

Transformer Contextualization & Re-Ranking

Sebastian Hofstätter

sebastian.hofstaetter@tuwien.ac.at

 /s_hofstaetter

Today

Transformer Contextualization & Re-Ranking

- 1 Contextualized Representation Learning
 - Self-attention /w Transformers
 - Pre-trained models (BERT et al.)
- 2 Re-Ranking with Transformers
 - Using BERT
 - Efficient Transformer-Kernel architectures

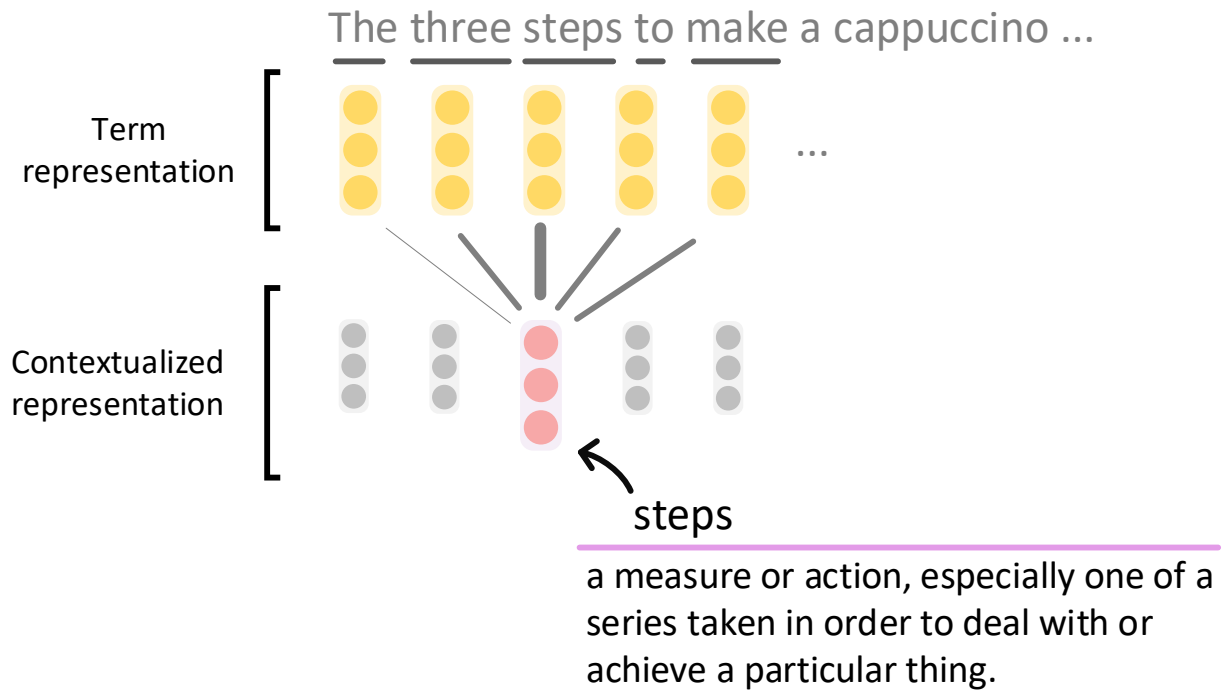
Now we Reached the State-of-the-art

- Finally! 🍰
- Fast moving field:
 - Most of the contents of this lecture did not exist in the beginning of 2018
 - And by July we probably have a new SOTA technique
- Many questions are open
 - We try to answer a few here
- General direction: More computation = better results,
possible by better hardware & more data

Contextualized Representation Learning

Self-attention attending to itself

Contextualization via Self-Attention



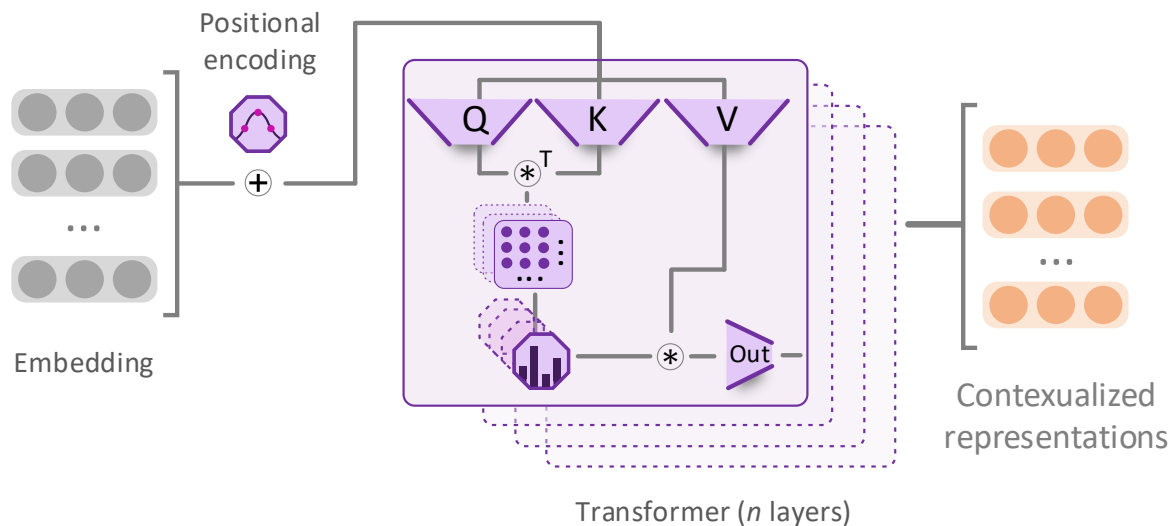
- Learn meaning based on surrounding context for every word occurrence
- This *contextualization* combines representations
- Context here is local to the sequence (not necessary a fixed window)
- Is computationally intensive $O(n^2)$
 - Every token attends to every other token

Transformer

- Transformers contextualize with multi-head self-attention
 - Every token attends to every other token $O(n^2)$ complexity
- Commonly Transformers stack many layers
- Can be utilized as encoder-only or encoder-decoder combination
- Do not require any recurrence
 - The attention breaks down to a series of matrix multiplications over the sequence
- Initially proposed in translation
 - Now the backbone of virtually every NLP advancement in the last years

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, et al.
Attention is all you need. In NeurIPS. 2017.

