Transformers in Action

Applying transformer models to understand customer interactions

Gil Noh
Machine Learning Engineer
OMQ GmbH

Short introduction on OMQ

- OMQ: providing intelligent customer service software.
- Customer support is everywhere.
  - eCommerce, public transports, universities, utility companies, service providers …
- The common problem: customer support agents are answering the same issues again and again.
- Our approach: apply AI technology, automatically answer the repeating issues. ⇒ help both support-agents and end-customers.
- Vision: “Answer once”
  - “You answer the same question only once, the machine will follow up all the last.”

Kontaktieren Sie unseren Kundenservice

Bei Fragen und Anregungen hilft Ihnen unser Team vom Kundenservice gerne weiter. Bei den meisten Anliegen kann Ihnen unser interaktiver Self-Service aber schon im Handumdrehen Antworten liefern. Probieren Sie es doch mal aus!

Hallo, ich habe einen Platten

Wie melde ich einen Schaden am Fahrrad?

Self service software by OMQ

Keine Antwort gefunden? Schicken Sie Ihre Nachricht direkt an den Kundenservice:

Thema

Bitte wählen Sie ein Thema.

Name

Kundennummer

E-Mail*

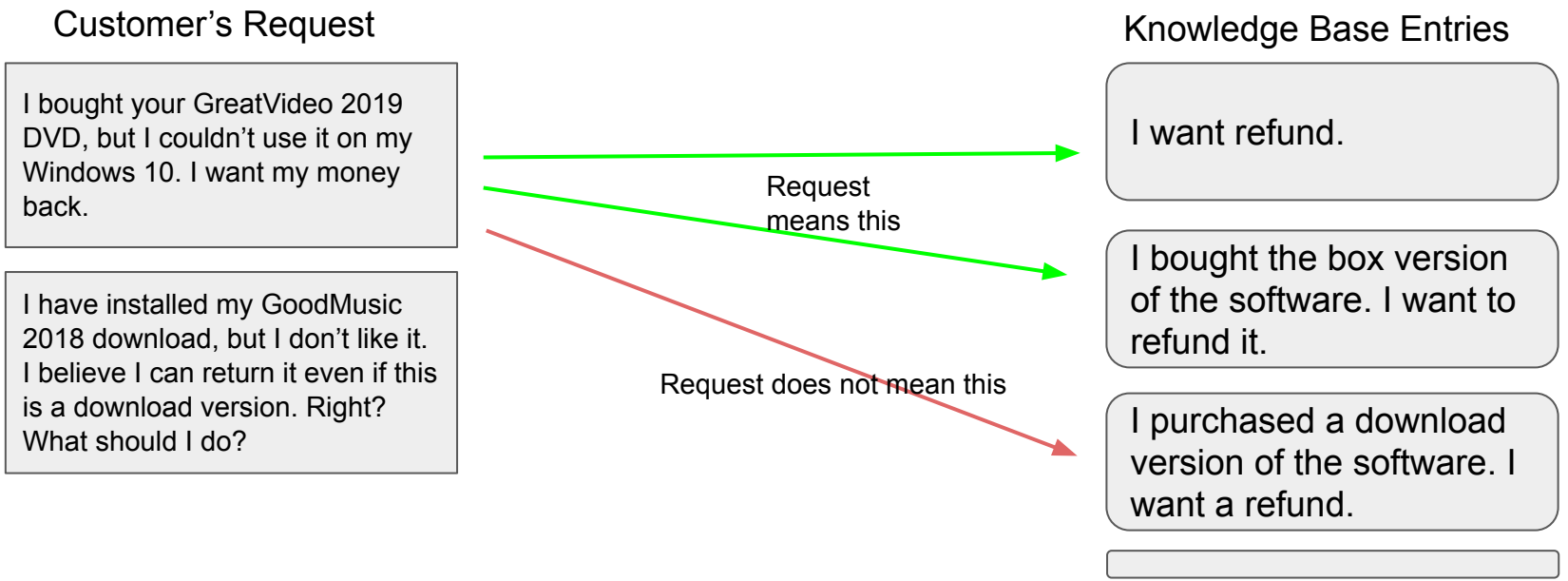
Telefon

Central knowledge base

- **Search:** *Versandkosten Ausland*
- **Question:** *Kann ich es mir ins Ausland liefern lassen? Wie hoch sind die Versandkosten.*
- **Email**: *Hallo Supportteam,*
  vielen Dank für die Antwort zu meiner Versandfrage.
  Ich habe aber noch eine Frage zu der Bezahlung
  …
  Support schrieb am 4.05. …
  > Hallo Mark,

