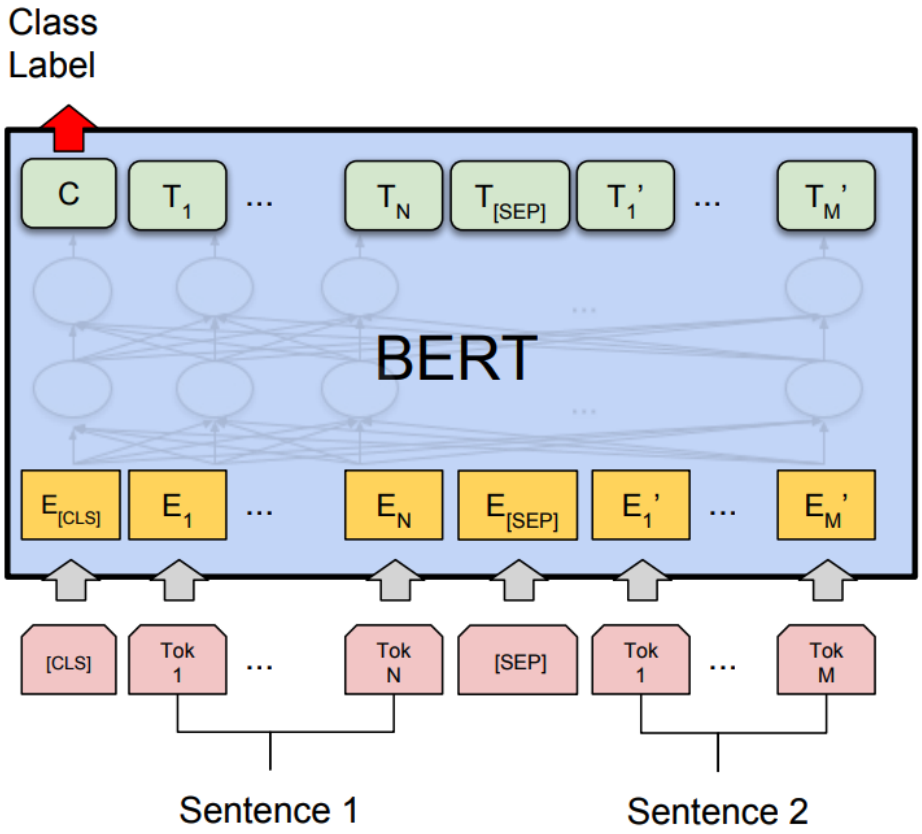
TECHNICKÁ UNIVERZITA V LIBERCI
Fakulta mechatroniky, informatiky
a mezioborových studií

Aplikace jazykových modelů

František Kynych
9. 12. 2021 | MVD

Opakování z předchozí přednášky

BERT


Class
Label
C
T1
...
TN
T[SEP]
T1'
...
TM'
BERT
E[CLS]
E1
...
EN
E[SEP]
E1'
...
EM'
[CLS]
Tok 1
...
Tok N
[SEP]
Tok 1
...
Tok M
Sentence 1
Sentence 2

Část I.:
Rozšíření BERT modelu

RoBERTa

- A Robustly Optimized BERT Pretraining Approach
- Robustnější trénování BERT modelu
  - Větší množství dat
  - Úprava maskování vstupních tokenů
- Trénovací data
  - BERT model (16 GB - BookCorpus + Wikipedia)
  - Původní data +
    - CC-NEWS (76 GB)
    - OpenWebText (38 GB)
    - Stories (31 GB)

RoBERTa

- Maskování tokenů
  - Statické (BERT)
    - Náhodně maskujeme tokeny v trénovacím datasetu
    - Maskování je předpřipraveno a uloženo na disk

    ⇒ V každé epoše máme stejný text vícekrát, ale s jinou maskou

    ⇒ V různých epochách zpracováváme stejný text
  - Dynamické (RoBERTa)
    - V každé epoše generujeme náhodně maskování

    ⇒ V různých epochách jsou masky na různé poloze

Masking	SQuAD 2.0	MNLI-m	SST-2
reference	76.3	84.3	92.8
Our reimplementation:			
static	78.3	84.3	92.5
dynamic	78.7	84.0	92.9