Transformer – Architecture



- We embed (subword) tokens
- We add a positional encoding
- In each Transformer-Layer:
 - Project each vector with 3 linear layers to **Query**, **Key**, **Value**
 - Transform projections to another multi-head dimension
 - Matrix-multiply Query & Key
 - Get Q-K attention via softmax
 - Multiply attention with Values and project back to output

Nice detailed walkthrough code + paper:
<https://nlp.seas.harvard.edu/2018/04/03/attention.html>

Self-Attention Definition

- The Transformer Self-Attention is defined as:

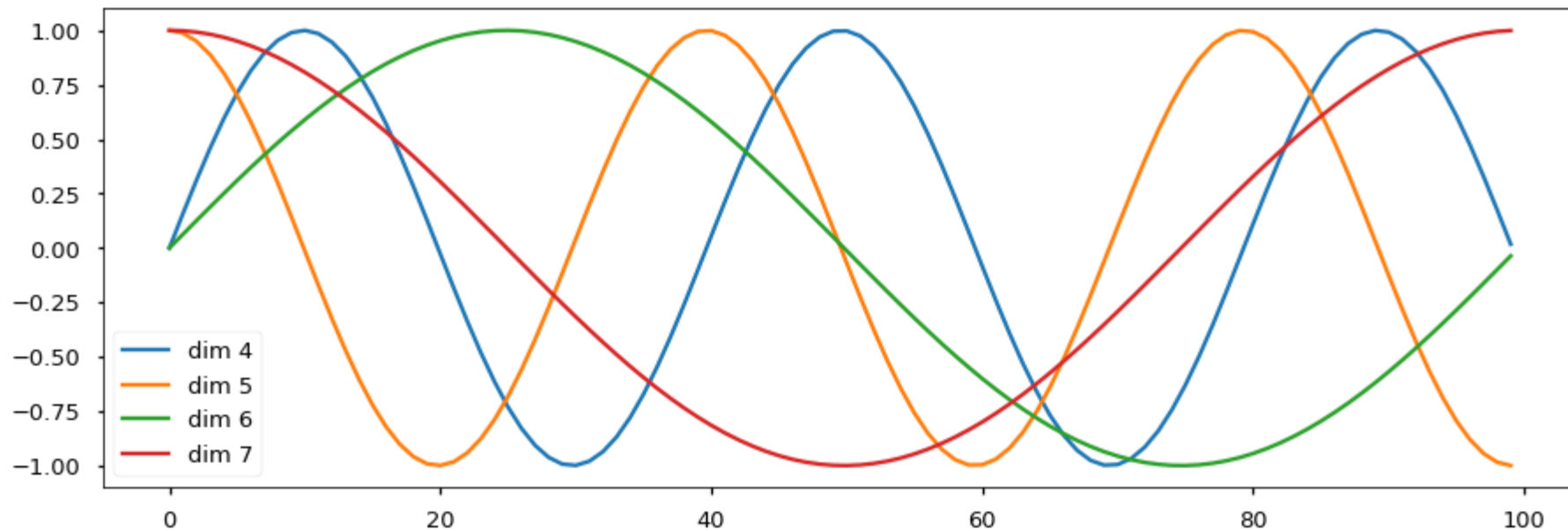
$$\text{SelfAttention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) * V$$

- Q, K, V are projections of the **same input** sequence
- This definition hides quite a bit of complexity, visible in the code

Q	Attention “Query”
K	Attention “Key”
V	Attention “Value”
d_k	Dimension of key embeddings

Transformer – Positional Encoding

- Transformers add sinusoid curves to the input, before the attention
 - Informs about relative position inside the sequence
 - Removes need for explicit recurrence patterns



Transformer - Variations

Overview of recent Transformer literature [Weng, 2020]

Attention is all you need [Vaswani et al., 2017]

Running self-attention on pre-segmented text [Al-Rfou et al., 2019, Hofstätter et al., 2020]

Localized Attention Span (Image Transformer) [Parmar et al., 2018]

Transformer-XL [Dai et al., 2019]

XLNet [Yang et al., 2019]

Gated Transformer-XL [Parisotto et al., 2019]

Reformer [Kitaev et al., 2019]

Reversible Residual Network [Gomez et al., 2017]

Routing Transformer [Roy et al., 2020]

Sparse Sinkhorn Attention [Tay et al., 2020]

Sparse Transformers [Child et al., 2019]

Megatron LM [Shoeybi et al., 2019]

Longformer [Beltagy et al., 2020]

Transformer-XH [Zhao et al., 2014]

Roberta [Liu et al., 2019]

Adaptive Attention Span [Sukhbaatar et al., 2019]

Adaptive Computation Time [Graves, 2016]

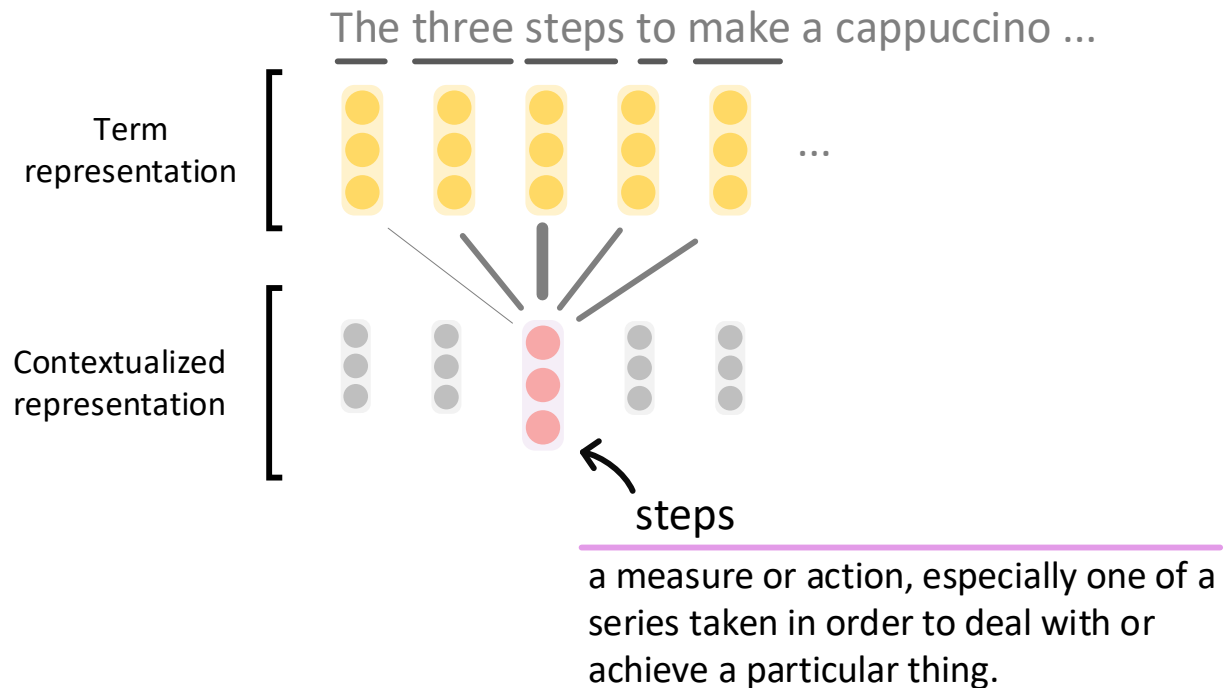
Universal Transformers [Dehghani et al., 2018]

- Non-exhaustive (probably) list of Transformer variants
- A lot focus on efficiency & long input
 - Break $O(n^2)$ runtime and memory requirement
- Incredible speed of innovation

More at:

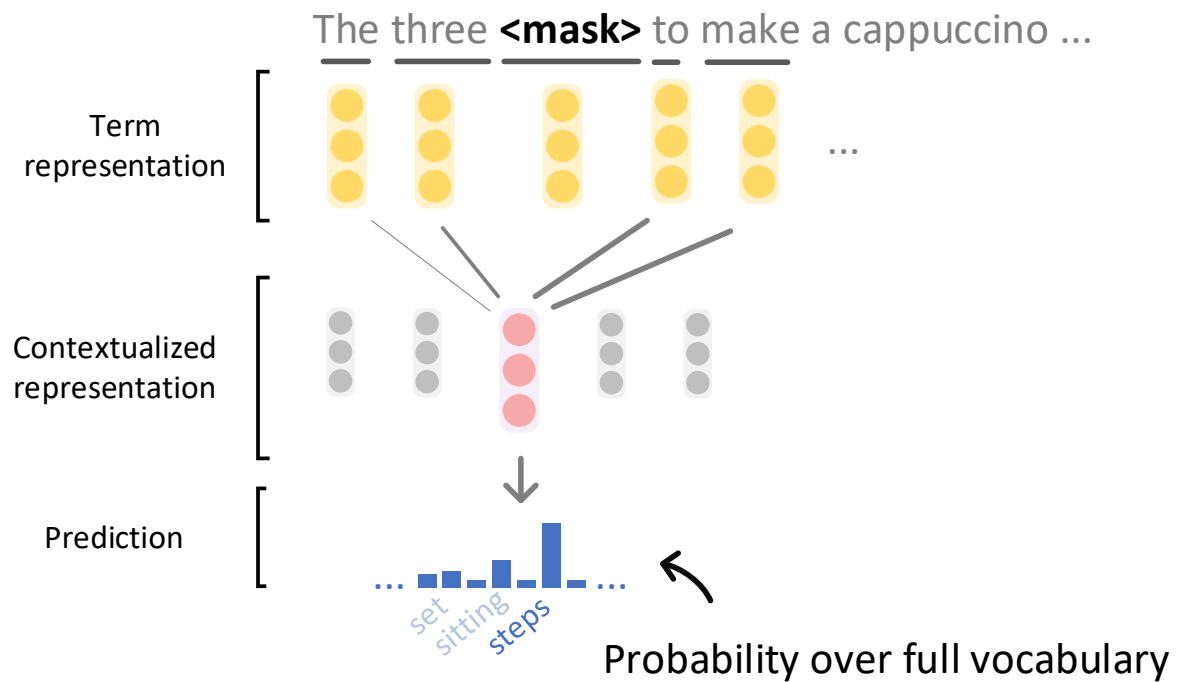
<https://lilianweng.github.io/lil-log/2018/06/24/attention-attention.html>

Masked Language Modelling



- Recall our example:
 - We want a good context-dependent representation of “steps”
- Unsupervised Pre-training:
 - Take text and mask random words
 - Try to predict original word
 - Update weights based on loss of prediction vs. actual word

Masked Language Modelling



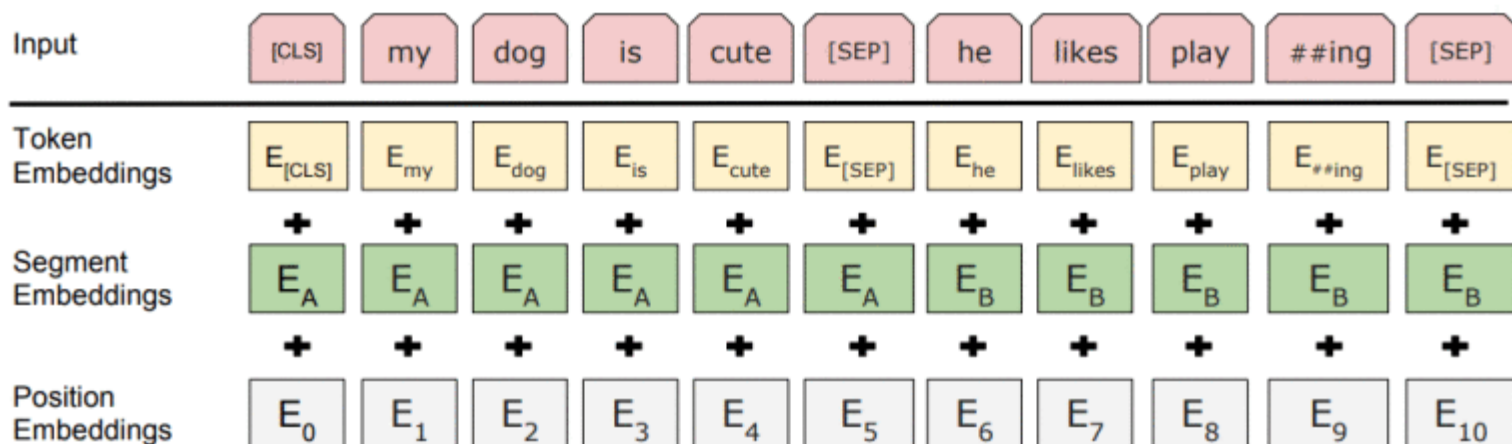
- Training procedure:
 - Take text and mask random words
 - Try to predict original word from context words
 - Update weights based on loss of prediction vs. actual word
- Loss requires prediction over vocabulary
 - Prohibitive for large vocabs
 - Models use WordPiece or BytePair splitting of infrequent terms

BERT

- **Bidirectional Encoder Representations from Transformers**
- Large effectiveness gains on *all* NLP tasks
- Ingredients:
 - WordPiece Tokenization & Embedding (small vocab, covers infrequent terms)
 - Large model (many dimensions and layers – base: 12 layers and 768 dim.)
 - Special tokens (shared use between pre-training and fine-tuning)
 - **[CLS]** Classification token, used as pooling operator to get a single vector per sequence
 - **[MASK]** Used in the masked language model, to predict this word
 - **[SEP]** Used to indicate (+ sequence encodings) a second sentence
 - Long MLM pre-training (weeks if done on 1 GPU)

BERT - Input

- Either one or two sentences, always prepended with [CLS]
 - BERT adds trained position embeddings & sequence embeddings

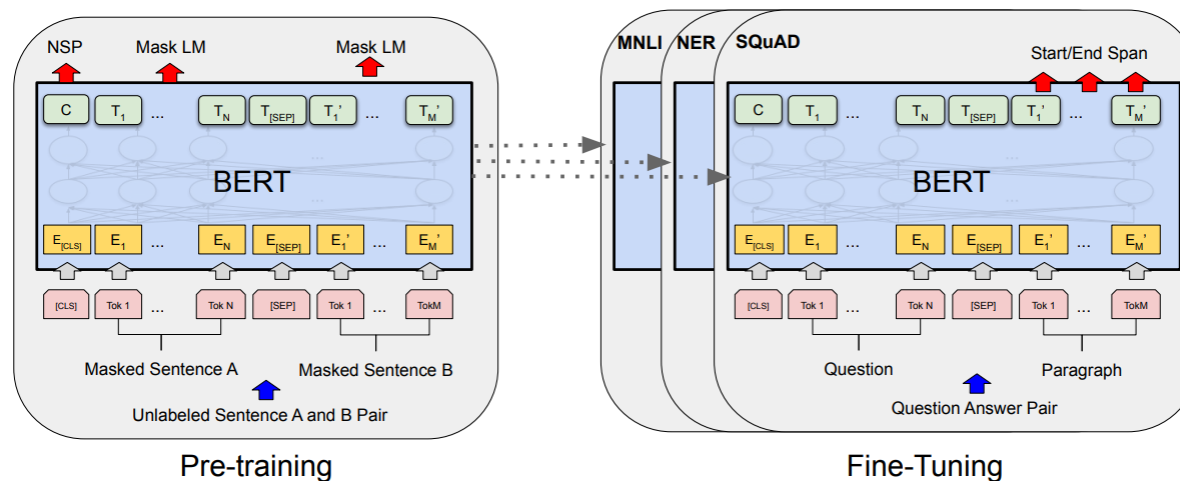


BERT - Model

- Model itself is quite simple: n Layers of stacked Transformers
 - Using LayerNorm, GeLU activations (like ReLU, but with a grace swing under 0)
 - Task specific heads on top to pool [CLS] or individual token representations
 - Every Transformer layer receives as input the output of the previous one
- The [CLS] token itself is only special because we train it to be
 - No mechanism inside the model that differentiates it from other tokens
- Novel contributions center around pre-training & workflow