Natural Language Understanding (NLU) task

- Is the given request text ***means*** this knowledge base case?
- Text-to-Text decision



Multiple representations

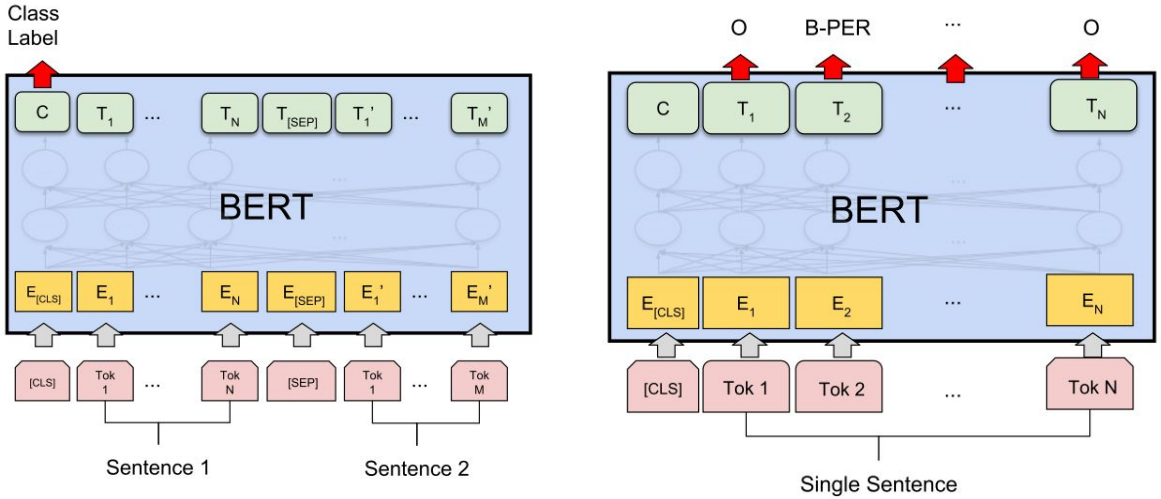
	Classical search (IR) model	**Shallow NN model: WordVectors + Composition**	**Deep NN model: Transformer-based**
Training method	(no training)	- Pre-trained sub-word vectors. - Fine-tuning of joint model on domain data self & supervised.	- Pre-trained transformers. - Fine-tuned on domain data, self-supervised & supervised.
Representation	- Bag of (lemmatized) tokens	- A vector in concept space. - Represents the topic of the sentence.	- A sequence of vectors. - Holds full syntactic and semantic information of the text.
Decision function	Classical IR ranking score	Similarity metric over two vectors.	Task layer (single layer) NN.

Transformers: *the* model for NLP

- (Zachary Lipton) “2019, the year (search for) neural architecture died.”
  - 2012-2018 = "how do we learn fn approx. for given data"
  - 2019-???? = "we have good fn approx., now what?”
- For NLP: Transformer models, pre-trained on large document.
  - “Doing Natural Language Processing (NLP)? Use transformers, it probably will work well.”
  - Google BERT, Facebook XLM, OpenAI GPT-2, …
- Transformer-based models took over NLP research
  - Replaced almost all state-of-the-art scores on many NLP tasks
  - Hard to find a research paper that is not using one or two Transformer-based model.

Transformers: *the* model for NLP

- Neural network model for sequences, works very well with language.
- Flexible - can be applied to many different tasks, can be scaled up / down.
- Pre-training: pretrained on large-data, fine-tuning on (small) target data.



Transformers: coming to productive systems

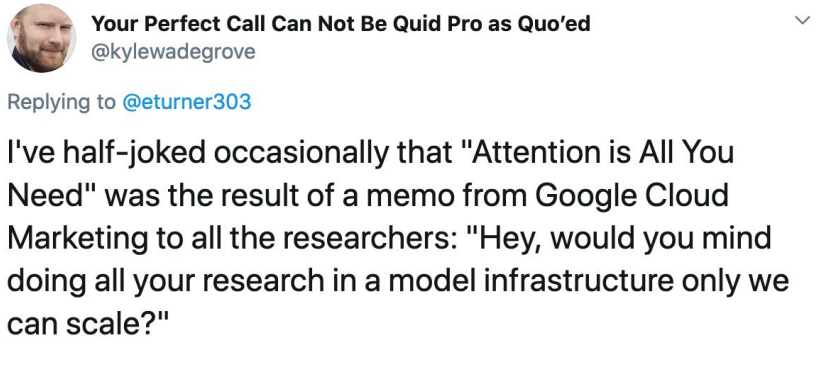
- Q: How good does it work for real-world applications?
- Q: What hinderance does it have, to be deployed for productive services?

