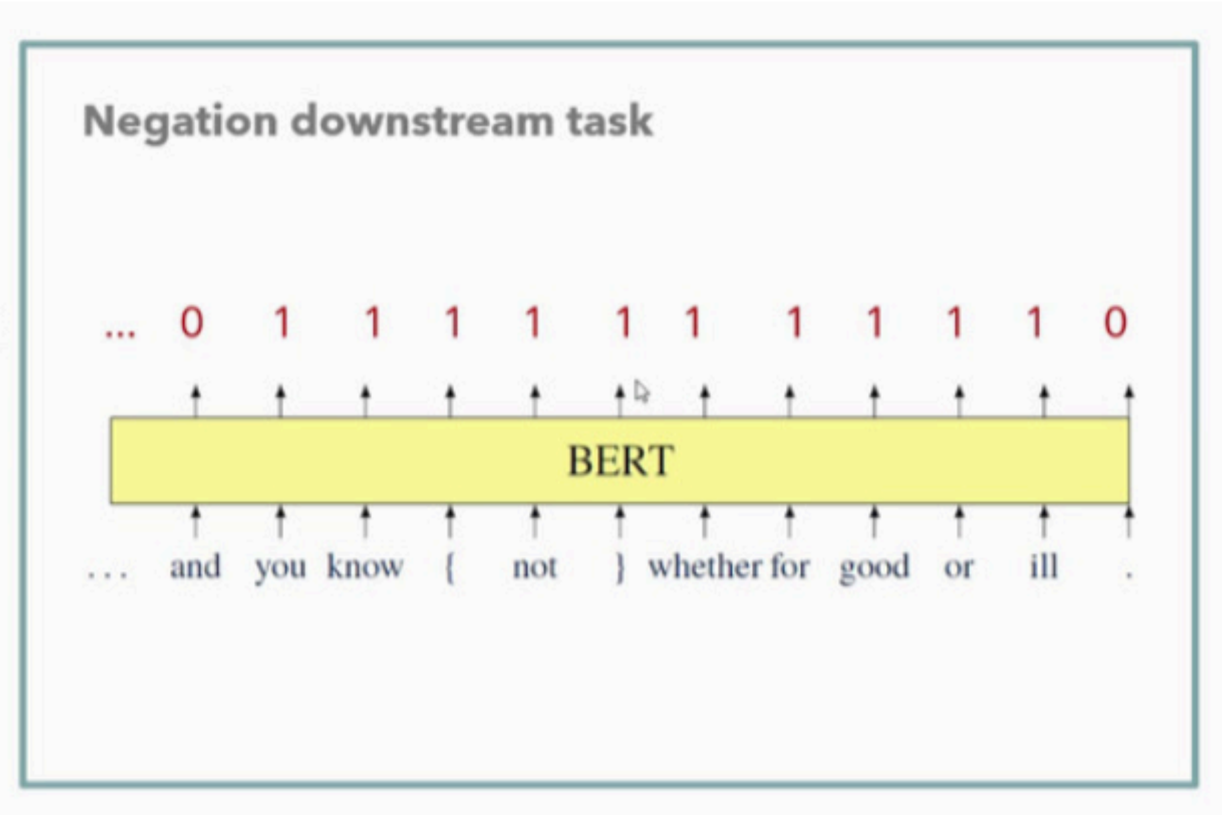
$$attendneg(i) = \begin{cases} 1 & \text{if } \operatorname*{argmax}_{j=1}^{n} a_{ij} = j_{neg} \\ 0 & otherwise \end{cases}$$