BERT - Workflow

- Someone with lots of compute or time pre-trains a large model
 - BERT uses Masked Language Modelling [MASK] and Next Sentence Prediction [CLS]
- We download it and fine-tune on our task



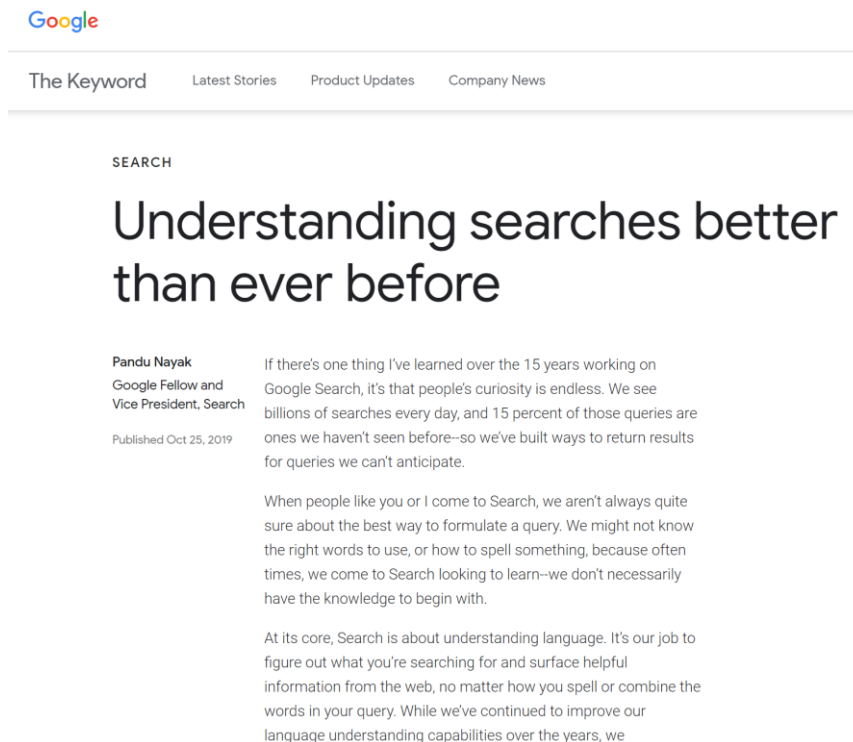
Beyond BERT

- Same as with Transformer variations, there are now many BERT variants
 - For many languages
 - Domains like biomedical publications
 - Different architectures, but similar workflow: Roberta, Transformer-XL, XLNet, Longformer ...
- Main themes for adapted architectures:
 - Bigger
 - More efficient
 - Allowing for longer Sequences (BERT is capped at 512 tokens in total)

Re-Ranking with Transformers

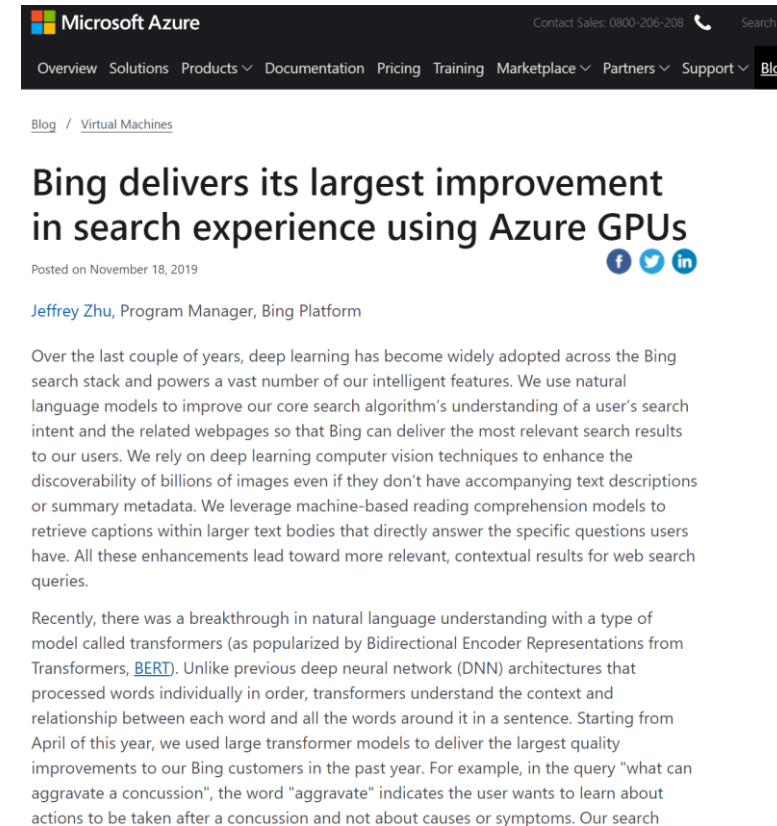
Can be slow and effective or fast and almost as effective

The Case for Contextualization in IR



The screenshot shows the Google homepage with a navigation bar containing 'The Keyword', 'Latest Stories', 'Product Updates', and 'Company News'. Below the navigation bar, the word 'SEARCH' is displayed. The main headline reads 'Understanding searches better than ever before'. To the left of the main text, the author's name 'Pandu Nayak' is listed, followed by his title 'Google Fellow and Vice President, Search' and the publication date 'Published Oct 25, 2019'. The main text begins with 'If there's one thing I've learned over the 15 years working on Google Search, it's that people's curiosity is endless. We see billions of searches every day, and 15 percent of those queries are ones we haven't seen before--so we've built ways to return results for queries we can't anticipate.' The text continues with 'When people like you or I come to Search, we aren't always quite sure about the best way to formulate a query. We might not know the right words to use, or how to spell something, because often times, we come to Search looking to learn--we don't necessarily have the knowledge to begin with.' The final paragraph states 'At its core, Search is about understanding language. It's our job to figure out what you're searching for and surface helpful information from the web, no matter how you spell or combine the words in your query. While we've continued to improve our language understanding capabilities over the years, we

Google (October 2019)



