Question Answering
&
Question Generation

Thomas Scialom, Paul-Alexis Dray

14/02/2019 – LIP6

Question Answering

Question Answering
SQuAD

Given a context and a question, the model infers an answer.

The answer can be a span of words, multiple choice or human generated.

In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under **gravity**. The main forms of precipitation include drizzle, rain, sleet, snow, **graupel** and hail... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals **within a cloud**. Short, intense periods of rain in scattered locations are called "showers".

What causes precipitation to fall?
gravity

What is another main form of precipitation besides drizzle, rain, snow, sleet and hail?
graupel

Where do water droplets collide with ice crystals to form precipitation?
within a cloud

Question Answering

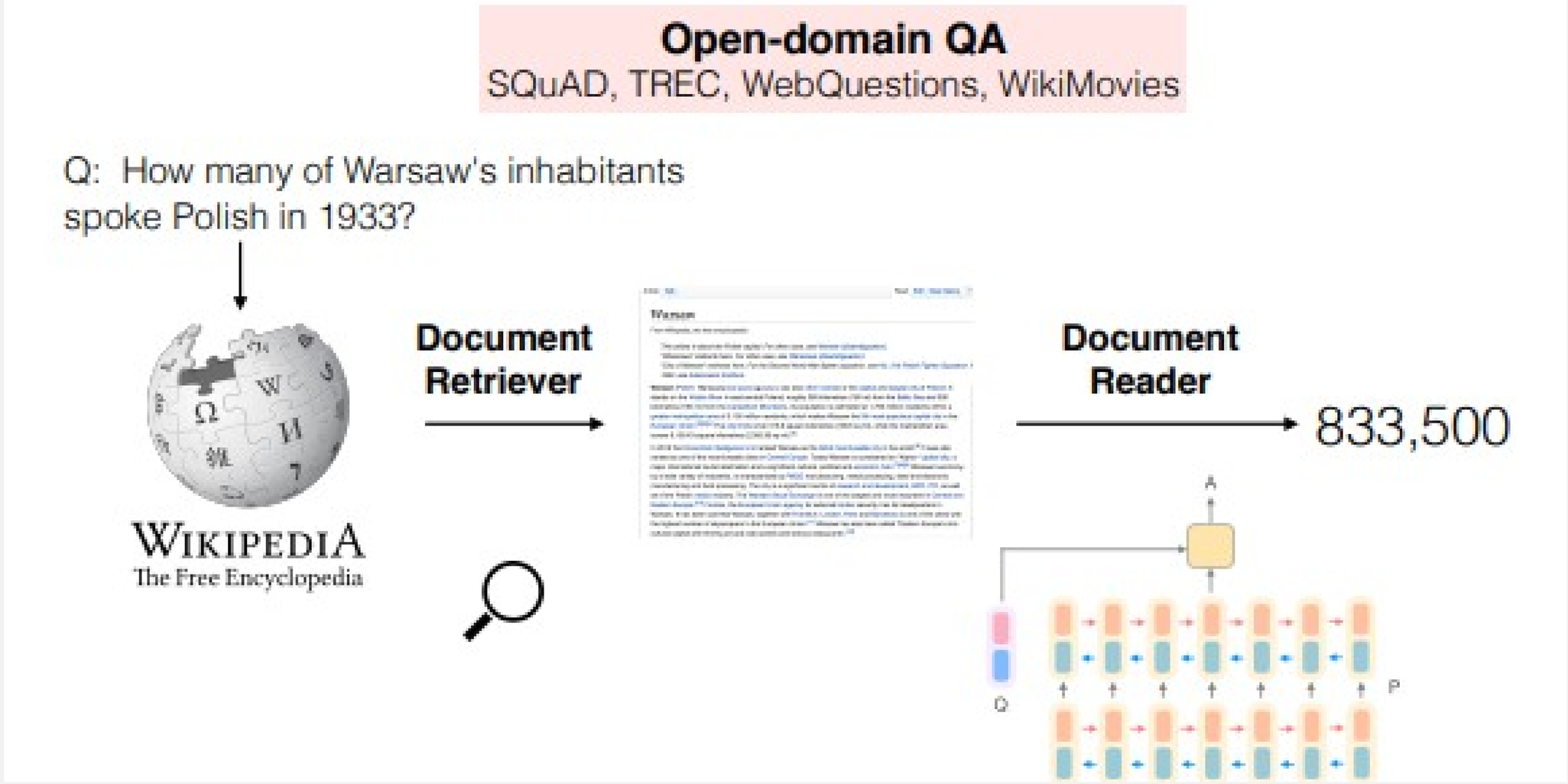
Many different datasets

Dataset	Segment	Question Source	Answer	# Questions	# Documents
NewsQA	No	Crowd-sourced	Span of words	100k	10k
DuReader	No	Crowd-sourced	Human generated	200k	1M
NarrativeQA	No	Crowd-sourced	Human generated	46,765	1,572 stories
SearchQA	No	Generated	Span of words	140k	6.9M passages
RACE	No	Crowd-sourced	Multiple choice	97k	28k
ARC	No	Generated	Multiple choice	7,787	14M sentences
SQuAD	No	Crowd-sourced	Span of words	100K	536
MS MARCO	Yes	User logs	Human generated	1M	8.8M passages, 3.2m docs.

Bajaj, P., Campos, D., Craswell, N., Deng, L., Gao, J., Liu, X., ... & Rosenberg, M. (2016). MS MARCO: A Human Generated MAchine Reading COmprehension Dataset. *arXiv preprint arXiv:1611.09268*.

Open Domain Question Answering

Open Domain QA



Chen, D., Fisch, A., Weston, J., & Bordes, A. (2017). Reading wikipedia to answer open-domain questions. *arXiv preprint arXiv:1704.00051*.

Open Domain QA

Bert is all you need ?

1/ Question Answering

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Jan 15, 2019	BERT + MMFT + ADA (ensemble) *Microsoft Research Asia*	85.082	87.615