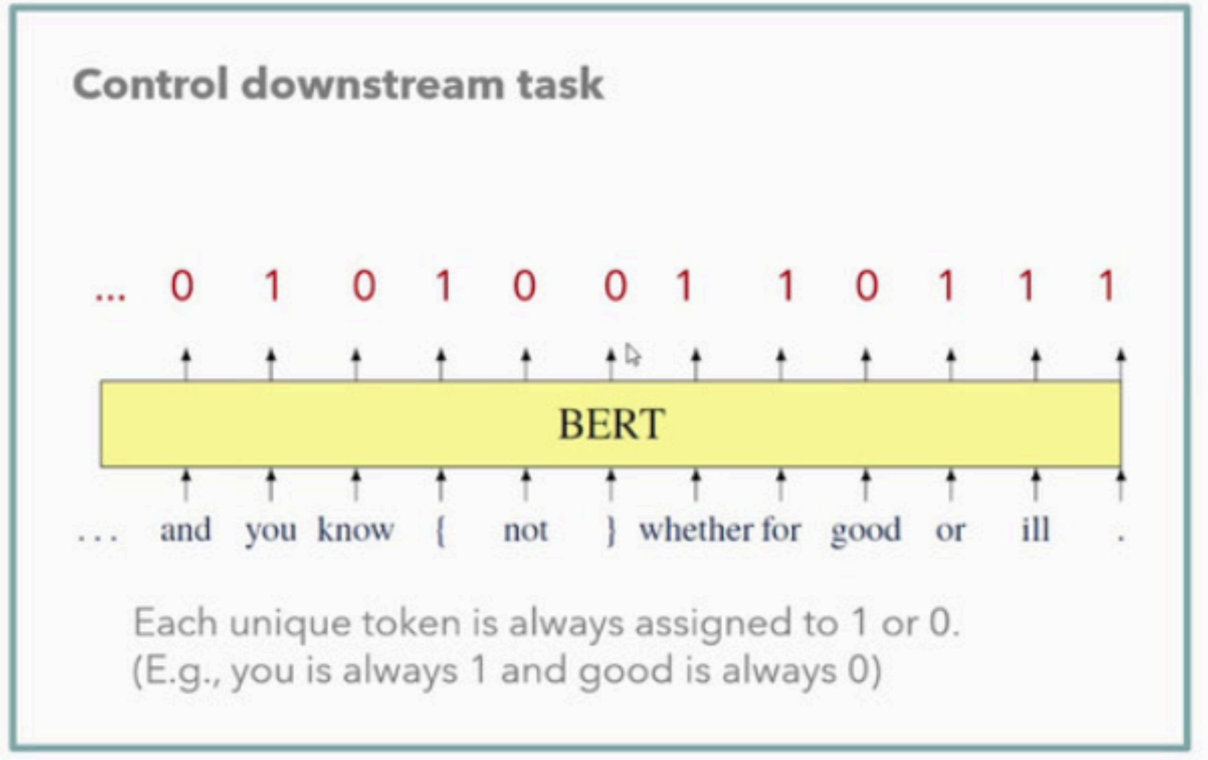


Step 1: hypothesize an interpretable representation of the phenomenon of interest.
Words in scope pay maximum Attention to the negation cue.

How does BERT's attention change when you fine-tune? An analysis methodology and a case study in negation scope

Step 2: Construct a related downstream task and a control task and fine-tune the models.