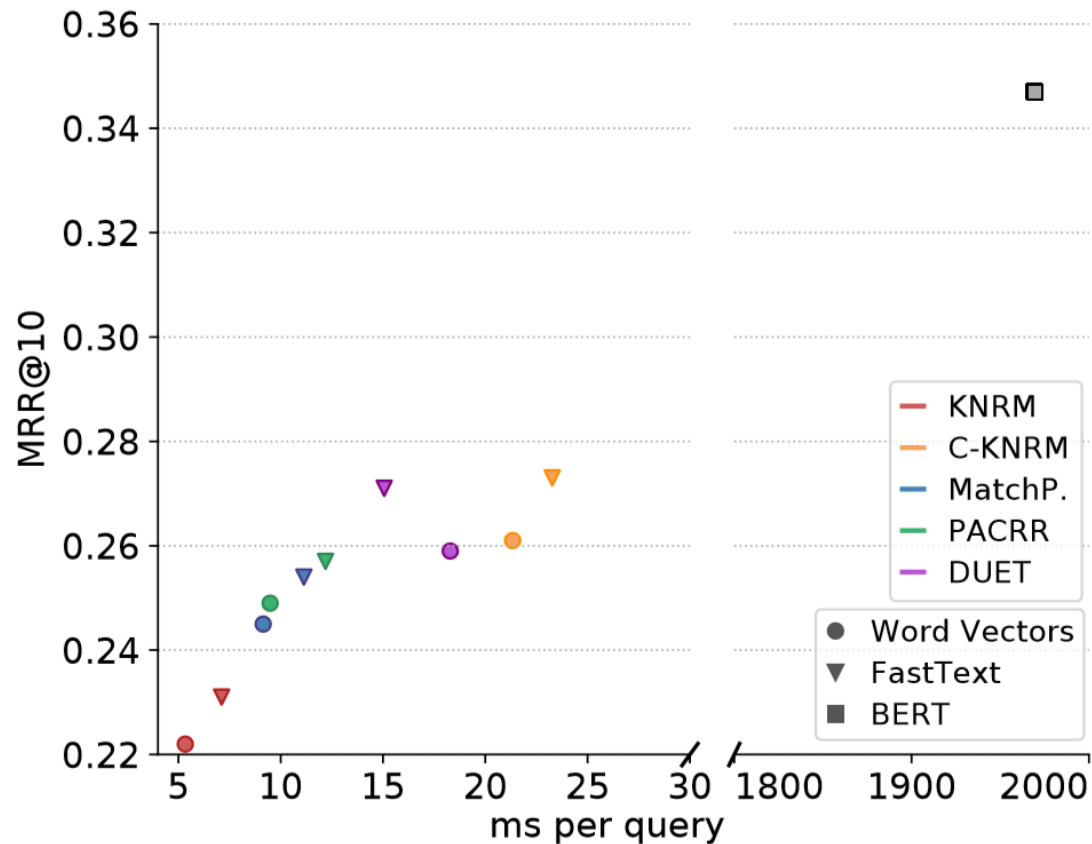
The screenshot shows the Microsoft Azure website with a dark header containing the 'Microsoft Azure' logo, a 'Contact Sales' link with a phone icon, and a search bar. The navigation bar includes 'Overview', 'Solutions', 'Products', 'Documentation', 'Pricing', 'Training', 'Marketplace', 'Partners', 'Support', and 'Blog'. Below the navigation bar, the breadcrumb 'Blog / Virtual Machines' is visible. The main headline reads 'Bing delivers its largest improvement in search experience using Azure GPUs'. Below the headline, the date 'Posted on November 18, 2019' and social media icons for Facebook, Twitter, and LinkedIn are shown. The author's name 'Jeffrey Zhu, Program Manager, Bing Platform' is listed. The main text begins with 'Over the last couple of years, deep learning has become widely adopted across the Bing search stack and powers a vast number of our intelligent features. We use natural language models to improve our core search algorithm's understanding of a user's search intent and the related webpages so that Bing can deliver the most relevant search results to our users. We rely on deep learning computer vision techniques to enhance the discoverability of billions of images even if they don't have accompanying text descriptions or summary metadata. We leverage machine-based reading comprehension models to retrieve captions within larger text bodies that directly answer the specific questions users have. All these enhancements lead toward more relevant, contextual results for web search queries.' The final paragraph states 'Recently, there was a breakthrough in natural language understanding with a type of model called transformers (as popularized by Bidirectional Encoder Representations from Transformers, [BERT](#)). Unlike previous deep neural network (DNN) architectures that processed words individually in order, transformers understand the context and relationship between each word and all the words around it in a sentence. Starting from April of this year, we used large transformer models to deliver the largest quality improvements to our Bing customers in the past year. For example, in the query "what can aggravate a concussion", the word "aggravate" indicates the user wants to learn about actions to be taken after a concussion and not about causes or symptoms. Our search

Microsoft (November 2019)

BERT Re-Ranking

- Scoring 1 query and 1 passage (from a candidate set)
- Concatenating the two sequences to fit BERT's workflow
 - [CLS] query [SEP] passage
 - Pool [CLS] token to predict score
 - Train with pairwise ranking loss
- Works awesome out of the box
 - Major jumps in effectiveness across collections and domains
 - But, of course, comes at the cost of performance and virtually no interpretability

BERT In-Efficiency



- Evaluated on 250 docs / query on short MSMARCO-Passage (*max 200 tokens*)
- IR-specific networks are fast, but moderately effective
- Transformer-based BERT is very effective, but very slow
 - + Infrastructure cost (blocking 1 GPU for 2 seconds at a time)
- Tradeoff well studied in classical learning-to-rank, but unexplored in neural models

Sebastian Hofstätter and Allan Hanbury. 2019.

Let's measure runtime! Extending the IR replicability infrastructure to include performance aspects . In OSIRRC @ SIGIR.

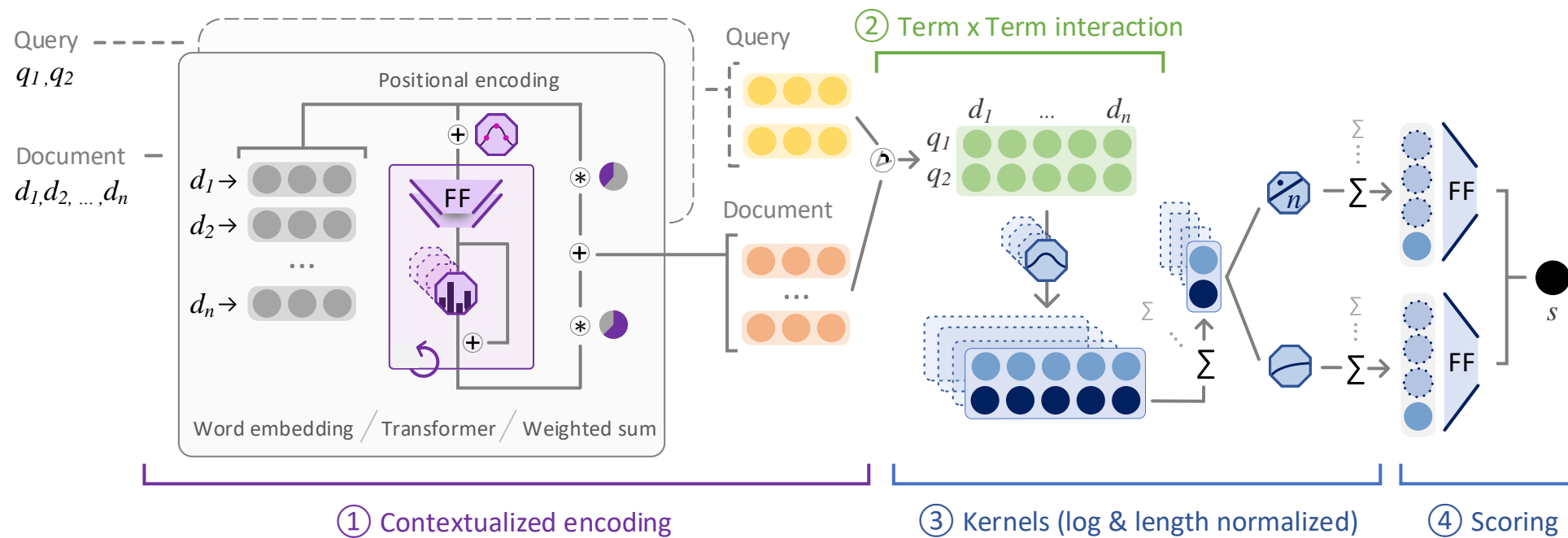
Efficiency – Effectiveness Tradeoff

- Large influence on how we employ a model
- Main outside factor: how many documents to re-rank
- Faster models can re-rank more documents in the same time as slower ones
- Evaluate based on a time budget
 - Allows us to simultaneously evaluate effectiveness & efficiency in a realistic setting