2 Jan 10, 2019	BERT + Synthetic Self-Training (ensemble) *Google AI Language* https://github.com/google-research/bert	84.292	86.967
3 Dec 13, 2018	BERT finetune baseline (ensemble) *Anonymous*	83.536	86.096
4 Dec 16, 2018	Lunet + Verifier + BERT (ensemble) *Layer 6 AI NLP Team*	83.469	86.043
4 Dec 21, 2018	PAML+BERT (ensemble model) *PINGAN GammaLab*	83.457	86.122

SQuAD 2 leaderboard https://rajpurkar.github.io/SQuAD-explorer/

Open Domain QA

Bert is all you need ?

2/ Information Retrieval

By replacing our aggregator with BERT, we improve performance by 50-100% in all three datasets (RL-10-Sub + BERT Aggregator). This is a remarkable improvement given that we used BERT without any modification from its original implementation. Without using our reformulation agents, the performance drops by 3-10% (RM3 + BERT Aggregator).

Nogueira, R., Bulian, J., & Ciaramita, M. (2018). Learning to Coordinate Multiple Reinforcement Learning Agents for Diverse Query Reformulation. *arXiv preprint arXiv:1809.10658*.

Open Domain QA

Bert is all you need ?

3/ Multilingual transfer learning

Thomas Scialom @ThomasScialom · 2 févr.

Indeed we did so at @RecitalAI. Yes it works well BUT little resources natively available in non-eng. Here the BERT multilingual results. Note that our french Squad is a machine translation of the En. version => might be biased and not perfect.

Traduire le Tweet

finetuning	evaluation	EM	F1
EN	EN	81.60	88.88
EN	FR	56.85	73.70
FR	FR	63.53	77.64
EN+FR	FR	64.17	78.50

Table 1. Results obtained using the original SQuAD dataset (EN) and its automatically translated French version.

1 4 15

Matt Gardner @nlpmattg · 2 févr.

Wow, that's shockingly good, except the machine translation does muddy it a bit - not clear if you would get the same impressive results with natural French.

Traduire le Tweet

Open Domain QA

Bert is all you need ?

Still a lot of limitations.

One of the most important is probably the fixed length context.

What come next ? BERT-XL ?