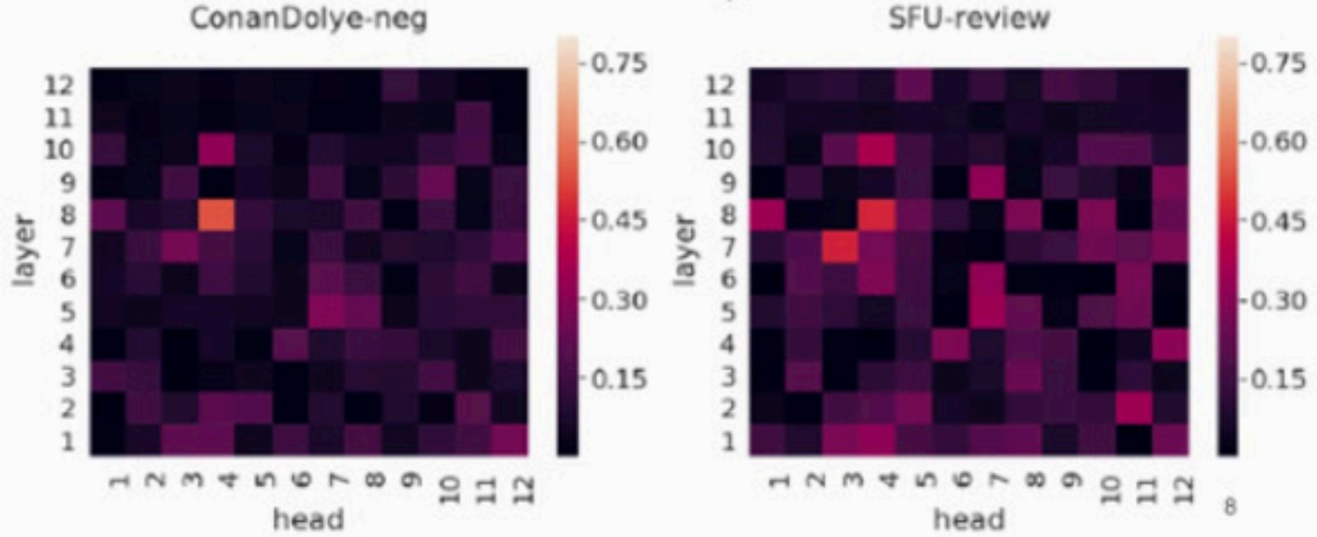




BEFORE FINE-TUNE: NEGATION SCOPE KNOWLEDGE IN SOME HEADS.

- Many heads in the pretrained BERT-base do not encode the negation scope in this way.
- Some attention heads could outperform the offset baseline.
- There is a consistency of attention head performance across different datasets (Kendall tau correlation: 0.415)

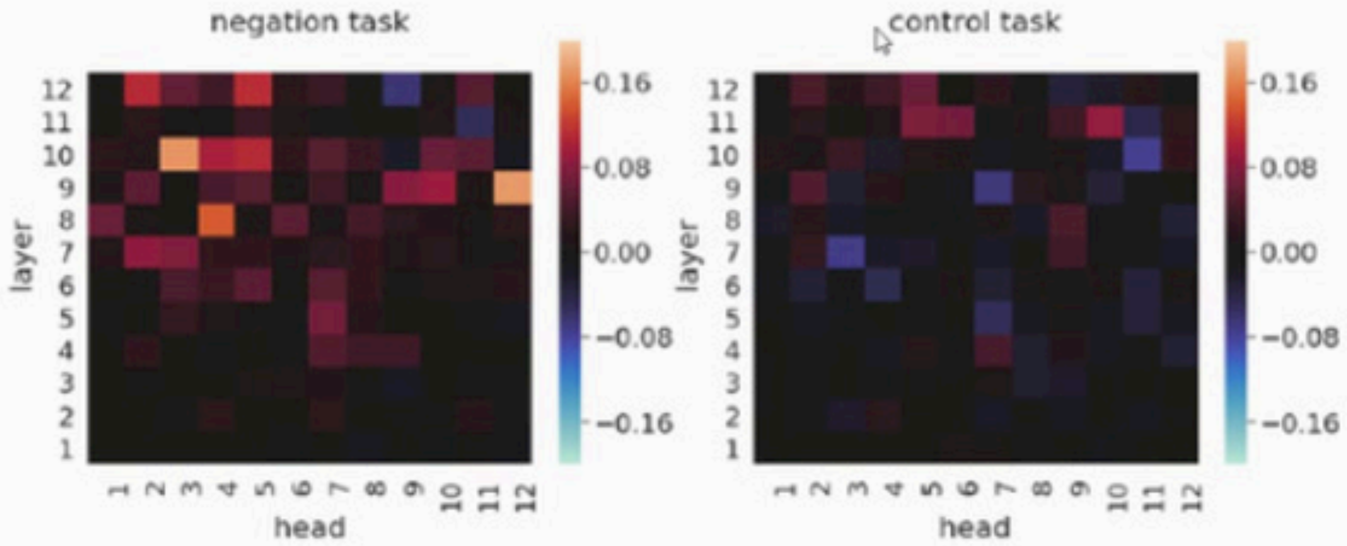
Models		F_1
baseline	average fixed offset	15.2
baseline	best fixed offset (-1)	20.4
attention	average head	9.0
attention	best head (8-4)	**53.8**



AFTER FINE-TUNE: ATTENTION PATTERN ENHANCED IN NEGATION TASK.