TK: Transformer-Kernel Ranking

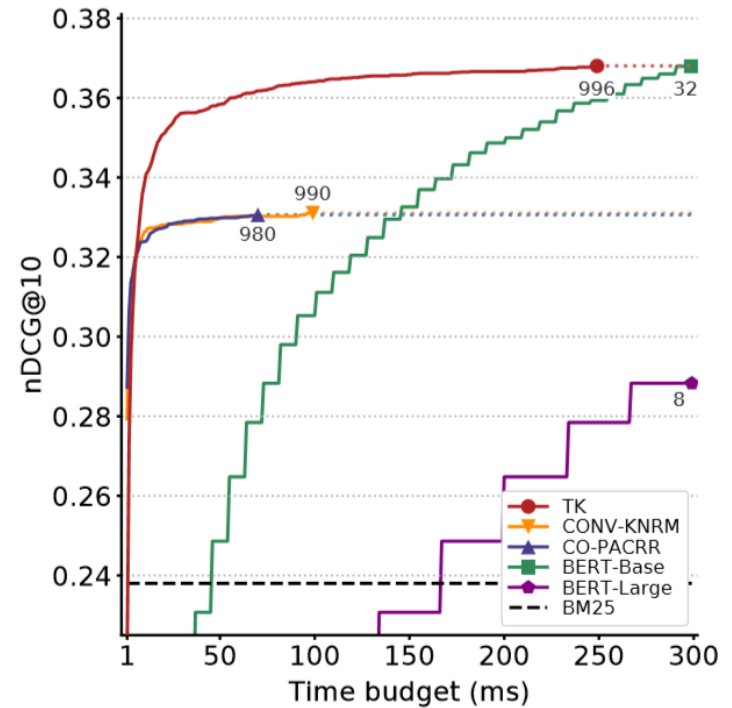
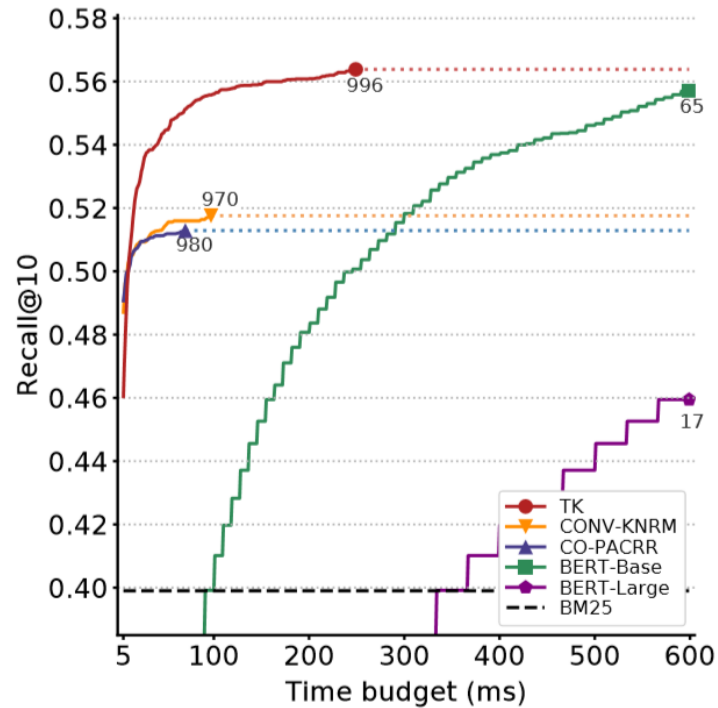
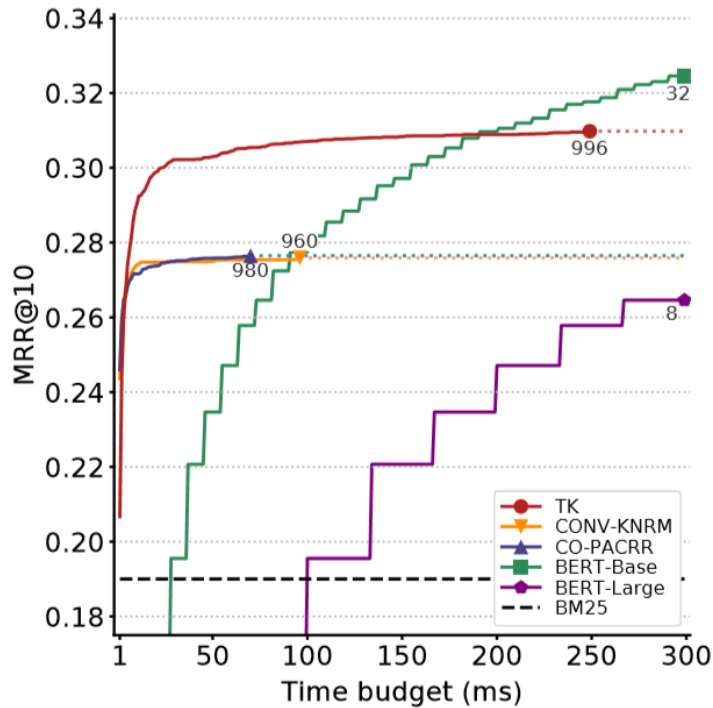
- Desired properties: Lightweight, interpretable, effective
- We proposed the TK model
 - Combine Transformer-contextualization with kernel-pooling
 - Strong results compared to IR-specific models
 - State-of-the-art model for time-budget constrained environments
- Use Transformer-blocks as contextualization layer
 - Create hybrid contextualization by merging context & non-context
- Limit the number of Transformer layers, as each additional layer takes considerable amount of time with diminishing returns

TK: Transformer-Kernel Ranking



TK: Transformer-Kernel Ranking

- More documents in the same time = better results




(a) MSMARCO-Passage


Understanding TK

neural-ir-explorer – TK using 2 Transformer layers @ MSMARCO-Passage (Dev)

< | when and where did paella originate ?

Min. similarity  or Kernels 1 0.9 0.7 0.5 0.3 0.1 -0.1


Rank Score (length & log) Log-Kernel scores: 1 to -0.1 & rest

★ ① -23.57 (1.58 & -25.15) -4.38 -0.70 4.41 -2.19 -2.69 0.02 -0.65 -18.96 


Paella has ancient roots , but its modern form originated in the mid - 19th century in the area around Albufera lagoon on the east coast of Spain , adjacent to the city of Valencia . [4] Many non - Spaniards view paella as Spain 's national dish , but most Spaniards consider it to be a regional Valencian dish .

