Answer real questions using Wikipedia content

Axel Didier, Frédéric Assmus, Pierre Goncalves

UNIVERSITÉ DE LORRAINE

Institut des sciences du Digital
Management & Cognition

Core topic

- Kaggle competition : Give to questions one short and one long answer coming from Wikipedia
- Input : Questions linked with Wikipedia articles on which they are based on
- Actual goal : Fine-tuning of BERT model to have the best score possible at answering

Our actual subject

- Work on a **more practical**, user-oriented subject
- Base subject : Was more research oriented, we already knew the articles/excerpt in which to search for an answer
- Our subject : Having **only the question** given by the user and **finding a relevant excerpt** to give to BERT along with the question.

Applications ? Uses ?

- Answer naive factual questions without being aware of the context
- Google Search & Bing Search

What we'll have to do + tools

- Process the user's question : **NLTK, SpaCy**
  ↪️ Extract the subject and other relevant information : **NLTK, SpaCy**
- Find the most relevant Wikipedia article based on this information : **Wikipedia API, RegEx**
- Slice the article and extract the most relevant excerpt : **Wikipedia API, RegEx**
- Use the excerpt in pair with the initial question as input of a BERT model and see the results : **BERT, metrics**
- Find ways to improve these results

Current state of our project

- We took in hand a BERT fine-tuned model on SQuAD
- First tries with subject processing with NLTK
- First tries with article extraction using Wikipedia API
- Already get answer
- Global checking of the possibilities for the different steps

Our plans

- Next step : Try SpaCy instead of NLTK for question processing
- Going further in each step in an organised way
- Find a way to evaluate the quality of the answer
- If everything goes well - add functionalities
    - long answer with highlight on the short answer
    - Webpage or Jupyter Widgets

Questions ?

- Answers