- Fine-tuning on the negation-scope task enhance the hypothesized attention patterns.
- Fine-tuning on the control task does not show the similar enhancement.
- On average, fine-tuning on the negation-related task introduces more obvious change than the control task.
- This provides evidence that the hypothesized encoding of negation scope is related to the downstream task.

Attention Head	Fine-Tune	$F_1 \pm$ sd
Average	None	9.0
Average	Control	9.0 ± 0.4
Average	Negation	11.1 ± 1.2
Best (8-4)	None	53.8
Best (8-4)	Control	53.1 ± 6.7
Best (8-4)	Negation	67.7 ± 7.9



F1 change after fine-tuning for each attention heads over the negation task and control task on the ConanDolye datasets

- If an attention head has a high performance before fine-tuning, will it increase in performance more than other attention heads?
- The rich do not get richer: attention heads that had the top F_1s in the pretrained model do not have the top-ranked improvements after fine-tuning on negation scope.

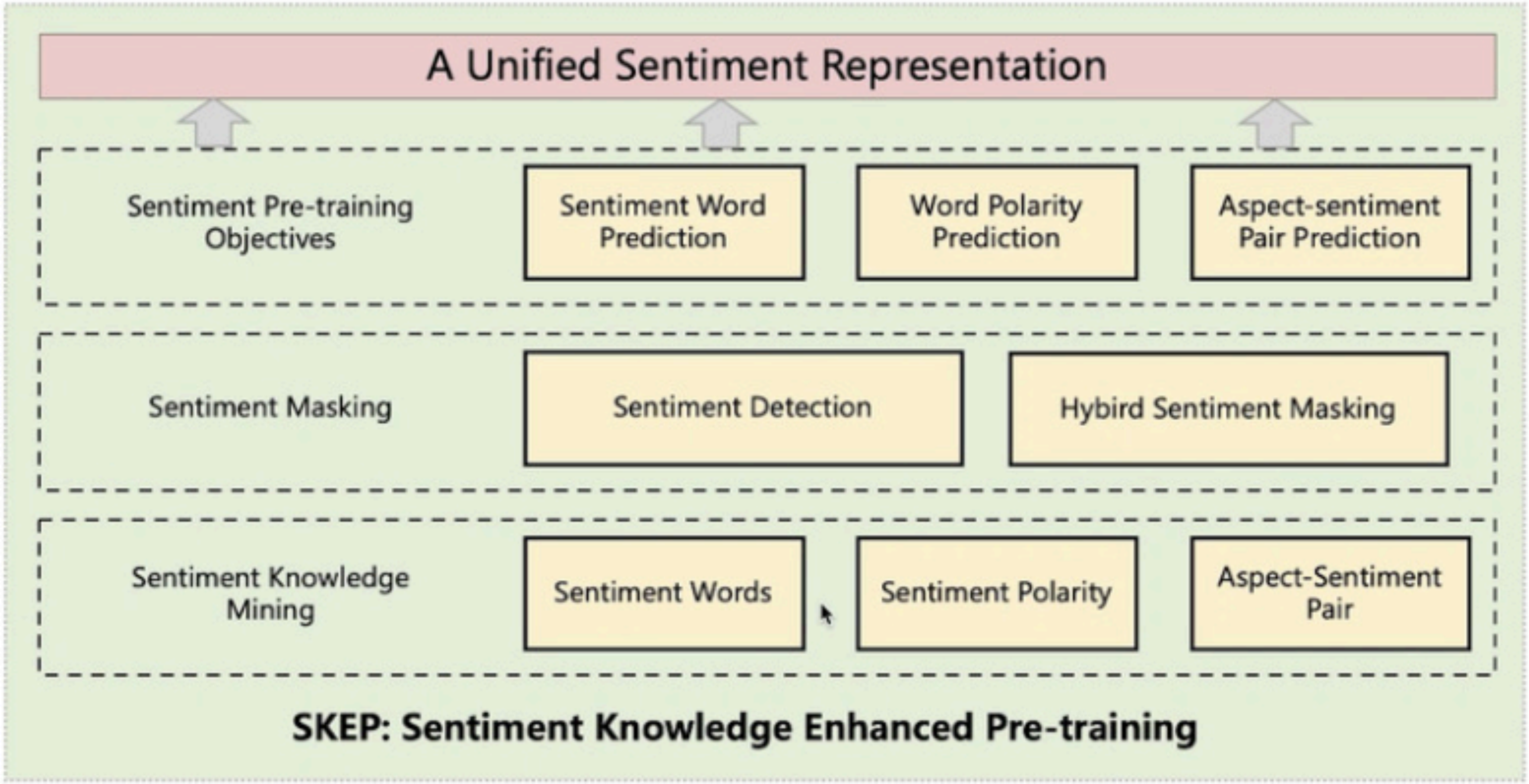
Negation change	Precision			Recall			F_1		
	τ	pos/neg	sig	τ	pos/neg	sig	τ	pos/neg	sig
BERT-base	-0.065	0/3	3/10	0.096	5/0	5/10	0.085	5/0	5/10
BERT-large	-0.098	2/5	7/10	-0.132	0/7	7/10	-0.132	0/8	8/10
RoBERTa-base	-0.134	0/7	7/10	-0.107	0/5	5/10	-0.113	0/6	6/10
RoBERTa-large	-0.155	0/8	8/10	-0.142	0/8	8/10	-0.144	0/8	8/10