② -23.90 (1.20 & -25.10) -4.45 -0.75 4.07 -2.05 -2.44 0.01 -0.55 -18.94 


Paella is a world - famous dish , which originated in the region of Valencia , in eastern Spain . It is now widely eaten in all provinces of Spain , as well as every continent of the world . Like so many other popular recipes , Valencian paella was initially a peasant dish .

③ -24.83 (1.72 & -26.55) -4.95 -1.03 4.00 -2.20 -2.97 0.01 -0.53 -18.89 

Paella is a dish that originated from Valencia , Spain . The main ingredients are rice , and saffron . Everyone has a different recipe for paella . Some paella is made from all seafood , some from all meat , and some are a mix of meat and seafood . Here is my recipe for paella . It is easy to make and very authentic tasting .

④ -25.13 (1.66 & -26.79) -5.14 -1.12 3.65 -2.01 -2.53 0.03 -0.74 -18.93 


Paella . Originating in Valencia , paella is a rice dish prepared with seafood . Of all the foods in Spain , this is the most popular . In this dish , savory yellow rice is combined with tomatoes , onions , peas , shellfish , squid , clams and chicken drumsticks .

⑤ -25.53 (1.82 & -27.34) -5.56 -1.25 3.87 -2.12 -2.61 0.03 -0.80 -18.91 

Authentic Paella Valenciana . I lived in Spain for two years where I was taught the art of making the Paella which originated in Valencia . I have n't found anything on here which is even close to authentic , so I thought I would add this recipe for those who would like to try a taste of Spain .

- Demo application to showcase TK
 - Displays internal similarity & kernel results
- Users can browse around the results
 - Get an overview over the queries
 - Dig deep into a single result
- Complements metric-based evaluation
- Allows users to develop a “feeling” for the test collection & model used

Sebastian Hofstätter, Markus Zlabinger and Allan Hanbury. 2020. Neural-IR-Explorer: A Content-Focused Tool to Explore Neural Re-ranking Results. In ECIR.

Sort    Collapse clusters Prefix filter

Let's start exploring

Here is what you see around you and what you can do with it:

- At the top you can sort the queries: randomly, ascending or descending (based on the rank of the first relevant document). You can also expand the clusters to see all queries.
- We clustered the queries based on their mean contextualized encoding. Each card holds the queries for a cluster.
- At the top of each card is: the median best rank of the neural model, the difference to the first-stage baseline, and a manual summary of the queries in that cluster
- Each query line contains: the best rank of the neural model, the difference to the first-stage baseline, and the query
- Simply click on a query to go to the result view and see details on the query result

⑤ 3 where is location

- 1 1 what airport is in wilder ky
- 1 0 where is alepotrypa cave
- 1 0 where is azaz
- 1 2 where is bell buckle tn
- 1 4 where is boston georgia
- 1 0 where is henry's plant farm
- 1 0 where is last name hollis derived from
- 1 1 where is lima beads located
- 1 3 where is mathura

⑨ 5 general knowledge questions (trivia)

- 1 0 do vhi swiftcare do blood tests?
- 1 1 what do partnerships file tax in michigan
- 1 6 what does android sdk tools do
- 1 0 what eventually replaced the cottage industry
- 1 9 what federal statute gives the epa authority to regulate pesticides
- 1 0 what navy installation support camp david
- 1 3 how can deforestation directly affect living organisms
- 1 13 how do active transport and passive transport differ
- 1 0 when do atoms become excited
- 1 0 definition do classic
- 1 0 more queries (click to expand)

Live Demo available at

<https://neural-ir-explorer.ec.tuwien.ac.at/>

② 0 phone number

- 1 1 texas roadhouse glen mills pa phone number
- 1 0 the miners state bank routing number
- 1 0 usf admissions office phone number
- 1 0 vermont casting group phone number
- 1 0 dr azadpour phone number
- 1 1 dr. richard spech npi number
- 1 0 colorado routing number loveland colorado
- 1 0 green horizon mini storage contact number
- 1 0 cox business omaha phone number
- 1 3 dcu electronic routing number
- + 100 more queries (click to expand)

④ 2 location questions

- 1 0 hotels in thornton co
- 1 3 does azusa pacific university negotiate salary
- 1 0 troy student population
- 1 0 canada most dense area
- 1 0 carbon reactivation facilities california
- 1 0 what is the zip code in arrowhead lakes
- 1 0 honey in south carolina
- 1 0 honolulu chinese new year celebration
- 1 0 which bbc radio station specializes in sports commentaries
- 1 0 how old is dalton rapattoni
- + 172 more queries (click to expand)