RoBERTa

- Predikce další věty
  - BERT
    - Dvě věty na vstupu - predikce, zda jdou za sebou
    - S 50% pravděpodobností vybereme dvě věty ze stejného dokumentu nebo z různých dokumentů
  - RoBERTa

Model	SQuAD 1.1/2.0	MNLI-m	SST-2	RACE
Our reimplementation (with NSP loss):				
SEGMENT-PAIR	90.4/78.7	84.0	92.9	64.2
SENTENCE-PAIR	88.7/76.2	82.9	92.1	63.0
Our reimplementation (without NSP loss):				
FULL-SENTENCES	90.4/79.1	84.7	92.5	64.8
DOC-SENTENCES	90.6/79.7	84.7	92.7	65.6
$\text{BERT}_{\text{BASE}}$	88.5/76.3	84.3	92.8	64.3
$\text{XLNet}_{\text{BASE}}$ (K = 7)	–/81.3	85.8	92.7	66.1
$\text{XLNet}_{\text{BASE}}$ (K = 6)	–/81.0	85.6	93.4	66.7

RoBERTa

- Odebrání predikce další věty a dynamické maskování zlepšilo výsledky
- Použití větší batch size (2k) pomohlo snížit perplexitu a vylepšit výsledky modelu

Model	data	bsz	steps	SQuAD (v1.1/2.0)	MNLI-m	SST-2
RoBERTa						
with BOOKS + WIKI	16GB	8K	100K	93.6/87.3	89.0	95.3
+ additional data (§3.2)	160GB	8K	100K	94.0/87.7	89.3	95.6
+ pretrain longer	160GB	8K	300K	94.4/88.7	90.0	96.1
+ pretrain even longer	160GB	8K	500K	**94.6/89.4**	**90.2**	**96.4**
$\text{BERT}_{\text{LARGE}}$						
with BOOKS + WIKI	13GB	256	1M	90.9/81.8	86.6	93.7
$\text{XLNet}_{\text{LARGE}}$						
with BOOKS + WIKI	13GB	256	1M	94.0/87.8	88.4	94.4
+ additional data	126GB	2K	500K	94.5/88.8	89.8	95.6

ALBERT

- A Lite BERT for Self-Supervised Learning of Language Representations
- Vylepšení BERT architektury, které posunulo state-of-the-art výsledky

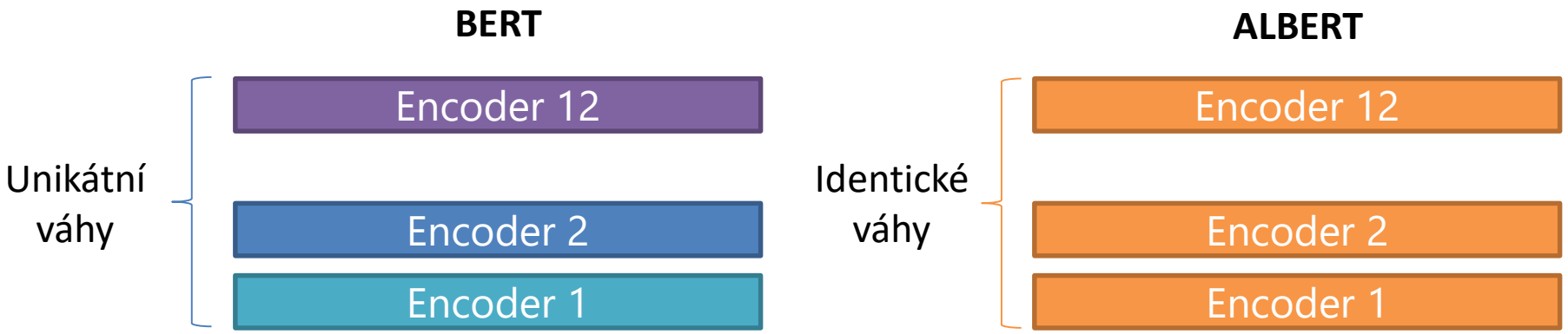
- Nevýhoda BERT modelu
  - Velké množství parametrů