Kendall rank correlation (τ) between the change of an attention head after fine-tuning on the negation task and its performance in the pretrained model.

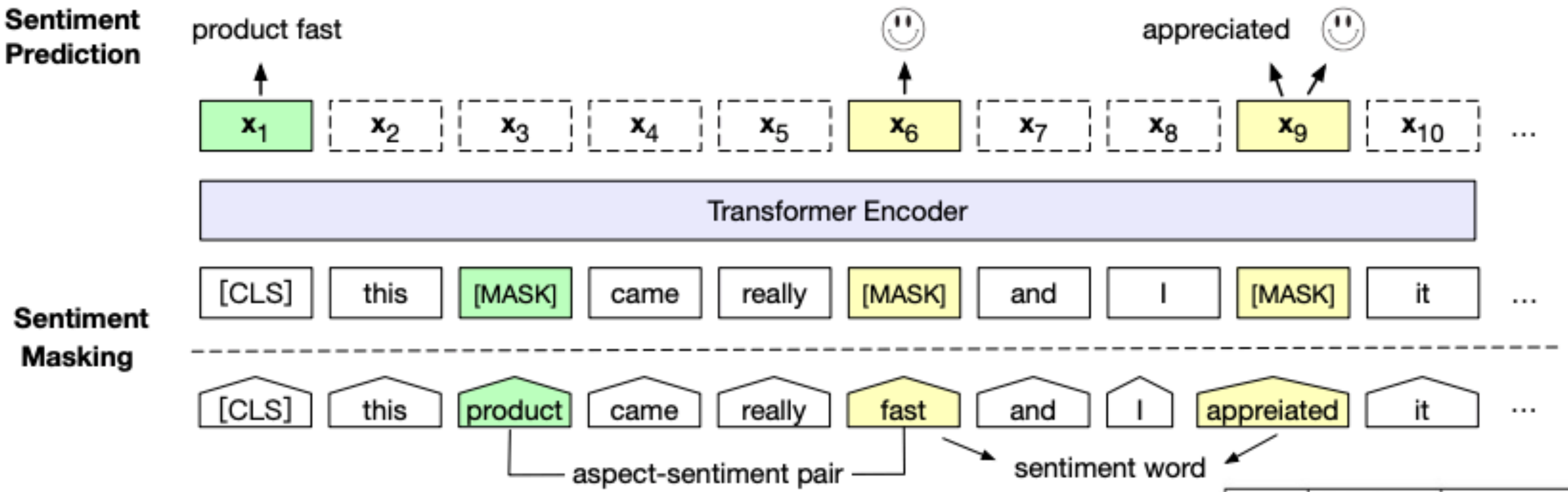
SKEP: Sentiment Knowledge Enhanced Pre-training for Sentiment Analysis

