Friday Meeting on Zoom

NLP最新の動向①
～言語モデル～

北大　言語メディア学研究室
M2 白藤大幹

Language Model (言語モデル)の発展

- 言語モデル（LM）の発展が最近は目覚ましい
- BERTの登場（2018/10）以降、最新の研究論文では数多くのタスクでBERTがSoTAを達成。今ではBERTと比較しないってどうよ…って感じになっている。
- LM: 大容量データで事前学習されたモデルを各種タスクにfine-tuneして、色々なタスクに適応できるようになっている（**転移学習**）。
- メリット：小さなデータでも十分に学習できるようになった！

Language Model (言語モデル)の発展

- 有名なLM: BM25、BERT（2018/10）、XLNET（2019/6）
- ⬇⬇⬇では結局、言語モデルって何？⬇⬇⬇
- 単語列の生成確率を付与するモデル（要は、人が用いるであろう言葉らしさを確率としてモデル化したもの）
- 例）N-gramモデル（結構古いモデルだけど、比較対象として使われることがある）
  [w1, w2,,, wn]という単語列のi番目の単語wiの生起確率P(wi)は直前のN-1単語に依存しているという仮説に基づいたモデル

T5

by Google

(Published in Oct 2019)

LMの最新手法①
T5 (Text-to-Text Transfer Transformer)

- 2019/10にGoogleから提案されたモデル。
- NLPにおいてほぼ全てのタスクが「入力も出力もテキスト」という形に落とし込めるという仮説のもと、モデルを作成。

- モデル: T5
- データ：C4 (Colossal Clean Crawled Corpus)

LMの最新手法①
T5 (Text-to-Text Transfer Transformer)

フレームワーク（T5）

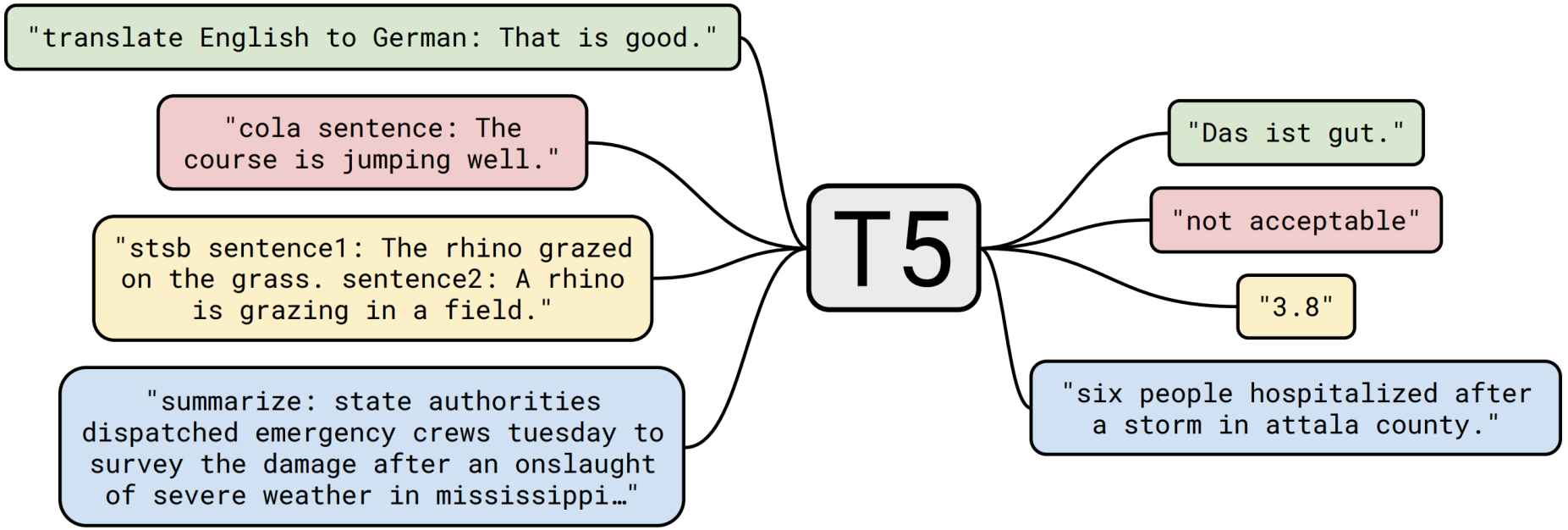
- **入力と出力が常にテキスト**であるtext-to-text統合フォーマットへの再構成を提案
- 機械翻訳、文書要約、質問回答、分類タスク（感情分析など）などのタスクで同一のモデル、損失関数、ハイパーパラメーターを使用できる。