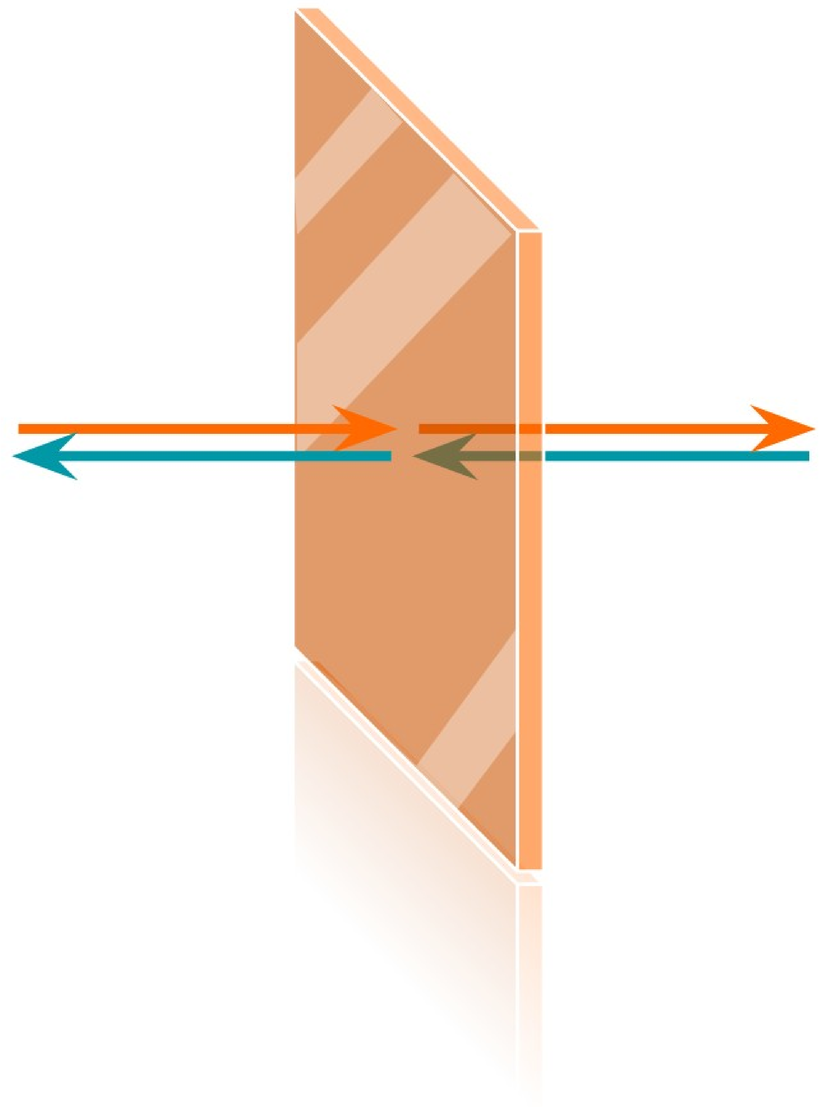
Question Generation

Question Generation
Relation with QA

question answering (QA) →

question generation (QG) →

Context



Question

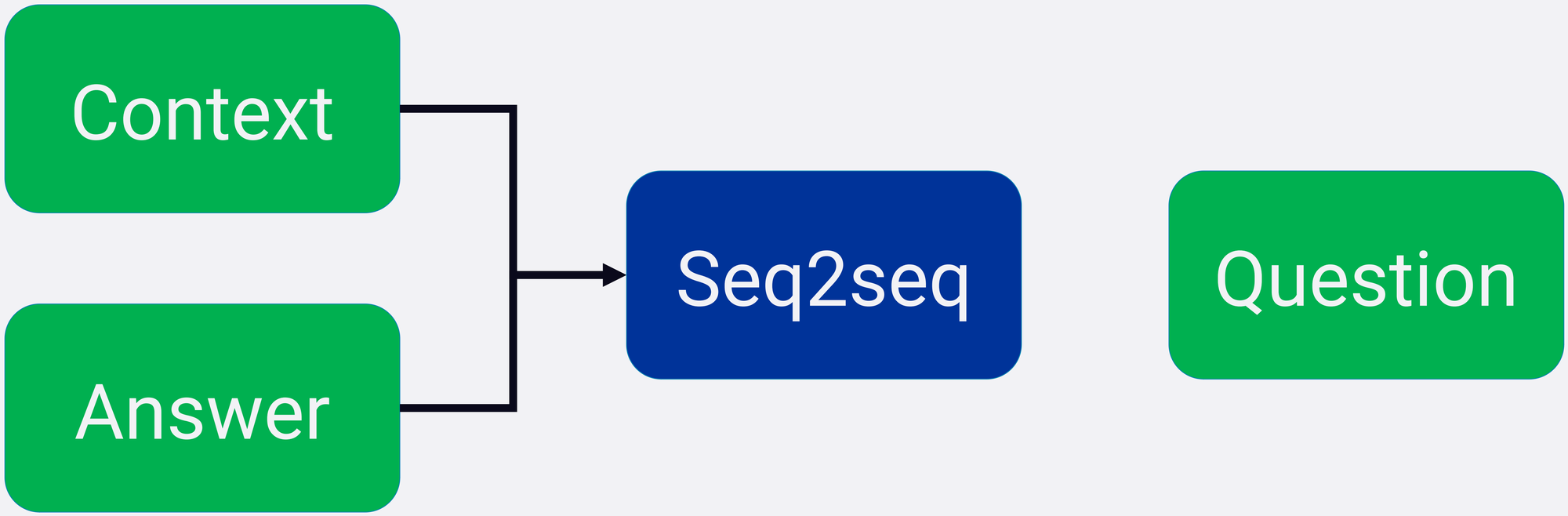
What is the function of lines of longitude and latitude?