- Paper: https://www.aclweb.org/anthology/2020.acl-main.374.pdf
- Video: https://slideslive.com/38929227/skep-sentiment-knowledge-enhanced-pretraining-for-sentiment-analysis

SKEP: Sentiment Knowledge Enhanced Pre-training for Sentiment Analysis



SKEP: Sentiment Knowledge Enhanced Pre-training for Sentiment Analysis



step1	Aspect-sentiment Pair	At most 2 aspect-sentiment pairs are selected to mask.
step2	Sentiment word	The total number of masked token is limited to be less than 10%
step3	Common Token	If the number of masked token in step2 is insufficient, then run step3

The three prediction objectives on top are jointly optimized:

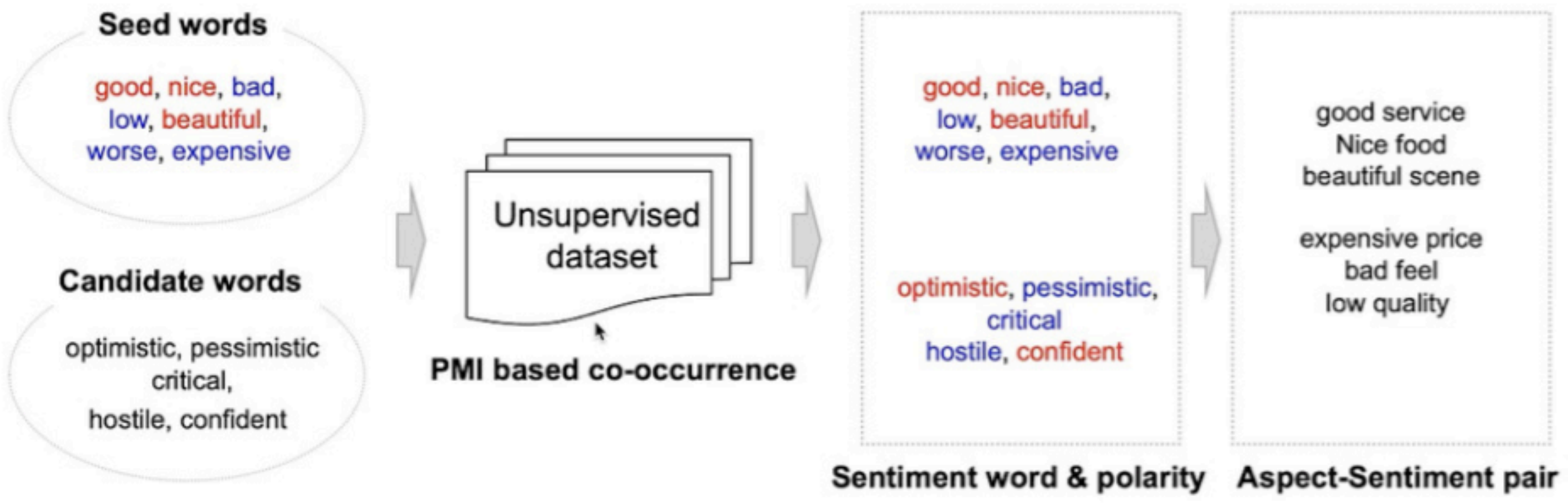
Sentiment Word (SW) prediction (on x9),

Word Polarity (SP) prediction (on x6 and x9),

Aspect-Sentiment pairs (AP) prediction (on x1).