  ⇒Výpočetně náročné trénování i inference
- Snížení parametrů BERT modelu (o 89 %)

⇒ Při menším počtu parametrů můžeme používat ještě větší model

ALBERT

- Cross-Layer Parametr Sharing
  - Sdílení vah napříč modelem
  - Učí se pouze jeden encoder a další vrstvy používají stejné váhy



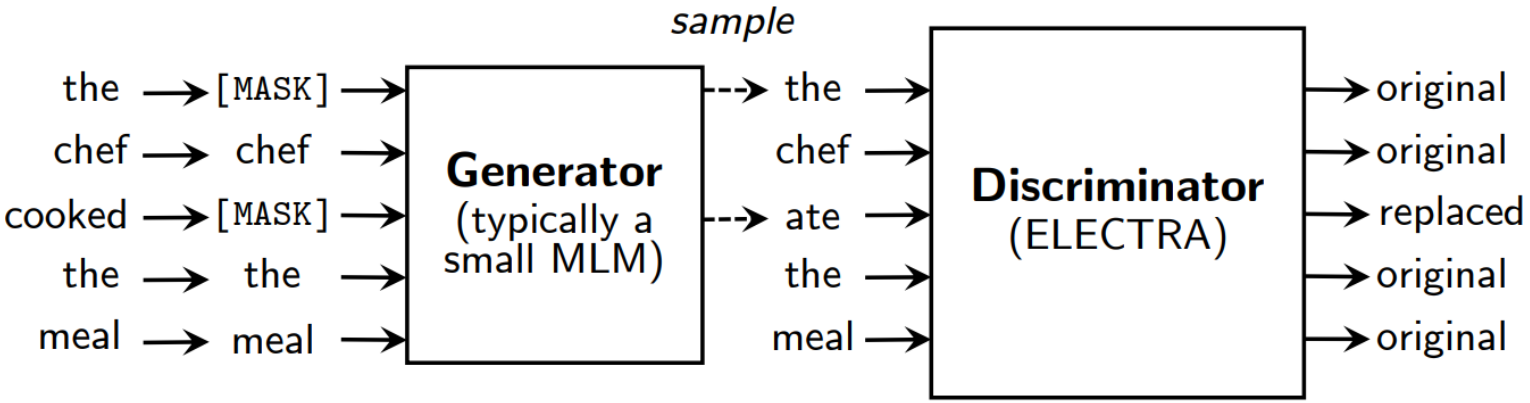
ALBERT

Model		Parameters	Layers	Hidden	Embedding	Parameter-sharing
BERT	base	108M	12	768	768	False
	large	334M	24	1024	1024	False
ALBERT	base	12M	12	768	128	True
	large	18M	24	1024	128	True
	xlarge	60M	24	2048	128	True
	xxlarge	235M	12	4096	128	True

- Sentence Order Prediction
  - Vybereme dvě věty z dokumentu
    - Ve správném pořadí -> pozitivní pár
    - Zaměníme jejich pořadí -> negativní pár

ELECTRA

- Efficiently Learning an Encoder that Classifies Token Replacements Accurately
- Úprava maskování vstupních tokenů u BERT modelu:
  - Generátor
    - Učí se stejně jako BERT model (nepoužívá se klasifikace pořadí vět)
  - Diskriminátor
    - Replaced Token Detection

