ДОДАТОК А

Апробація результатів роботи

APPLICATION OF GATED UNITS TO BERT-BASED MODELS

Andrii Zaiev
Student of Computer Science Faculty
Kharkiv National University of Radioelectronics

Scientific advisor: Oleksii Turuta
Assoc. Prof., PhD, Department of Software Engineering
Kharkiv National University of Radioelectronics
Ukraine

Introduction. Recently models based on BERT (Bidirectional Encoder Representations from Transformers) [1] significantly improved the results for many natural language processing tasks, including joint intent classification and slot labeling [2]. Further developments in this area focus on improving models accuracy via using different approaches to BERT implementation (e.g. ALBERT[3] or DistilBERT[4]).

On the other hand, there is a set of existing methods for solving same tasks using recurrent neural networks (RNN). While they are mostly become overshadowed by more recent approaches, they can still provide valuable insights to BERT-based models' improvement if applied correctly. One of such methods is the usage of slot-gated units [5], which was state-of-the-art before BERT appeared.

The goal of intent classification task is to find an intent – class – for provided utterance, while for slot filling task labels should be assigned to the words from an utterance. Therefore, joint task focuses on solving both subtasks simultaneously for the same utterance. The most common applications of this to practical tasks are chat-bots and voice assistants.

When applied to joint intent classification and slot labeling task, gate unit allows to interconnect slot-labeling module with the output of intent classificatory. The

motivation for this is that slot labels may depend on user's intent and so it can be used as a feature for slot labels classifier [5].

We propose a modified version of slot-gated unit which can be used as a final layer of BERT-based model and demonstrate that the usage of this unit allows to improve model's performance on ATIS and SNIPS datasets.

Proposed model. Original slot-gated unit cannot be applied directly to an output of BERT as it operates with RNN-specific hidden states and separate attention values, while in Transformer (internal architecture of BERT) [6] attention values are already included in hidden state. Moreover, outputs of BERT operate in different dimensionality, so unit should use matrix calculus internally.

The proposed version of slot-gate unit is defined as:

$$g = \sum_{i=1}^{n} v \cdot \tanh(h_i + W_g h_0)$$
$$P_i(L) = softmax(h_i g W_L + b_L), i = 1, \dots, n \quad (1)$$

where:

h_0 – *output for [CLS] class token,*

h_i – *output for input token with index i,*

W_g, W_L – *learnable weight matrices,*

v – *learnable weight vector,*

g – *intent-slot relation coefficient,*

b_L – *learnable bias,*

$softmax$ – *matrix form of softmax function.*

This unit was incorporated in current state-of-the-art model for joint intent classification and slot labeling [2] as the replacement of linear softmax layer for slot labeling there. As that model experiments with conditional random field (CRF) usage we included them in our model too.

Experiments and analysis. We evaluated models' performance on standard benchmark dataset ATIS using following configurations:

- Joint BERT;
- Joint BERT + CRF;

– Joint BERT + SlotGate;

– Joint BERT + SlotGate + CRF.

The results (table 1) show that when used in combination with CRF classifier model with slot-gate unit overperforms existing models by overall sentence accuracy metric but is inferior on separate subtasks.

Table 1

Models performance comparison

Model	Intent accuracy	Slot F1	Sentence accuracy
Joint BERT	**0.978**	0.955	0.875
Joint BERT + CRF	0.973	**0.960**	0.878
Joint BERT + SlotGate	0.972	0.956	0.877
Joint BERT + SlotGate + CRF	0.974	0.956	**0.880**

[created by author]

Conclusion. We propose an application of a modified version of slot-gated unit to BERT-based model for solving joint intent classification and slot labeling tasks. Experiments show that such usage of slot-gated unit allows to improve overall score on a sentence level. On subtask level original Joint BERT models outperform it, so it is applicable only for joint tasks. Further work includes evaluating this technic on other datasets for joint tasks and with optimized BERT models – for example, ALBERT and DistilBERT.