SKEP: Sentiment Knowledge Enhanced Pre-training for Sentiment Analysis

- We use a simple and effective mining method based on Pointwise Mutual Information (Turney, 2002)



SKEP: Sentiment Knowledge Enhanced Pre-training for Sentiment Analysis

	Sentence-Level		Aspect-Level		Opinion Role	
Model	SST-2	Amazon-2	Sem-L	Sem-R	MPQA-Holder	MPQA-Target
Previous SOTA	**97.1**[1] $_{*}$	97.37[2]	81.35[3]	87.89[4]	83.67/77.12[5]	81.59/73.16[5]
RoBERTa$_{base}$	94.9	96.61	78.11	84.93	81.89/77.34	80.23/72.19
RoBERTa$_{base}$ + SKEP	96.7	96.94	81.32	87.92	84.25/79.03	82.77/74.82
RoBERTa$_{large}$	96.5	97.33	79.22	85.88	83.52/78.59	81.74/75.87
RoBERTa$_{large}$ + SKEP	97.0	**97.56**	**81.47**	**88.01**	**85.77/80.99**	**83.59/77.41**

SKEP: Sentiment Knowledge Enhanced Pre-training for Sentiment Analysis

Model	Sentence-Level		Aspect-Level		Opinion Role	
	SST-2 dev	Amazon-2	Sem-L	Sem-R	MPQA-Holder	MPQA-Target
RoBERTa_{base}	95.21	96.61	78.11	84.93	81.89/77.34	80.23/72.19
+ Random Token	95.57	96.73	78.89	85.77	82.71/77.71	80.86/73.01
+ SW	96.38	96.82	80.13	86.92	82.95/77.63	81.18/73.15
+ SW + WP	96.51	96.87	80.32	87.25	82.97/77.82	81.09/73.24
+ SW + WP + AP	96.87	96.94	81.32	87.92	84.25/79.03	82.77/74.82
+ SW + WP + AP-I	96.89	96.93	81.19	87.71	84.01/78.36	82.69/74.36

- Random Token and SW+WP+AP: sentiment knowledge is helpful
- SW+WP+AP: diverse knowledge results in better performance
- AP vs AP-I: multi-label classification is efficient
- Verify the necessity of incorporating sentiment knowledge for PTMs

SKEP: Sentiment Knowledge Enhanced Pre-training for Sentiment Analysis

From	Model	Sentence Samples	Prediction
SST-2	RoBERTa	*altogether , this is successful as a film , while at the same time being a most touching reconsideration of the familiar masterpiece .*	positive
	SKEP	*altogether , this is successful as a film , while at the same time being a most touching reconsideration of the familiar masterpiece .*	positive
Sem-L	RoBERTa	*I got this at an amazing price from Amazon and it arrived just in time .*	negative
	SKEP	*I got this at an amazing price from Amazon and it arrived just in time .*	positive

Table shows the attention distribution of final layer for the [CLS] token when they adopt the SKEP model to classify the input sentences.

CamemBERT: a Tasty French Language Model

- Paper: https://www.aclweb.org/anthology/2020.acl-main.645.pdf
- Video: https://slideslive.com/38929193/camembert-a-tasty-french-language-model

CamemBERT: a Tasty French Language Model

CamemBERT is trained on OSCAR.

- **OSCAR** is a clean **extract** of **Common Craw** (Ortiz et al., 2019).
- **Open-source and freely available** at oscar-corpus.com.
- **French** data: **138GB of text**, 32.7B tokens, 59.4M documents.
- **Heterogeneous** data with **diverse styles** and **domains**.

