Using Regression Models to pinpoint Relevant Content in Research Papers

Moreno La Quatra

5th SmartData@PoliTO Workshop

September 27th 2019

Outline

- Problem formulation
- Our approach
- Considerations and conclusions

The structure of a research paper

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

Jacob Devlin **Ming-Wei Chang** **Kenton Lee** **Kristina Toutanova**
Google AI Language
{jacobdevlin,mingweichang,kentonl,kristout}@google.com

Abstract

We introduce a new language representation model called **BERT**, which stands for **B**idirectional **E**ncoder **R**epresentations from **T**ransformers. Unlike recent language representation models (Peters et al., 2018a; Radford et al., 2018), BERT is designed to pre-train deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers. As a result, the pre-trained BERT model can be fine-tuned with just one additional output layer to create state-of-the-art models for a wide range of tasks, such as question answering and language inference, without substantial task-specific architecture modifications.

BERT is conceptually simple and empirically powerful. It obtains new state-of-the-art results on eleven natural language processing tasks, including pushing the GLUE score to 80.5% (7.7% point absolute improvement), MultiNLI accuracy to 86.7% (4.6% absolute improvement), SQuAD v1.1 question answering Test F1 to 93.2 (1.5 point absolute improvement) and SQuAD v2.0 Test F1 to 83.1 (5.1 point absolute improvement).

1 Introduction

Language model pre-training has been shown to be effective for improving many natural language processing tasks (Dai and Le, 2015; Peters et al., 2018a; Radford et al., 2018; Howard and Ruder, 2018). These include sentence-level tasks such as natural language inference (Bowman et al., 2015; Williams et al., 2018) and paraphrasing (Dolan and Brockett, 2005), which aim to predict the relationships between sentences by analyzing them holistically, as well as token-level tasks such as named entity recognition and question answering, where models are required to produce fine-grained output at the token level (Tjong Kim Sang and De Meulder, 2003; Rajpurkar et al., 2016).

There are two existing strategies for applying pre-trained language representations to downstream tasks: *feature-based* and *fine-tuning*. The feature-based approach, such as ELMo (Peters et al., 2018a), uses task-specific architectures that include the pre-trained representations as additional features. The fine-tuning approach, such as the Generative Pre-trained Transformer (OpenAI GPT) (Radford et al., 2018), introduces minimal task-specific parameters, and is trained on the downstream tasks by simply fine-tuning *all* pre-trained parameters. The two approaches share the same objective function during pre-training, where they use unidirectional language models to learn general language representations.

We argue that current techniques restrict the power of the pre-trained representations, especially for the fine-tuning approaches. The major limitation is that standard language models are unidirectional, and this limits the choice of architectures that can be used during pre-training. For example, in OpenAI GPT, the authors use a left-to-right architecture, where every token can only attend to previous tokens in the self-attention layers of the Transformer (Vaswani et al., 2017). Such restrictions are sub-optimal for sentence-level tasks, and could be very harmful when applying fine-tuning based approaches to token-level tasks such as question answering, where it is crucial to incorporate context from both directions.

In this paper, we improve the fine-tuning based approaches by proposing BERT: **B**idirectional **E**ncoder **R**epresentations from **T**ransformers. BERT alleviates the previously mentioned unidirectionality constraint by using a "masked language model" (MLM) pre-training objective, inspired by the Cloze task (Taylor, 1953). The masked language model randomly masks some of the tokens from the input, and the objective is to predict the original vocabulary id of the masked