References:

1. Devlin, J., Chang, M.-W., Lee, K. & Toutanova K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Volume 1 (Long and Short Papers). June, 2019, Minneapolis, United States of America.
2. Chen, Q., Zhuo, Z. & Wang, W. (2019). BERT for Joint Intent Classification and Slot Filling. *ArXiv*. Retrieved from https://arxiv.org/abs/1902.10909

3. Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P. & Soricut, R. (2019). ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. *Proceedings of the 2019 International Conference on Learning Representations (ICLR 2019)*. May, 2019. New Orleans, United States of America.
4. Sanh, V., Debut, L., Chaumond, J. & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *Proceedings of the 5th Workshop on Energy Efficient Machine Learning and Cognitive Computing - NeurIPS 2019*. December, 2019. Vancouver, Canada.
5. Goo, C.-W., Gao, G., Hsu, Y.-K., Huo, C.-L., Chen, T.-C., Hsu, K.-W. & Chen Y. (2018). *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Volume 2 (Short Papers). June, 2018. New Orleans, United States of America.
6. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L. & Polosukhin, I. (2017). Attention Is All You Need. *Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017)*. December, 2017. Long Beach, United States of America.

ДОДАТОК Б

Слайди презентації

ХНУРЕ
Кафедра ПІ
Атестаційна робота магістра

Дослідження методів аналізу природньої мови для заповнення форм в чат-ботах

ВИКОНАВ:
СТ. ГР. ІПЗМ-18-1 ЗАЄВ А.О.

КЕРІВНИК:
ДОЦ. КАФ. ПІ, К.Т.Н., ДОЦ. ТУРУТА О.П.

Інформація про дослідження

Мета дослідження – розробка алгоритму для отримання з тексту природною мовою даних, що необхідні для заповнення заданої форми.

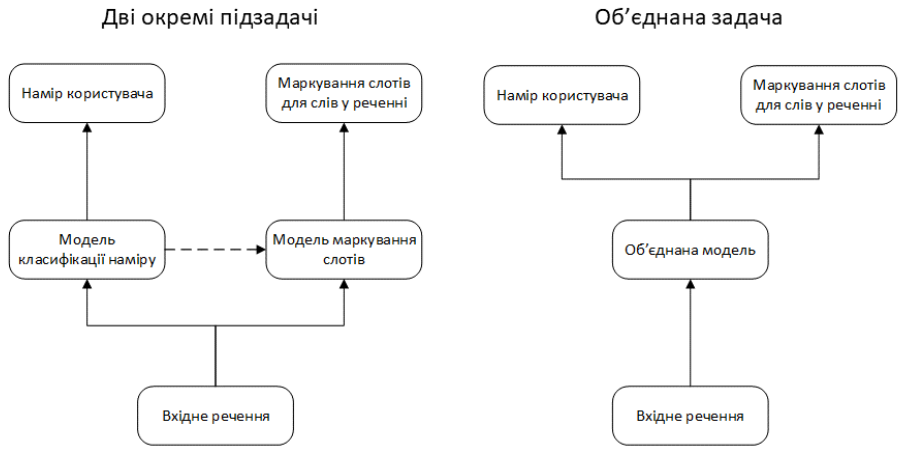
Об’єкт дослідження – методи аналізу природньої мови.

Предмет дослідження – методи заповнення форм даними з тексту природною мовою.

Постановка задачі

1. Дослідити існуючі підходи до задачі заповнення форм у застосуванні до чат-ботів. Вхідними даними задачі є повідомлення користувача природною мовою у довільному форматі, вихідними – тип форми, яку необхідно заповнити, та дані, якими вона має бути заповнена.
2. Розробити модель, що дозволила б розв'язувати задачу заповнення форм з більшою точністю ніж існуючі методи на заданих наборах даних.
3. Провести дослідження впливу внутрішніх параметрів моделі та компонентів, що в ній застосовуються, на метрики її роботи.