大規模な事前学習データセット（C4）

- 高品質・多様・大規模なラベルなしデータセット。Common Crawlは大規模だけど品質が低く、Wikipediaは高品質だがスタイルが同じ。
  → そこで作られたのがC4！**Wikipediaの２倍のデータ量＋重複排除、不完全な文の破棄などのクリーニング作業**
  → **事前学習中に過学習することなくモデルサイズを大きくすることができた！**

LMの最新手法①
T5 (Text-to-Text Transfer Transformer)

以下のように、翻訳や質問回答などに活用できる。
回帰タスクに適用して、**数値自体ではなく数値の文字列表現を予測するように学習**することもできる。



LMの最新手法①
T5 (Text-to-Text Transfer Transformer)

転移学習におけるモデル・事前学習方法論など

- 一般に言語モデルにおいて、encoder-decoder modelよりもdecoder modelの方が劣っていることが筆者らのサーベイにより判明。
- 事前学習の目標は「空白入力スタイルのノイズ除去」が最も効果的。
- ラベルなしである理由は、タスク依存の小さなデータセットで事前学習する事は有害な過学習に繋がる可能性があるため。

LMの最新手法①
T5 (Text-to-Text Transfer Transformer)

良いところ

- 回答時に外部知識を参照できない質問回答問題で優れた性能を発揮！

- 今までの生成モデルであるGPT-2は・・・「継続タスク」
  すなわち、人間が書いた文章に続く文を予測するモデル。これに対して、T5は「空欄入力タスク」（継続タスクを一般化させたタスクとも捉えられる）。

今後の展望

- 破損したテキストのノイズ除去を目標にモデルを事前学習していた。しかし、これは最も効率的な方法では無いのではないか、と筆者らは考えている。

T-NLG

by Microsoft

(not published yet)

LMの最新手法②

T-NLG (Turing-Natural Language Generation)

- 2020年3月にマイクロソフトから発表された言語モデル。言語モデルにおける学習の最適化の必要性・モデルの大きさの重要性を示した。
- 170億のパラメーター（GPT-2の2倍以上の数！）。
- 質問回答タスクや要約タスクにおいて数々のSoTAを達成。

	LAMBADA (acc) strict	WikiText-103 (test adj. ppl)
Open AI GPT-2 1.5B	52.66 (63.24)*	17.48
Megatron-LM 8.3B	66.51	10.81
T-NLG 17B	67.98	10.21

*Open AI used additional processing (stopword filtering) to achieve higher numbers than the model achieved alone. Neither Megatron nor T-NLG use this stopword filtering technique.

Figure 1 below shows how T-NLG performs when compared with Megatron-LM on validation perplexity.

LMの最新手法②
T-NLG (Turing-Natural Language Generation)

Turing Natural Language Generation (T-NLG) is a 17 billion parameter language model by Microsoft that outperforms the state of the art on many downstream NLP tasks. We present a demo of the model, including its freeform generation, question answering, and summarization capabilities, to academics for feedback and research purposes.

- 上の文章はT-NLGがその解説記事を要約した文章。とてもシステムが生成したとは思えない🙃

  解説記事のURL: https://www.microsoft.com/en-us/research/blog/turing-nlg-a-17-billion-parameter-language-model-by-microsoft/

LMの最新手法②

T-NLG (Turing-Natural Language Generation)

- より自然な自然言語生成のためには、より大きなモデルが必要という昨今の情勢を受けて、170億個のパラメータを持つ言語モデルを作成した。

- 活用されたもの：DeepSpeed（最適化手法）、ZeRO（並列化オプティマイザ）

- 基本的なモデル：Transformerベース

- 既存の要約手法や質問回答システム：既存のコンテンツを文書から抽出して、それを代替の回答や要約として利用
  → これでは支離滅裂な文章が出来上がっていた【問題点】
  → **T-NLGを用いると、これが解決できる！！！**

LMの最新手法②
T-NLG (Turing-Natural Language Generation)

- モデル並列度を、ZeROを搭載したDeepSpeedにより、16から4にまで削減し、バッチサイズを４倍に増やせた。【結果】計算量を1/3に削減！

- T-NLGの詳細説明
  78 Transformer Layer with a hidden size of 4256 and 28 attention heads.
  Generation loss for 300,000 steps of batch size 512 on sequences of 1024 tokens.