Answer

allow us to find an absolute location of a point on earth.

Question Generation

Given a **context** and an **answer**, the model infers a **question**.



Question Generation
Models

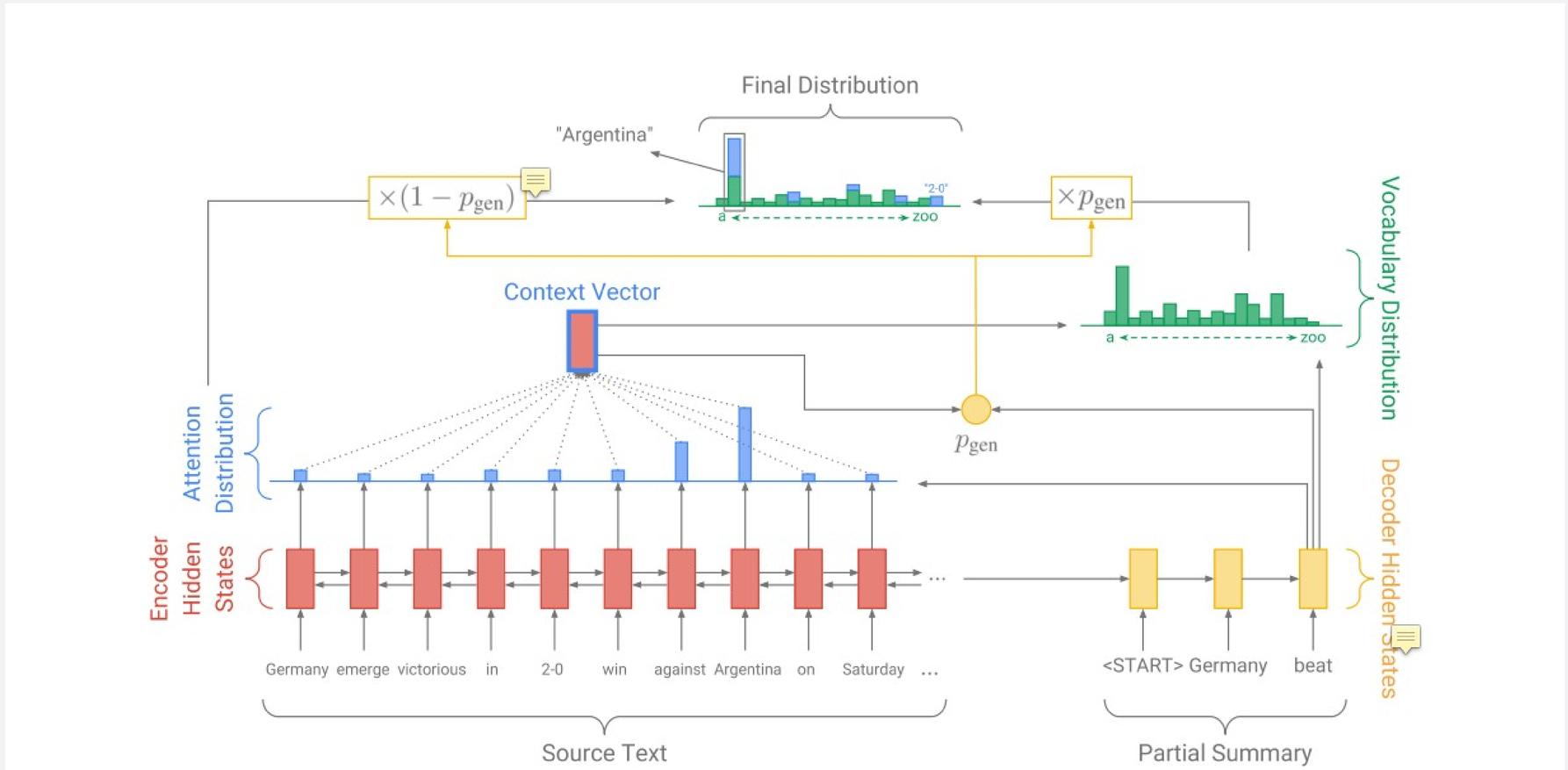


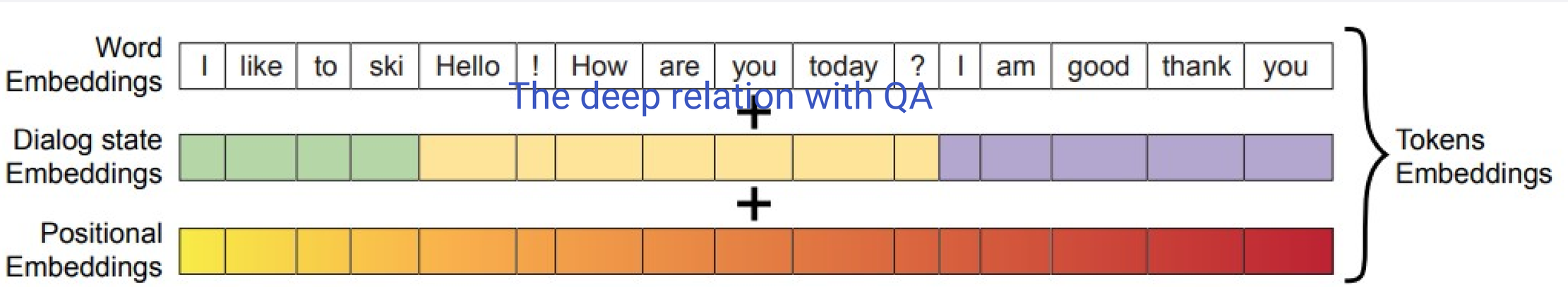
Figure 3: Pointer-generator model. For each decoder timestep a generation probability $p_{\text{gen}} \in [0, 1]$ is calculated, which weights the probability of *generating* words from the vocabulary, versus *copying* words from the source text. The vocabulary distribution and the attention distribution are weighted and summed to obtain the final distribution, from which we make our prediction. Note that out-of-vocabulary article words such as *2-0* are included in the final distribution. Best viewed in color.

See, A., Liu, P. J., & Manning, C. D. (2017). Get to the point: Summarization with pointer-generator networks. *arXiv preprint arXiv:1704.04368*.

Question Generation

Futur models ?

A Transfer Learning Approach with Positional Answer Embeddings similar to Wolf & al ?