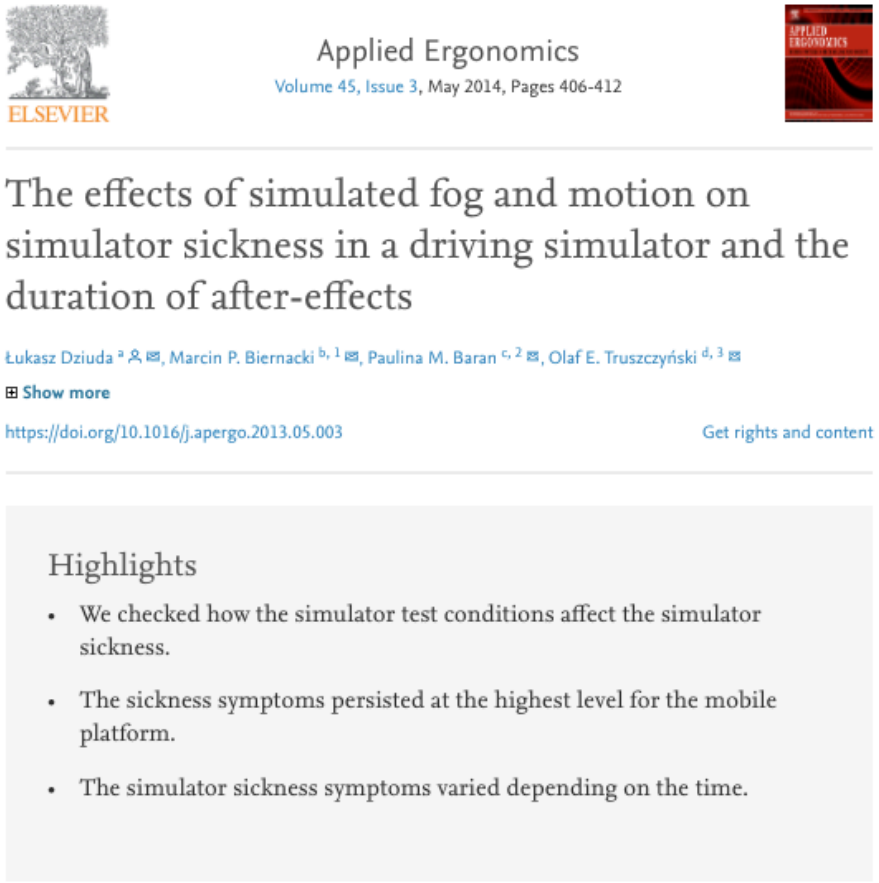
Title

Authors' information

Abstract: short summary of the paper

The full text of the research paper (divided in sections)

And highlights…



ELSEVIER

Applied Ergonomics

Volume 45, Issue 3, May 2014, Pages 406-412

The effects of simulated fog and motion on simulator sickness in a driving simulator and the duration of after-effects

Łukasz Dziuda [a], Marcin P. Biernacki [b, 1], Paulina M. Baran [c, 2], Olaf E. Truszczyński [d, 3]

Show more

https://doi.org/10.1016/j.apergo.2013.05.003 Get rights and content

Highlights

- We checked how the simulator test conditions affect the simulator sickness.
- The sickness symptoms persisted at the highest level for the mobile platform.
- The simulator sickness symptoms varied depending on the time.

- "Highlights are three to five **result-oriented** points."
- "They provide readers with an at-a-glance overview of the **main findings** of your article."
- "Think of them as a quick snippet of the results—short and sweet."

Source: https://www.elsevier.com/authors/journal-authors/highlights

Abstract vs Highlights

Abstract	Highlights
It summarizes the **full content** of the paper.	They summarize the main findings of the paper, with **focus on the results**.
It is **self-contained**.	They are in the **bullet point format**.
It is a **general purpose** summary of the paper.	They are created with the purpose of highlighting **specific part** of the paper.

We need of a **supervised algorithm** rather than a general-purpose summarizer.

Highlights

The paper highlights are manually provided by the authors. Editors start asking for highlights almost a decade ago.

- A large number of publications **does not have highlighted sentences**.
- Authors could receive **suggestion** to write their own highlights.