LMの最新手法②
T-NLG (Turing-Natural Language Generation)

質問回答タスク

- zero shot question：文脈のない質問
  → 事前学習で得た知識を用いて回答を生成する。

When did WW2 end?	WW2 ended in 1945.
How many people live in the US?	There are over 300 million people living in the US.

- direct question answering：
  多くの質問において、文脈の中に質問の回答となる部分がある。目標は回答（ユーザが得たい情報の提供）をなすことである。

Question	Who was Jason Mraz engaged to?
Passage	Mraz was engaged to singer/songwriter and long-time close friend **Tristan Prettyman** on Christmas Eve 2010; they broke off the engagement six months later.
“Direct” Answer	Jason Mraz was engaged to Tristan Prettyman.

LMの最新手法②
T-NLG (Turing-Natural Language Generation)

文章要約タスク

- 文章要約の種類：
  ① 要約でなく文書から少数の文章を抽出する「抽出型」
  ② NLGモデルを用いて要約を生成する「抽象型」

- T-NLGの目標：
  既存の文をコピーするのではなく、電子メール、ブログ記事、Excelなどの幅広いテキスト文書に対して、人間のような抽象的な要約を書く！

- 上の目標の課題：人間は全てのデータに対して要約をするわけではないので、教師データが欠如している！
  → **T-NLGでは教師データを比較的使わずに要約することができる**

LMの最新手法②
T-NLG (Turing-Natural Language Generation)

文章要約タスク

- T-NLGを各タスクにおいてfine-tuneした後、ROUGE-Lの評価基準を用いて、今までの手法と比較。
  ※生成された要約がオリジナルのuni-gramやbigramとどれだけ正確に一致しているかを示す。

- なお、PEGASUSとは2019/12に発表された要約特化型の言語モデル。

Rouge-L	T-NLG	Previous SOTA	PEGASUS Large
arXiv	30.8	23.6	25.7
WikiHow	31.5	26.5	33.4
BillSum	41.6	33.7	45.8
WikiSum	36.0	38.8	-

LMの最新手法②

T-NLG (Turing-Natural Language Generation)

文章要約タスク

1つ目

https://blogs.microsoft.com/blog/2020/01/16/microsoft-will-be-carbon-negative-by-2030/

2つ目

https://www.newyorker.com/magazine/2019/10/14/can-a-machine-learn-to-write-for-the-new-yorker

Link to original text	T-NLG Summary
"Microsoft will be carbon negative by 2030" by Brad Smith, Official Microsoft Blog"	Microsoft is committed to being carbon negative by 2030. We are launching an aggressive program to cut our carbon emissions by more than half by 2030, both for our direct emissions and for our entire supply and value chain. We are also launching an initiative to use Microsoft technology to help our suppliers and customers reduce their own carbon footprints and a new $1 billion climate innovation fund to accelerate the development of carbon reduction, capture, and removal technologies that will help us and the world become carbon negative. In addition to our aggressive carbon goals, we are launching a new Climate Innovation Fund to accelerate carbon reduction and removal opportunities. We are also launching a program to use our technology to improve the efficiency of our supply chain and reduce our own carbon footprint as well...
"The Next Word: Where will predictive text take us?" by John Seabrook, *The New Yorker*	"Smart" compose is a feature that suggests what going to type next, based on your past e-mails and your current thoughts. But what happens when you let the A.I. decide what you want to say next?

LMの最新手法②
T-NLG (Turing-Natural Language Generation)

【まとめ】

- 今の所、文章要約と質問回答しか例がないので、他のタスクでどれ程の精度を叩きだせるのかは不明。

- 筆者らは、今後は画像処理と合わせてマルチタスクにしていく模様。

【所感】

- まだ論文が出ていないので、詳細なモデル概説がなされていなかったことが残念である。arxiveに投稿されたらチェックしなくてはいけない。

参考URL: https://www.microsoft.com/en-us/research/blog/turing-nlg-a-17-billion-parameter-language-model-by-microsoft/

以上、研究の参考になれば幸いです。

次回はPEGASUSやELECTRAを紹介したい😀