The deep relation with QA

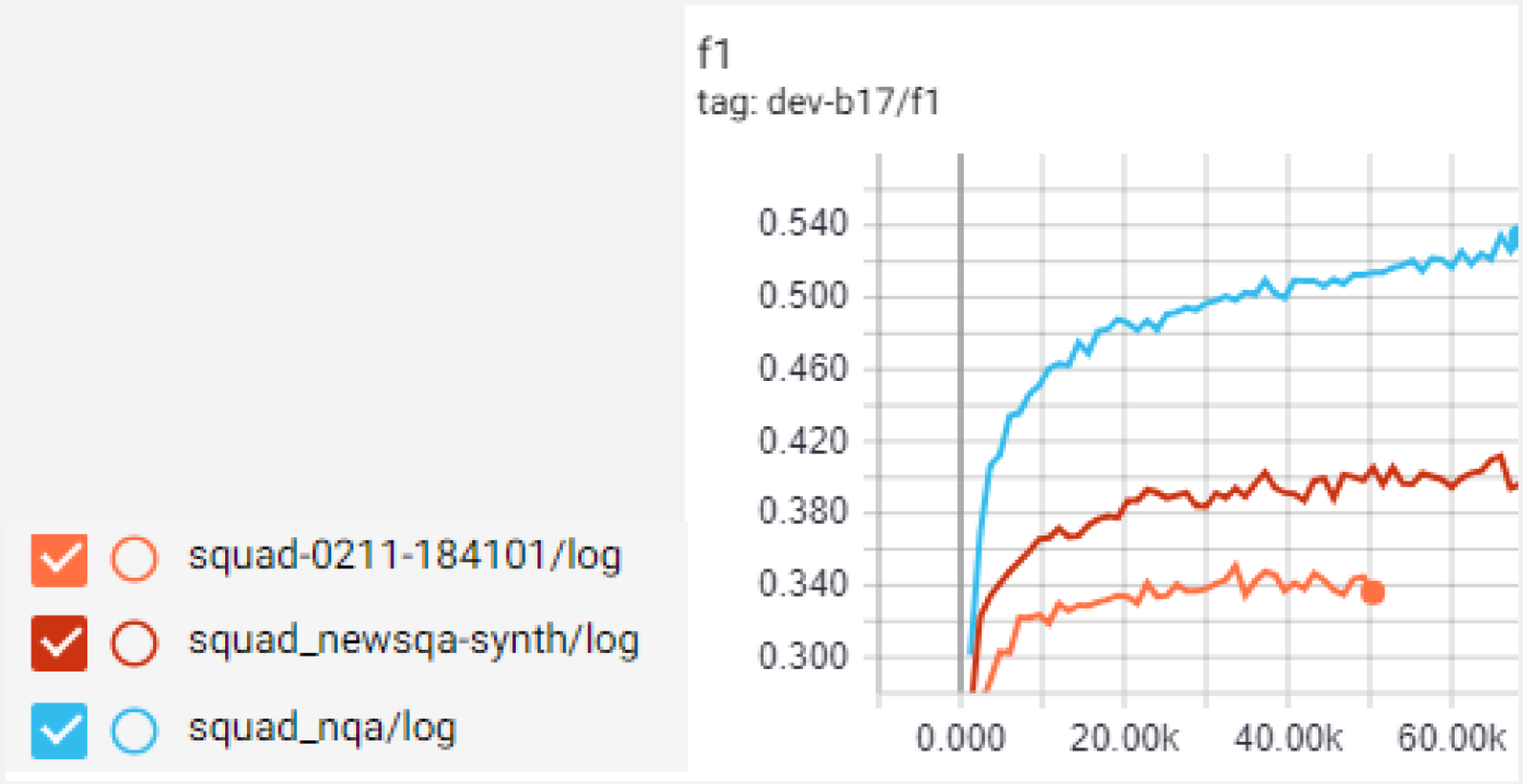
Figure 1: TranferTransfo's input representation. Each token embedding is the sum of a word embedding, a dialog state embedding and a positional embedding.

Wolf, T., Sanh, V., Chaumond, J., & Delangue, C. (2019). TransferTransfo: A Transfer Learning Approach for Neural Network Based Conversational Agents. *arXiv preprint arXiv:1901.08149*.

Question Generation for RC Transfer Learning

Question Generation
Synthetic QA Dataset

QA models does not perform well out of their domain. Generating synthetic dataset help adapting QA to the new domain ！



In progress :

- where does the improvement come from ?
- do we observe similar behavior with BERT ?

Question Generation with an QA metric

Question Generation
Proxy for evaluation

- BLEU doesn’t work well…

- QA probability seems to be a **good proxy:**

    - To know whether a question can be answered correctly given the context document (Yuan & al. Machine comprehension by text-to-text neural question generation)

    - To measure the difficulty of a question (Gao & al. Difficulty controllable question generation for reading comprehension)

QA score does not depend on a human annotated data. This is rare in NLP, on the contrary of BLEU or ROUGE metrics for instance.

How to leverage on it ? At training time ? At inference time ?