Tasks and baselines:

- **Part-Of-Speech Tagging (POS)**: mBERT, XLM, UDify, and UDPipe Future
- **Dependency Parsing**: Same as POS tagging
- **Named Entity Recognition (NER)**: CRF, BiLSTM-CRF, and mBERT.
- **Natural Language Inference**: mBERT, XLM, and XLM-R

Two evaluation settings:

- **Fine-tuned**: CamemBERT is fine-tuned on the downstream tasks
- **As Embeddings**: Freeze CamemBERT, use output embeddings as input to another model

CamemBERT: a Tasty French Language Model

- What type of training data?

  - **OSCAR vs. CCNet vs. Wikipedia**
    - **CCNet:** Common Crawl filtered with Wikipedia language model.

⇒ **Uniformity (Wikipedia)** is **detrimental** in all cases, largest gap in NLI

DATASET	SIZE	GSD		SEQUOIA		SPOKEN		PARTUT		AVERAGE		NER	NLI
		UPOS	LAS	UPOS	LAS	UPOS	LAS	UPOS	LAS	UPOS	LAS	F1	ACC.
Fine-tuning													
Wiki	4GB	98.28	93.04	98.74	92.71	96.61	79.61	96.20	89.67	97.45	88.75	89.86	78.32
CCNet	4GB	98.34	93.43	98.95	93.67	96.92	**82.09**	96.50	**90.98**	97.67	**90.04**	90.46	**82.06**
OSCAR	4GB	98.35	93.55	98.97	93.70	96.94	81.97	96.58	90.28	97.71	89.87	90.65	81.88
OSCAR	138GB	**98.39**	**93.80**	**98.99**	**94.00**	**97.17**	81.18	**96.63**	90.56	**97.79**	89.88	**91.55**	81.55

CamemBERT: a Tasty French Language Model

- How much training data?

  - **4GB vs. 138GB**

⇒ **Competitive results with** as few as **4GB of data** for a Base model!
Proves that strong models can be trained even on low resource languages or domain-specific datasets.
Similar findings in (Kaplan et al., 2020) for Base architecture.

Dataset	Size	GSD		Sequoia		Spoken		ParTUT		Average		NER	NLI
		UPOS	LAS	UPOS	LAS	UPOS	LAS	UPOS	LAS	UPOS	LAS	F1	Acc.
Fine-tuning													
Wiki	4GB	98.28	93.04	98.74	92.71	96.61	79.61	96.20	89.67	97.45	88.75	89.86	78.32
CCNet	4GB	98.34	93.43	98.95	93.67	96.92	**82.09**	96.50	**90.98**	97.67	**90.04**	90.46	**82.06**
OSCAR	4GB	98.35	93.55	98.97	93.70	96.94	81.97	96.58	90.28	97.71	89.87	90.65	81.88
OSCAR	138GB	**98.39**	**93.80**	**98.99**	**94.00**	**97.17**	81.18	**96.63**	90.56	**97.79**	89.88	**91.55**	81.55

Other works:

- Learning and Evaluating Emotion Lexicons for 91 Languages
  - https://www.aclweb.org/anthology/2020.acl-main.112.pdf
  - https://slideslive.com/38929408/learning-and-evaluating-emotion-lexicons-for-91-languages
- Perturbed Masking: Parameter-free Probing for Analyzing and Interpreting BERT
  - https://www.aclweb.org/anthology/2020.acl-main.383.pdf
  - https://slideslive.com/38929032/perturbed-masking-parameterfree-probing-for-analyzing-and-interpreting-bert
- Politeness Transfer: A Tag and Generate Approach
  - https://www.aclweb.org/anthology/2020.acl-main.169.pdf
  - https://slideslive.com/38929267/politeness-transfer-a-tag-and-generate-approach
- Leveraging Pre-trained Checkpoints for Sequence Generation Tasks
  - https://www.mitpressjournals.org/doi/pdf/10.1162/tacl_a_00313
  - https://slideslive.com/38929500/leveraging-pretrained-checkpoints-for-sequence-generation-tasks
- Mike Lewis's talk at RepL4NLP workshop: Beyond BERT
  - https://slideslive.com/38929793/beyond-bert