⑨ 0 weather and climate

- 1 2 weather in greenbelt md

⑭ 8 what is/are 2+ words

- 1 3 what are caged ibc tanks used for

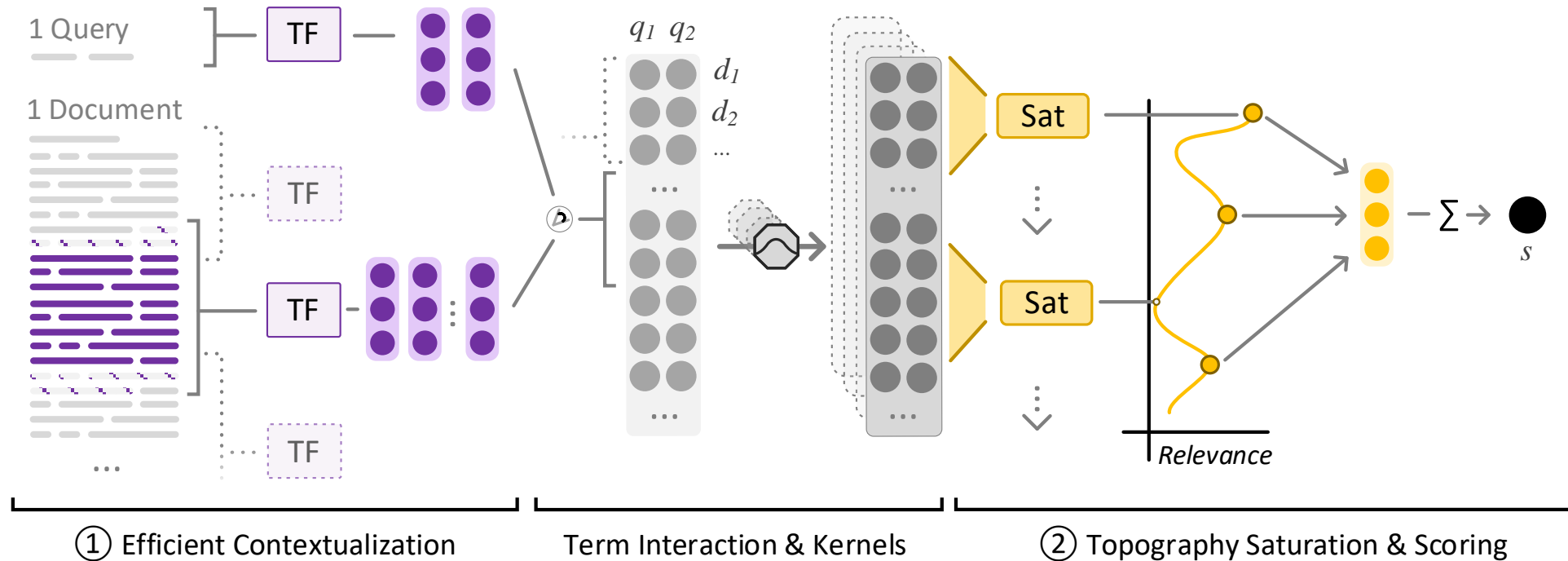
⑨ 7.5 historical dates & times

- 1 62 what month does winter start in new zealand

TKL: Transformer-Kernel for Long Documents

- Long documents (>200 tokens) are very slow
 - 300-dimensional vector per each word
 - Re-rank 100s of documents per query
 - Padding techniques are insufficient
- State-of-the-art models don't work
 - Do not contain a notion of region importance
 - Current best approach split a document and score sentences/paragraphs individually
- We proposed an extension to TK for Long documents (TKL)

TKL: Transformer-Kernel for Long Documents



TKL: Why long documents?

- We found that longer document input gives us better results
- But only if we do top region detection in TKL
- Main idea behind exercise 1:
 - Find out if the model was correct in this assumption
 - Can we prove that we need longer input

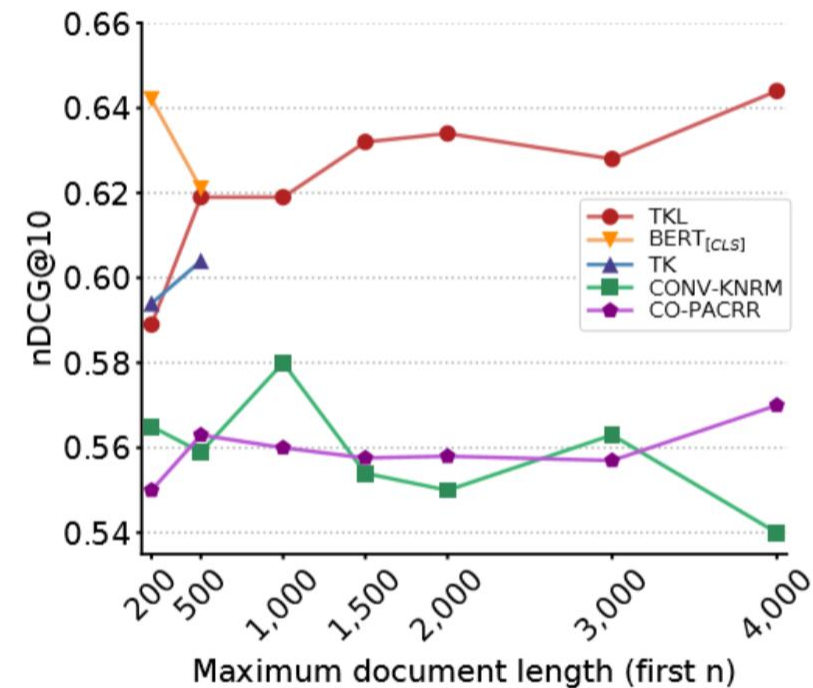
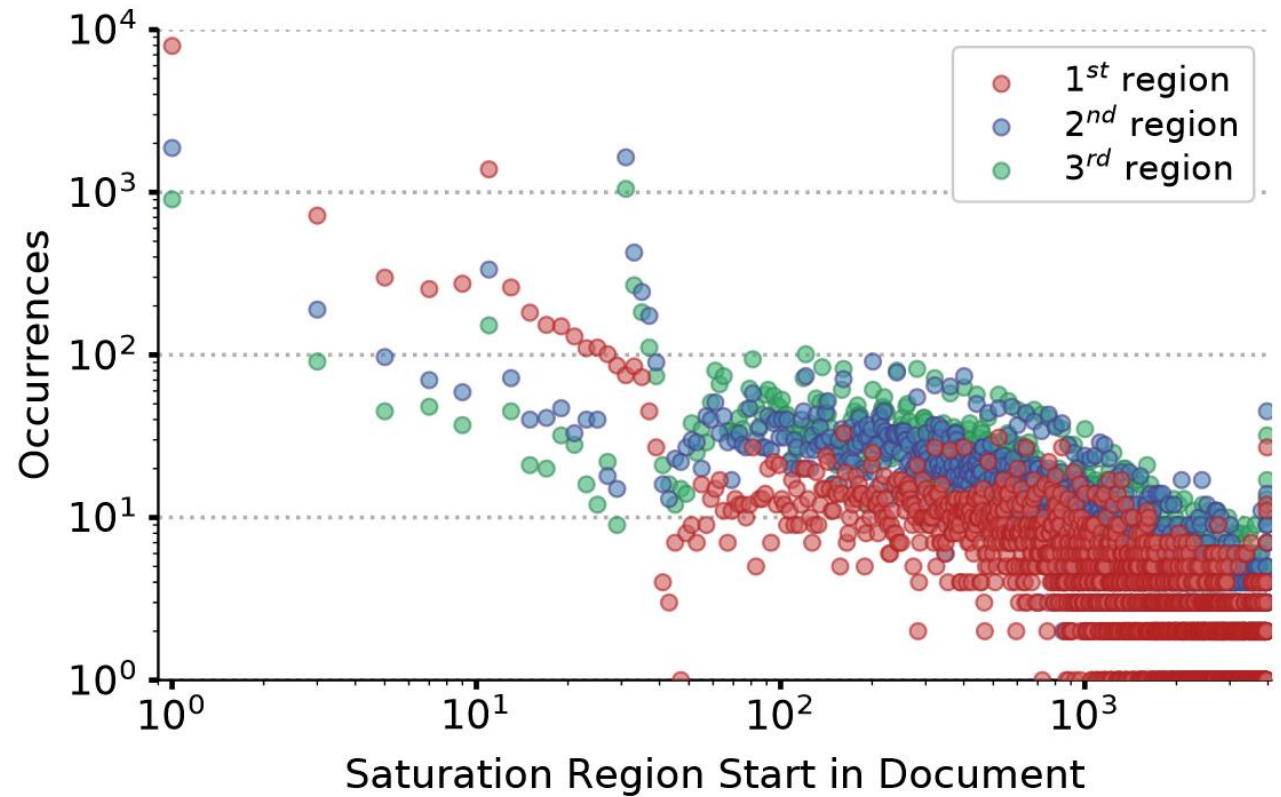


Figure 2: TREC-2019 results based on the document length.

TKL: Where is the relevance?

- TKL provides relevant regions location in the document
- Occurrences follow a Zipfian-Distribution
 - In the TREC-DL 2019 Document collection
- Could be used as snippet generation
 - Better user interfaces



Summary: Contextualization & Re-Ranking

- 1 Transformers apply self-attention to contextualize words
- 2 BERT provides enormous effectiveness jumps at the cost of speed
- 3 Combining Transformers and Kernels leads to a good compromise

- 1 Transformers apply self-attention to contextualize words
- 2 BERT provides enormous effectiveness jumps at the cost of speed
- 3 Combining Transformers and Kernels leads to a good compromise

Thank You