Question Generation

Proxy for evaluation

We want to generate questions with a high QA score.

RL and Beam Search are highly related here :

Question Generation

Proxy for evaluation

We want to generate questions with a high QA score.

RL and Beam Search are highly related here :

- **RL policy during training**: Yuan & al propose to fine tune the QG model with REINFORCE algorithm on QA score

	NLL	BLEU	F1	QA	PPL
Seq2Seq	45.8	4.9	31.2	45.6	153.2
Our System	**35.3**	10.2	39.5	65.3	175.7
+ PG (R_{PPL})	35.7	9.2	38.2	61.1	**155.6**
+ PG (R_{QA})	39.8	**10.5**	**40.1**	**74.2**	300.9
+ PG ($R_{\text{PPL+QA}}$)	39.0	9.2	37.8	70.2	183.1
Question LM	-	-	-	-	87.7
MPCM	-	-	-	70.5	-

Question Generation
Proxy for evaluation

We want to generate questions with a high QA score.

RL and Beam Search are highly related here :

- **RL policy during training**: Yuan & al propose to fine tune the QG model with REINFORCE algorithm on QA score
- **brute force at inference** : we could also in theory (of course just in theory!) generate all the possible sequences and sorted them by the QA score. It is similar to a beam search with number of beam equal to the vocabulary size power sequence length.

  Beam Search with decent beam size on a well trained model is a good solution

In progress : can we optimize smarter the beam search policy ?

Merci !