Enhancing BiDAF with BERT Embeddings, and Exploring Real-World Data

Priyanka Sekhar, Sumit Minocha, Ryan Mui
Department of Computer Science, Stanford University

Stanford | ENGINEERING
Computer Science

Motivation and Problem

Problem / Questions

- What effect on performance (SQuAD 2.0 leaderboard) does combining BERT embeddings with the BiDAF architecture have?
- How well do SOTA NLP architectures translate to industry and real-world settings?
- What are some of the industry pain points that we might run into while tuning/evaluating our performance?
- How do you qualitatively interpret the results (risks and strengths of automation)?

Datasets

Q: *How much does Benetint cost?*
A: *Hi Ammara! Our benetint cheek & lip stain is $30 USD! xo* https://www.benefitcosmetics.com/us/en/product/benetint-cheek-lip-stain?gclid=EAIaIQobChMIr8bspIe84AIVB9lkCh2RZAEYEAAYASAAEgIr6vD_BwE

Q: *Can you recommend something for me*
A: *No Answer*

Q: *Hello, I would like to know if the Lizy P eyebrow kit is available in Bolivia?*
A: *Hi gorgeous! Unfortunately it is not, but thank you for reaching out. To find the closest Benefit boutique location near you, please email us at customercare@benefitcosmetics.com! xo*
Q: *Do you have branches in Jordan?*
A: *Hi there! Unfortunately we no longer ship internationally but please email us at customercare@benefitcosmetics.com to find the Benefit location closest to you! xo*

Fig 1: Examples of Benefit DM interactions.

Our various BERT + BiDAF models (see Approach section) were trained on 2 datasets. 1) default project SQuAD 2.0 dataset. 2) a custom Benefit Cosmetics dataset (global cosmetics manufacturer).

Approach

Methods

- **BERT Embeddings**: BERT Base model consists of 12 encoder layers (transformer blocks). Extract and combine output from each of these blocks in different ways to construct word embeddings that perform best on the QA specific task at hand.[1]
- **BiDAF**: Modify BiDAF baseline[2] as shown in Figure 3, to intake the above mentioned BERT embeddings. Result is a BERT + BiDAF architecture.

Experimental Setup

- **Industry:** Heuristics used to turn raw csv files into QCA triplets, generate new dev / train sets and run BERT + BiDAF.
- **Transformer Block Variations:** examine 3 high performing combinations of embedding vectors generated from from the pre-trained Transformer's hidden layers. Extract 1) just the 1st layer, 2) just the last layer, and 3) sum of last 4 layers. Feed embeddings as input into our modified BiDAF model.
  - *4 experiments run:* Vanilla Baseline Submission, BERT Fine Tune, BiDAF with BERT, Industry
  - 2 epochs, lr=3e-5, doc_stride=128, seq_len = 384, b_s=6

Process

BERT Embeddings

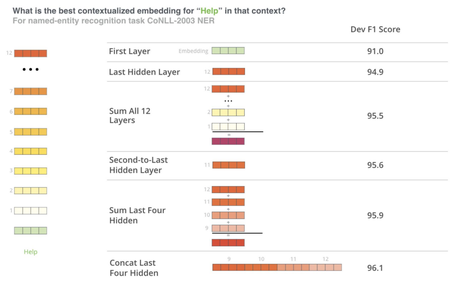

Fig 2: The different word embedding vectors, generated by concatenating different combinations of the pre-trained Transformer's hidden layer outputs

QA Architectures

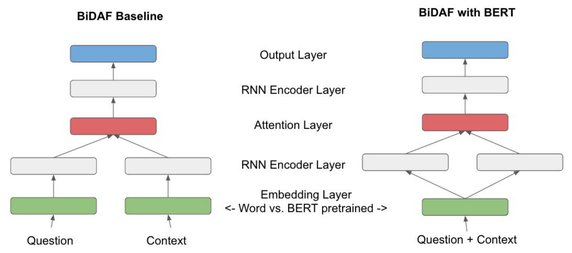

Fig 3: BiDAF architecture modification using BERT embeddings. Provided baseline takes as input a Q and C (which it converts to embeddings). Instead, take the combined Q and C and convert to BERT embeddings before passing them onto subsequent BiDAF layers.