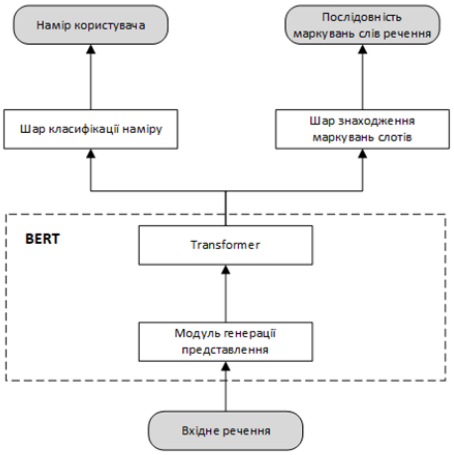
Задача заповнення форм



Реалізація моделі

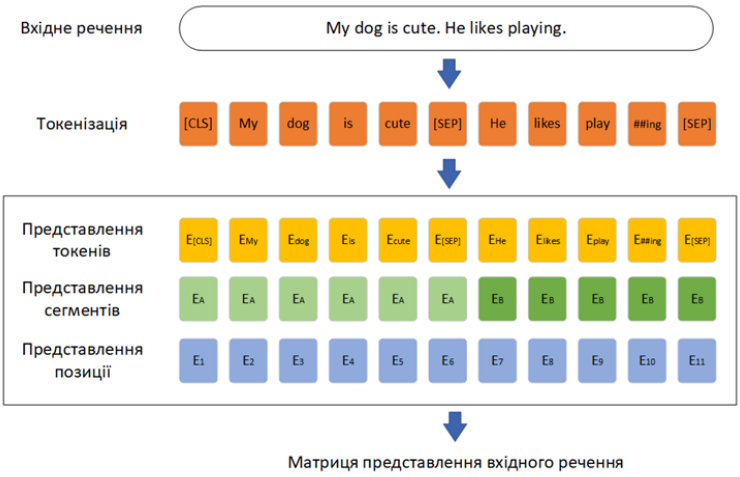
PyTorch

Transformers



5

Обробка вхідних даних



Засновано на Jacob Devlin et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

6

Шар маркувань слотів

Можливі конфігурації шару:

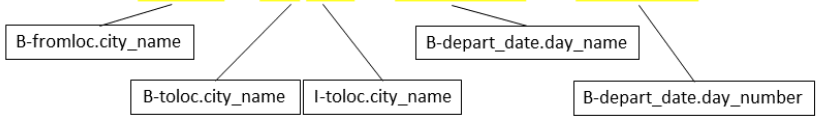
- Лінійний softmax шар – Baseline
- Лінійний CRF шар – CRF
- Softmax шар з вентильним вузлом – SlotGate
- CRF шар з вентильним вузлом – SlotGate+CRF

Набор даних ATIS

- Всього речень: 5871
- Кількість намірів: 21
- Кількість маркувань слотів: 120

Приклад:

<*atis_flight*> What are the flights from Tacoma to San Jose on Wednesday the nineteenth?



Метрики роботи різних архітектур

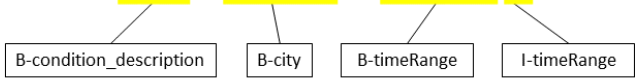
Набір даних	Архітектура	Точність наміру	Метрика F1 для слотів	Точність для речення	Час навчання, хв	Час передбачення, с
ATIS	Baseline	**0.978**	0.955	0.875	**1.835**	**6.120**
	CRF	0.973	**0.960**	0.878	2.320	10.986
	SlotGate	0.972	0.956	0.877	2.151	7.050
	SlotGate+CRF	0.974	0.956	**0.880**	2.618	11.203
SNIPS	Baseline	**0.984**	0.962	0.919	**4.423**	**4.720**
	CRF	**0.984**	0.959	0.911	7.272	8.851
	SlotGate	0.983	0.962	0.919	6.620	5.138
	SlotGate+CRF	**0.984**	**0.963**	**0.920**	7.513	9.437

Набор даних SNIPS

- Всього речень: 14,484
- Кількість намірів: 7
- Кількість маркувань слотів: 72

Приклад:

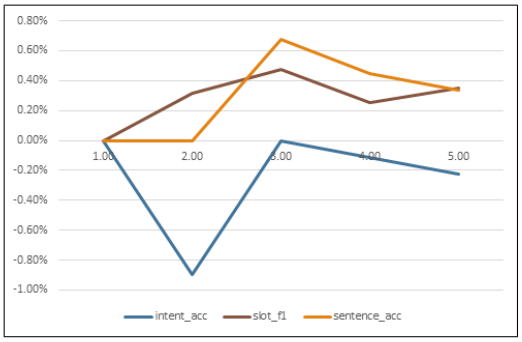
<*GetWeather*> Will it be humid in Beedeville on November 20?