- OMQ is progressively rolling out Transformer-based models in our products.
- In this talk, we would like to share some lessons we learned.

1. Computational cost: high but not a stopper

- Notorious for its heavy computation?
  - To be fair: training any model with that many parameters do tend to.
  - Fine-tuning don't necessarily be costly
- Inference time on CPU instance, is really what makes or breaks.
  - e.g. Several thousands of requests per minute, over 100+ different tenants
  - GPU/TPUs are still unacceptably high cost objects

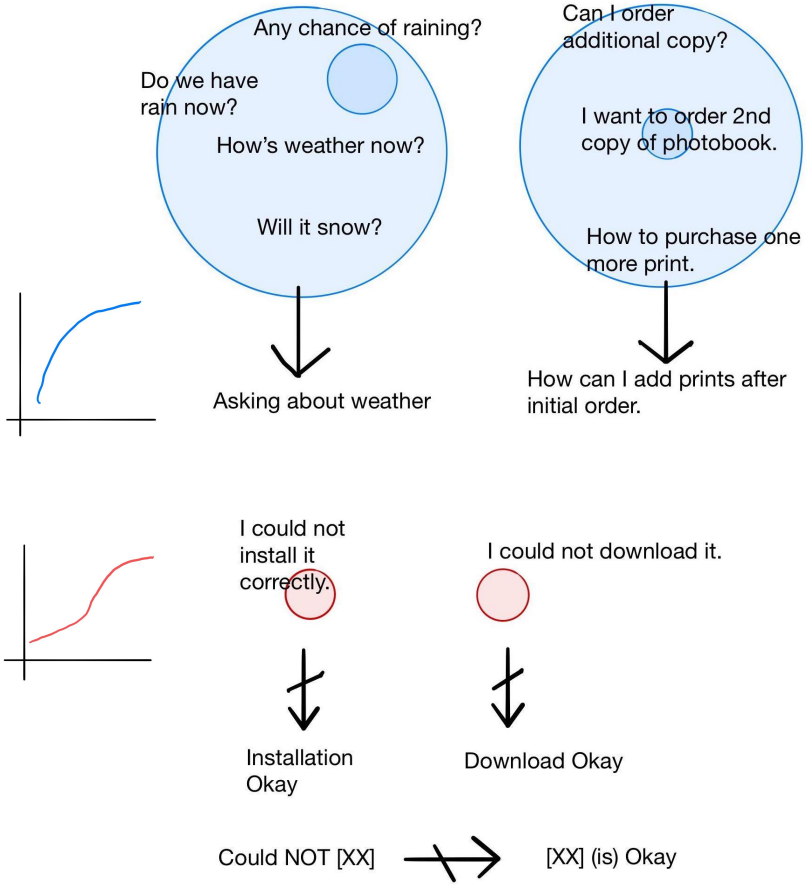
A Twitter joke, on FaceBook XLNet training cost, est. $245,000 with Google TPUs.



- Achievable: near-real time inference with reasonable computational cost
  - With some smart integrations

2. When you fine-tune: fast learn / slow learn

- One way to visualize effect of fine-tuning on pre-trained models: **“force-multiplier”**
- On what it knows already from pre-train
  - Topical relatedness
  - Paraphrasings
  - Simple reasoning
- Not on what it don’t know yet
  - Negations
  - Presuppositions
- Generalization is slower there
- Take away point
  - If the task require phenomenon not covered by pre-training: you still need good amount of data.