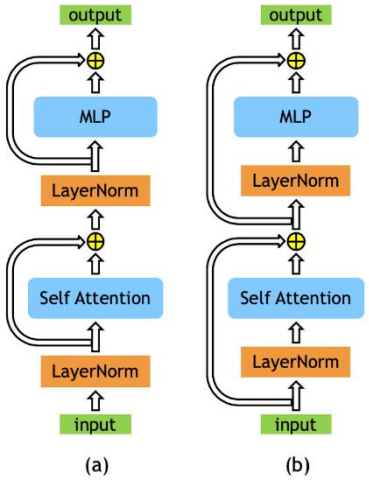


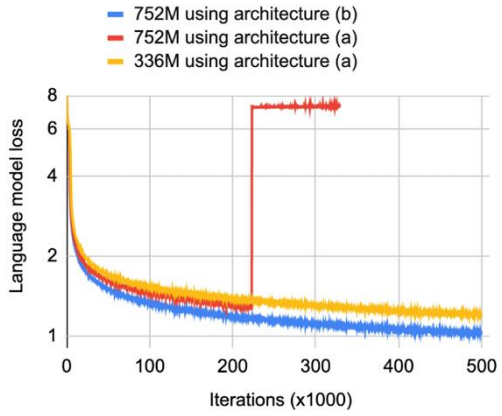
ELECTRA

Model	Train FLOPs	Params	SQuAD 1.1 dev		SQuAD 2.0 dev		SQuAD 2.0 test	
			EM	F1	EM	F1	EM	F1
BERT-Base	6.4e19 (0.09x)	110M	80.8	88.5	–	–	–	–
BERT	1.9e20 (0.27x)	335M	84.1	90.9	79.0	81.8	80.0	83.0
SpanBERT	7.1e20 (1x)	335M	88.8	94.6	85.7	88.7	85.7	88.7
XLNet-Base	6.6e19 (0.09x)	117M	81.3	–	78.5	–	–	–
XLNet	3.9e21 (5.4x)	360M	**89.7**	**95.1**	87.9	**90.6**	87.9	90.7
RoBERTa-100K	6.4e20 (0.90x)	356M	–	94.0	–	87.7	–	–
RoBERTa-500K	3.2e21 (4.5x)	356M	88.9	94.6	86.5	89.4	86.8	89.8
ALBERT	3.1e22 (44x)	235M	89.3	94.8	87.4	90.2	88.1	90.9
BERT (ours)	7.1e20 (1x)	335M	88.0	93.7	84.7	87.5	–	–
ELECTRA-Base	6.4e19 (0.09x)	110M	84.5	90.8	80.5	83.3	–	–
ELECTRA-400K	7.1e20 (1x)	335M	88.7	94.2	86.9	89.6	–	–
ELECTRA-1.75M	3.1e21 (4.4x)	335M	**89.7**	94.9	**88.0**	**90.6**	**88.7**	**91.4**

Megatron

- BERT i RoBERTa architektury ukázaly, že větší modely přinesou lepší výsledky
  - Je možné dosáhnout zlepšení s ještě většími modely?
- Předchozí experimenty s větším BERT modelem degradovaly úspěšnost
- Megatron řeší úpravu v architektuře, aby bylo možné natrénovat i větší modely





https://developer.nvidia.com/blog/language-modeling-using-megatron-a100-gpu/

Megatron

Model	Trained tokens (ratio)	MNLI m/mm accuracy (dev set)	QQP accuracy (dev set)	SQuAD 1.1 F1/EM (dev set)	SQuAD 2.0 F1/EM (dev set)	RACE m/h accuracy (test set)
RoBERTa	2	90.2 / 90.2	92.2	94.6 / 88.9	89.4 / 86.5	83.2 (86.5 / 81.8)
ALBERT	3	90.8	92.2	94.8 / 89.3	90.2 / 87.4	86.5 (89.0 / 85.5)
XLNet	2	90.8 / 90.8	92.3	95.1 / 89.7	90.6 / 87.9	85.4 (88.6 / 84.0)
Megatron-336M	1	89.7 / 90.0	92.3	94.2 / 88.0	88.1 / 84.8	83.0 (86.9 / 81.5)
Megatron-1.3B	1	90.9 / 91.0	92.6	94.9 / 89.1	90.2 / 87.1	87.3 (90.4 / 86.1)
Megatron-3.9B	1	**91.4 / 91.4**	**92.7**	**95.5 / 90.0**	**91.2 / 88.5**	**89.5 (91.8 / 88.6)**

https://developer.nvidia.com/blog/language-modeling-using-megatron-a100-gpu/

DistilBERT

- Knowledge Distillation
  - Přístup učitel-student trénování
    - Student (menší model) se trénuje reprodukovat chování učitele (většího modelu)
    - Student se neučí na one-hot vektory tříd, ale na výstupní pravděpodobnosti učitele
- Model je menší a rychlejší o 60 % než BERT