Results

Summary

Table 1: Embedding Experiments

Model	EM	F1
BERT Fine Tune Reported [9]	82.126	84.820
BERT Fine Tune Empirical	73.001	76.069
Baseline BiDAF Reported	55	58
Baseline BiDAF Empirical	57.906	61.493
Last Layer 2 epochs	53.241	53.870
First Layer 2 epochs	51.629	51.686
Sum Last Four Layers 2 epochs	52.929	54.895

Takeaways:

- Achieved comparable EM/F1 to vanilla baseline within 2 epochs, versus 30 (baseline), but requires more training.
- Embedding experiment with best EM score = last layer. Embedding experiment with best F1 score = sum of last 4.
- Fine Tuned BERT on industry + SQuAD 2.0 mixed datasets achieved ~70/75 EM/F1).
- However, BERT + BiDAF on the mixed industry/SQuAD train/dev sets, yielded EM/F1 less than 10

Example Model Predictions

Question	Prediction	Expected
In what country is Normandy	"France"	France
When was Scotland invaded by William?	"1072"	1072
What major conquest did Tancred play a roll in?	""	Jerusalem

Fig 4: BERT + BiDAF experiments did well with proper nouns (names, locations, etc.) and numerical answers (i.e. dates) but both BERT + BiDAF and baseline models erred on the side of "no answer" for more complex questions.

Q: "do i have to pay in dollars? im from the philippines. where are benefit cosmetics available?"

A: "hi... We are Turkey based company.we have all natural product range. we have a wide range of handmade natural soaps, olive oil based beauty and bodycare soaps, massage oils,massage creams,body lotions,body butter,haircare products, clay masks, world- known Turkish rose products and many other beauty products.Based on the quantity we can do private label production. We export to 10 countries. We look for new markets and new customers. Regards, http://www.hamamtamam.com/

Fig 5: Data we received was in the form of csv files and incompatible with models. We had to come up with heuristics to as accurately as possible construct QCA triplets. These heuristics performed relatively well but there were clearly preprocessing errors. This example, 2 Qs asked, answer ignores first.

Q: "hi.. for how much u r selling the holiday set?" A: "$72"

Q: "What is the price of the POREfessional value size kit?" A: "$54"

Fig 6: BERT+BiDAF model on industry data did see relative success on user questions related to price as well as shipping availability (product complaint related inquiries usually led to a "no answer").

Analysis, Challenges

Industry Qualitative

- 2 major challenges, can be attributed to decisions made in the preprocessing pipeline as well as in the nature of the data.
- Data was over-conversational (slang, etc.)
- Abundant misspellings lead to <unk> characters being very likely in default model. Character / sub-word embeddings implemented often led to non-sensical subword tokenizations that hindered model performance.
- Success on price and shipping availability (15% of Qs), so potentially lucrative solution for corporations.

Other Qualitative

- unforeseen challenges in making the pre-trained embeddings compatible with a representative model such as BiDAF
- tokenizer (mismatched y1s, y2s), embedding size limitation b/c of memory

Conclusion

Model-Specific

- BERT embeddings are powerful but not quite 'plug and play'. Potentially more effective with more standard processing pipelines and memory optimizations.

Industry

- Promising results for a narrow band of questions, but would have greatly improved scores with more structured labelling, preprocessing, and a clear bank of answers.

Sources

[1] Jacob Devlin et al. 2018. *BERT: Pre-training of deep bidirectional transformers for language understanding.*
[2] Minjoon Seo et al. 2018. *Bidirectional attention flow for machine comprehension.*