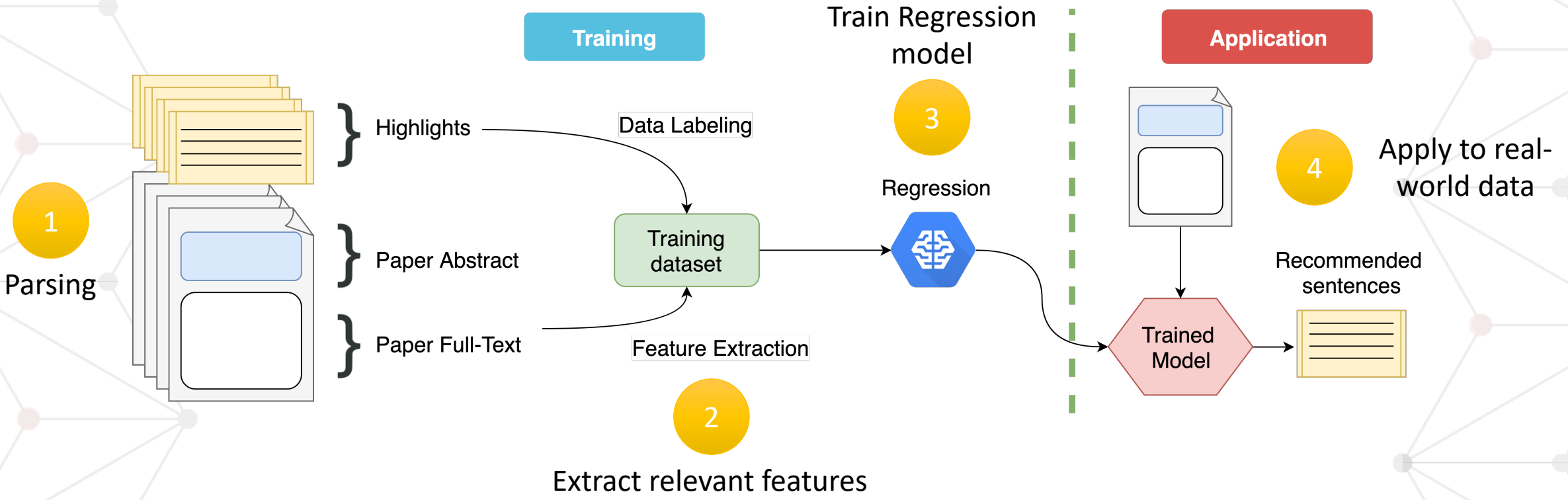
So… We aim at proposing a novel technique for the automatic extraction of highlights in research publications.

Our approach

The requirements

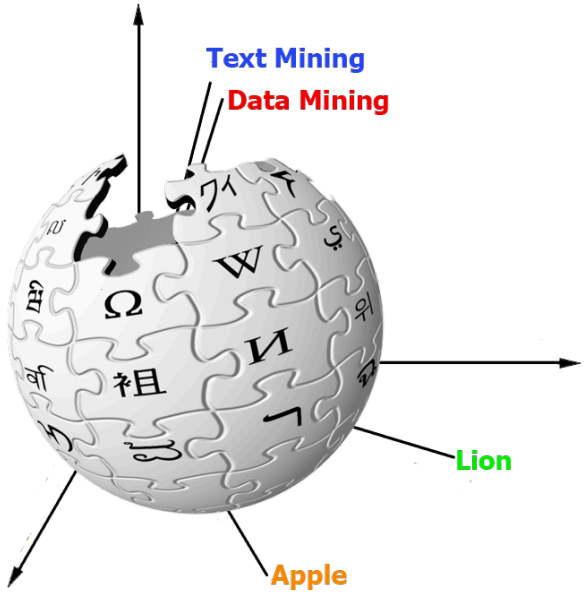
- We need to **identify** relevant sentences in the paper.
- They should be **targeted** at providing a quick overview of the main findings and the results obtained.
- We need a method able to give a **rank** such to propose 1-to-n sentences as highlights for the publication.

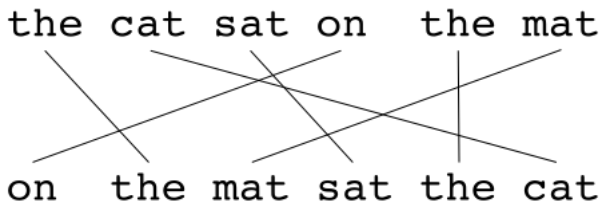
Method's overview



Features

Type	Name
Semantic	Sentence relevance in latent spaces (W2V-, S2V-PR), Sentence similarity with the abstract (Rouge, W2V, S2V).
Syntactic	Symbols count, POS-Tagging, Frequency-based word relevance.
Structural	Sentence position in the paper.





Real example

Example	
Ground-Truth	We checked how the simulator test conditions affect the simulator sickness.
	The sickness symptoms persisted at the highest level for the mobile platform.
	The simulator sickness symptoms varied depending on the time.
Our System	How does the severity of simulator sickness symptoms change over time?
	The analysis of results showed that the severity of simulator sickness symptoms was higher when using a fixed base simulator compared to the motion base one
	It turns out that the severity of most symptoms of simulator sickness is impacted upon both by the simulator task conditions and the time that has elapsed since performing the task in the simulator.

Considerations

- The algorithm extracts sentences **well-aligned to real-highlights**.
- If compared with the previous state of the art, it performs better and is capable to propose a **rank** of the sentences.
- Future works can integrate latest **deep learning methods** to exploit higher-level semantic connection between sentences.

Thank you!

Moreno La Quatra

5th SmartData@PoliTO Workshop

September 27th 2019