Model	**Score**	CoLA	MNLI	MRPC	QNLI	QQP	RTE	SST-2	STS-B	WNLI
ELMo	68.7	44.1	68.6	76.6	71.1	86.2	53.4	91.5	70.4	56.3
BERT-base	79.5	56.3	86.7	88.6	91.8	89.6	69.3	92.7	89.0	53.5
DistilBERT	77.0	51.3	82.2	87.5	89.2	88.5	59.9	91.3	86.9	56.3


10000
7500
5000
2500
0
MegatronLM
8300
NVIDIA
OpenAI
GPT-2
1500
Transformer
ELMo
465
MT-DNN
330
UNIVERSITY of WASHINGTON
Grover-
Mega
1500
XLM 665
340
XLNET
Carnegie
Mellon
University
RoBERTa
355
DistilBERT
66
ELMo
94
OpenAI
GPT
110
Google AI
BERT-Large
340
April 2018
July 2018
October 2018
January 2019
April 2019
July 2019

TECHNICKÁ UNIVERZITA V LIBERCI
Fakulta mechatroniky, informatiky
a mezioborových studií

Část II.:
Aplikace jazykových modelů

GLUE Benchmark

- General Language Understanding Evaluation

- Kolekce datasetů pro trénování, vyhodnocení a analýzu NLP systémů
- Obsahuje 9 datasetů s různými úlohami
  - Různé velikosti datasetů
  - Různá složitost
- Poskytuje také žebříček nejlepších architektur (link)

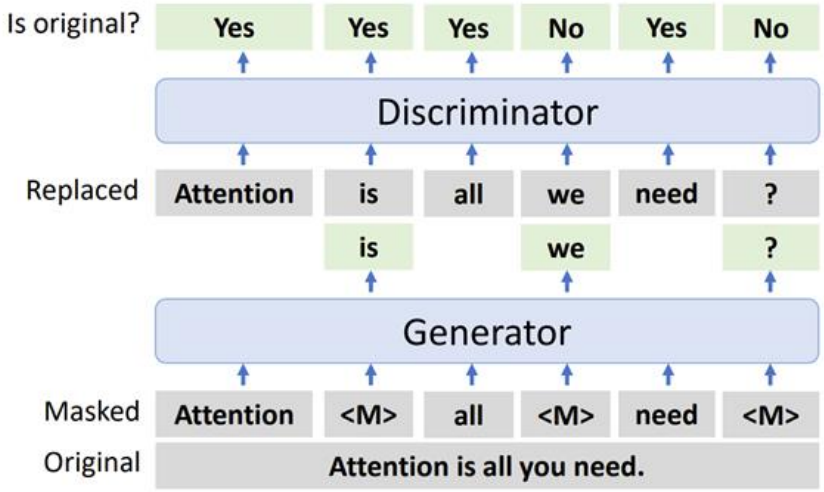
- Důvod vytvoření GLUE:
  - Chceme jazykové modely, které budou fungovat dobře na většině problémů a ne ty, které budou pouze vyladěny na jednu konkrétní úlohu