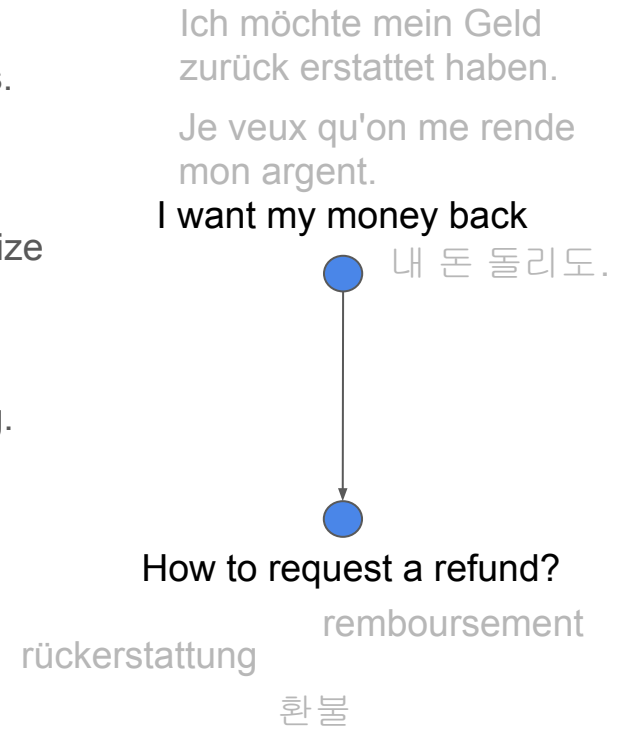
3. Do include self-supervised training of your domain

- Standard steps of using Pre-trained Transformers
  - Step 1) Pick a decent enough pre-trained transformer
  - Step 1.5) Further train on your domain text with self-supervised task.
  - Step 2) Fine-tune on your task data
- Self-supervised task of pre-training
  - Easily generated from unlabeled raw corpus
  - e.g. predict masked word, predict next sentence
- Domain data has yet more to give
  - Web scale is great, but there are things that can be picked up only by the domain text.

“gummi” in different context

4. Multilingual transformers are of great benefits

- Multilingual Transformers
  - Transformer models that are pre-trained on multiple languages.
  - Ability to make multilingual representation.
- Disparity of data over languages
  - Customer support text of a German company: German (data size 100%), English (40%), Swedish (10%), Chinese (1%), …
- Multilingual transformers performs surprisingly well.
  - “Force-multiplier” effect is even more impressive for minor lang.
- There are some pitfalls.
  - Not all language-pairs are equal.
  - Vocab-size matters.



5. Data preparation is important (as always)

- *“Your model will do anything … (to optimize loss function), it will ruthlessly do cheating or make short-cuts, if you give it half a chance.”*
- Transformers have the ability to generalize over syntax & semantics,
  - But if you give it a chance, it will cheat (!) and stuck on simplest explanation.
  - e.g. asking a person name? Answer with the second name of last sentence.
- Pitfall: unintended bias with smaller data
  - Smaller data ⇒ easier to introduce such holes
  - Transformers are relatively stronger against various issues e.g. overfitting, forgetting, …
  - But, it “knows” more, it can be also more creative at spotting such a hole
- To overcome: well designed training (data) preparation
  - We are never free from making careful training design.

The new representation: improvements

- The new model handles better on phrase/sentence structure
  - Ich möchte noch eien 2. Fotobuch dem Warenkorb hinzufügen. (I want to add 2nd photobook to the cart.)
    => "Kann ich mehrere Produkte gleichzeitig bestellen?" (Can I order multiple products at the same time?)
  - Kann ich mein letztes Fotobuch noch einmal gedruckt bekommen? (Can I get my last photobook printed again?)
    =>"Wie kann ich Produkte nachbestellen?" (How can I make a repeat-order?)
- Better handling of nuance
  - I've installed GoodMusic 2020, but I am having error code 36. *I am using Window 10, and installation went without issue.*
- Cross-lingual matching
  - [EN] "How to unlock previously purchased downloads?"
  - (System reports) 14 Dutch requests last week seems to be matching this [EN] question with high score. But there is no knowledge base entry for this in [NL]

Conclusion

- Pretrained Transformers: the model of NLP as of now.
  - Can handle almost all linguistic phenomenon (… given that you provide enough data)
  - Pre-training: works like a force-multiplier. Small amount of data to work better (with caution!)
  - Nicely catches up variety in natural language expressions.
- Transformer-based representations in OMQ
  - Currently rolling out for selected customers + services
  - Scheduled to be served for all our customers in mid-2020
- This is a very interesting time!
  - Fast-paced research progress, with practical code & models.
  - We expect even more powerful / clear / applicable methods and models to come.
- We aim to deliver latest progress in NLP research to the market, and share the progress.

Thanks!

- Visit our website for more information: **omq.ai**
- The talk video, and materials will be shared on our website - blog.

- Follow us on social media, to keep track of our progress:
  - Twitter: @OMQ_AI
  - Linked-in: OMQ (https://www.linkedin.com/company/omq)
  - Instagram: @omq_ai