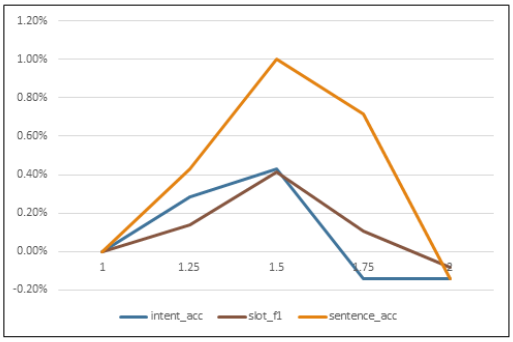
Налаштування моделі

$$Loss = \alpha \times Loss_I + \beta \times Loss_L,\ k = \frac{\beta}{\alpha}$$

ATIS



SNIPS



12

Метрики роботи різних моделей BERT

Набір даних	Базова модель	Точність наміру	Метрика F1 для слотів	Точність для речення	Час навчання, хв	Час передбачення, с	Фізичний розмір, МБ
ATIS	BERT	**0.978**	0.955	0.875	1.835	6.120	418.112
	ALBERT	0.976	**0.956**	**0.879**	2.384	7.013	655.370
	DistilBERT	0.973	0.951	0.865	**0.957**	**3.542**	**253.602**
SNIPS	BERT	**0.984**	0.962	0.919	4.423	4.720	420.400
	ALBERT	**0.984**	**0.965**	**0.922**	5.758	5.631	656.120
	DistilBERT	0.981	0.960	0.906	**2.571**	**2.504**	**253.535**

11

Висновки

В результаті даної роботи розроблено модель для вирішення задачі заповнення форм на основі тексту природньою мовою. Вона базується на нейромережевій моделі природньої мови BERT і дозволяє передбачувати намір користувача та параметри (слова або словосполучення), що відповідають цьому наміру.

Подальший розвиток даної роботи може здійснюватися у декількох напрямках:

- Адаптація моделі до використання з іншими мовами (у тому числі за рахунок багатомовної моделі BERT)
- Дослідження роботи моделі на більш складних наборах даних, у тому числі таких, що потребують підтримувати послідовність з декількох речень зі збереженням контексту
- Розширення базової задачі на випадки, коли дані форми необхідно додатково перетворювати у інший – машинний – формат (наприклад, для дати та часу)

ДОДАТОК В

Відгук

ХАРКІВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ РАДІОЕЛЕКТРОНІКИ
Факультет комп'ютерних наук

ВІДГУК

на атестаційну роботу магістра
Заєва Андрія Олександровича, ІПЗм-18-1
спеціальність 121 – Інженерія програмного забезпечення
освітньо-наукова програма «Інженерія програмного забезпечення»
Тема атестаційної роботи «Дослідження методів аналізу природньої мови для заповнення форм в чат-ботах»

Студент Заєв А.О. виконував атестаційну роботу магістра протягом двох років, досліджував методи аналізу природної мови та архітектуру текстових асистентів (чат-ботів).

В роботі Заєв А.О. самостійно виконав аналіз існуючих методів аналізу природної мови, оцінки іменованих сутностей, визначення ключової інформації, продемонстрував високий рівень підготовленості до самостійної роботи, використовував методи наукових досліджень, показав уміння користуватися науково-технічною літературою, ресурсами мережі Інтернет, виявив глибокі знання в області алгоритмізації та мов програмування. Робота виконана якісно, самостійно, під час дослідження та розробки студент показав знання та вміння використовувати сучасні інструменти машинного навчання.

В ході роботи було проаналізовано сучасні підходи як до аналізу природної мови у цілому, так і до вирішення конкретної задачі заповнення форм зокрема, розроблено удосконалену модель для розв’язання об’єднаної задачі класифікації наміру та маркування слотів на основі синтезу елементів існуючих моделей, які раніше сумісно не використовувалися.

Магістрант гр. ІПЗм-18-1 Заєв А.О. готовий до самостійної інженерної діяльності. Атестаційну роботу можна подати до захисту в ЕК за спеціальністю 121-«Інженерія програмного забезпечення», освітньо-науковою програмою «Інженерія програмного забезпечення».

«_____» _____ 2020 р.
Керівник атестаційної роботи магістра
Турута О.П., доцент каф. ПІ, к.т.н., доцент