GLUE Benchmark

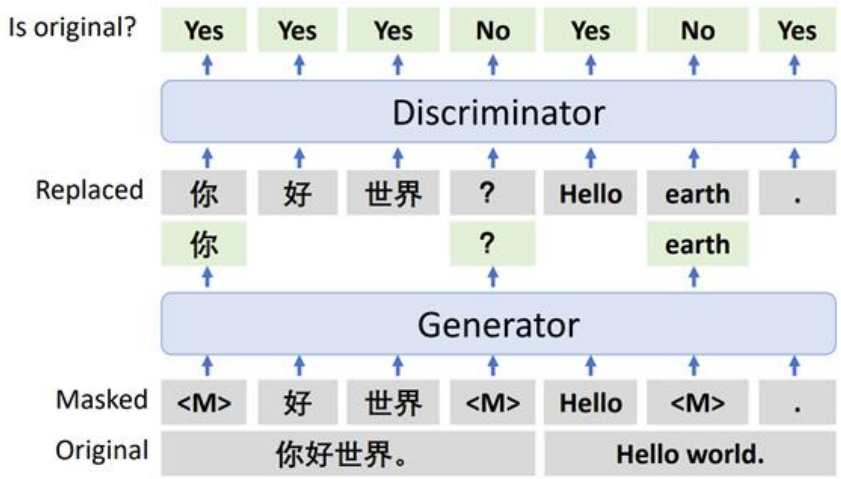
Dataset	Description	Data example	Metric
CoLA	Is the sentence grammatical or ungrammatical?	"This building is than that one." = **Ungrammatical**	Matthews
SST-2	Is the movie review positive, negative, or neutral?	"The movie is funny , smart , visually inventive , and most of all , alive ." = **.93056 (Very Positive)**	Accuracy
MRPC	Is the sentence B a paraphrase of sentence A?	A) "Yesterday , Taiwan reported 35 new infections , bringing the total number of cases to 418 ." B) "The island reported another 35 probable cases yesterday , taking its total to 418 ." = **A Paraphrase**	Accuracy / F1
STS-B	How similar are sentences A and B?	A) "Elephants are walking down a trail." B) "A herd of elephants are walking along a trail." = **4.6 (Very Similar)**	Pearson / Spearman
QQP	Are the two questions similar?	A) "How can I increase the speed of my internet connection while using a VPN?" B) "How can Internet speed be increased by hacking through DNS?" = **Not Similar**	Accuracy / F1
MNLI-mm	Does sentence A entail or contradict sentence B?	A) "Tourist Information offices can be very helpful." B) "Tourist Information offices are never of any help." = **Contradiction**	Accuracy
QNLI	Does sentence B contain the answer to the question in sentence A?	A) "What is essential for the mating of the elements that create radio waves?" B) "Antennas are required by any radio receiver or transmitter to couple its electrical connection to the electromagnetic field." = **Answerable**	Accuracy
RTE	Does sentence A entail sentence B?	A) "In 2003, Yunus brought the microcredit revolution to the streets of Bangladesh to support more than 50,000 beggars, whom the Grameen Bank respectfully calls Struggling Members." B) "Yunus supported more than 50,000 Struggling Members." = **Entailed**	Accuracy
WNLI	Sentence B replaces sentence A's ambiguous pronoun with one of the nouns - is this the correct noun?	A) "Lily spoke to Donna, breaking her concentration." B) "Lily spoke to Donna, breaking Lily's concentration." = **Incorrect Referent**	Accuracy

GLUE Benchmark

- Momentálně vede Microsoft s Turing universal language representation model (T-ULRv5)
  - Model inspirován ELECTRA přístupem (generátor, diskriminátor)



(a) Multilingual replaced token detection (MRTD)



(b) Translation replaced token detection (TRTD)

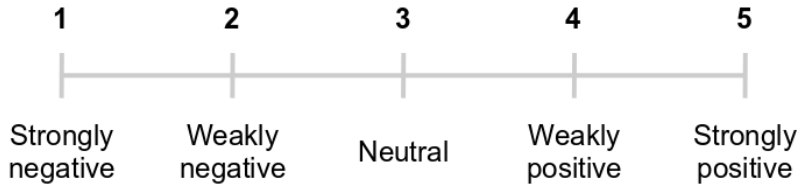
https://www.microsoft.com/en-us/research/blog/microsoft-turing-universal-language-representation-model-t-ulrv5-tops-xtreme-leaderboard-and-trains-100x-faster/

GLUE Benchmark

- Součást testovaných úloh
  - Gramatika věty
  - Analýza sentimentu
  - Parafráze textu
  - Podobnost textu
  - Má otázka odpověď?
  - Rozpor, důsledek (2 věty na vstupu)
  - Koreference (na co se „to“ odkazuje)

Analýza sentimentu

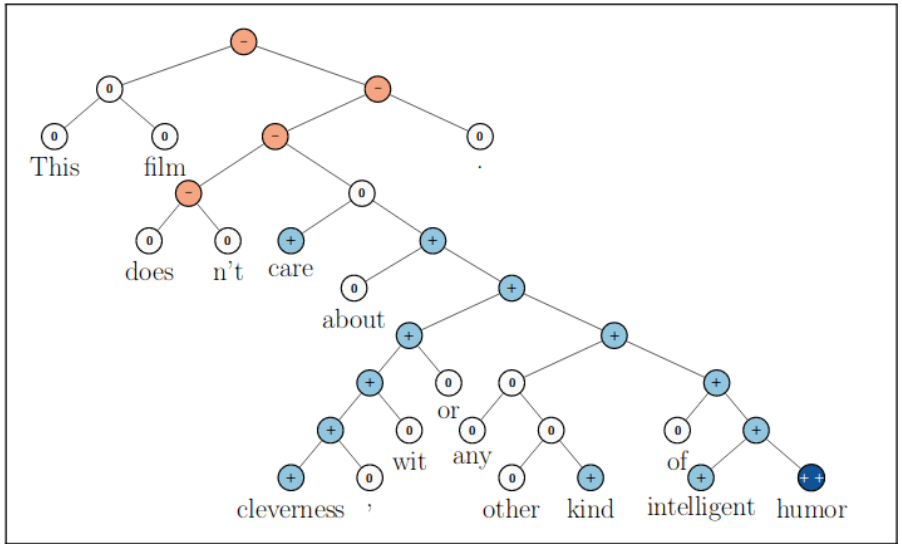
- Vytěžení a identifikace subjektivní informace
  - Pozitivní / negativní sentiment
  - Sentiment může být rozdělen i na více kategorií (složitější úloha)



- Přístupy:
  - Rule-based
  - Feature-based
  - **Embedding-based**

Analýza sentimentu

- Stanford Sentiment Treebank (SST-5)
  - 11 855 vět s hodnocení 1-5
  - Získáno z recenzí filmů
  - Labely ve formě stromové struktury
    - Z důvodu trénování jejich Recursive Neural Tensor Network (2015)



Analýza sentimentu + BERT

- Použití předtrénovaného BERT modelu
  - Možnost využití HuggingFace a transformers knihovny
- Přidání jedné plně propojené neuronové sítě (jedna nebo více vrstev)
  - Binární klasifikace pro sentiment 0/1
  - Klasifikace do tříd pro sentiment 1-5
- Natrénování přidaných vrstev na cílové aplikaci

Question Answering

- Automatické zodpovězení otázky na základě poskytnutého textu
  - Model obdrží otázku a zároveň text, ve kterém může odpověď vyhledat
  - Model může mít dostupnou databázi znalostí
  - Close-domain / Open-domain

Question Answering

- SQuAD (Stanford Question Answering Dataset)
  - Obsahuje otázky, odpovědi a texty Wikipedia článků
  - SQuAD 1.1
    - 100 000+ párů otázek a odpovědí na 500+ článků
  - SQuAD 2.0
    - SQuAD 1.1
    - Přidáno 50 000+ otázek, které nelze zodpovědět, ale vypadají podobně těm, které odpověď v textu mají

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Jun 04, 2021	IE-Net (ensemble) *RICOH_SRCB_DML*	**90.939**	**93.214**

Question Answering + BERT

- Složitější vytvoření predikce
  - Celý text se většinou nevejde na vstup modelu
    - Potřeba rozdělení textu na části, predikce, agregace

1. Na vstup modelu přivedeme otázku a text
2. Vytvoříme slovní embedding pro vstupní text (S×D)
3. Přidání jedné plně propojené neuronové sítě (jedna nebo více vrstev)
   - Výstupem jsou dva vektory o dimenzi S
     - Start vektor a End vektor s pravděpodobnostmi
     - Největší hodnoty znamenají největší pravděpodobnost začátku / konce odpovědi (na daném indexu textu)

S … délka textu, D … velikost slovního embeddingu

Užitečná literatura / kurzy

- GLUE benchmark
  - Papers with code – GLUE
- Články
  - RoBERTa
  - ALBERT
  - ELECTRA
  - Megatron
  - DistilBERT