



Language Understanding

10 - Paraphrasing and Question
Answering

Hossein Zeinali

What Is Paraphrasing?

- **Paraphrases**

- Sentences or phrases that convey approximately the same meaning using different words.

- **Paraphrasing** is the natural language processing task of detecting and generating paraphrases.

Words	Phrases	Sentences
fortunate lucky favorable advantageous opportune timely convenient felicitous	for example one such example an instance in particular including specifically however more concretely	The parrot is dead It's not pinin, it's passed on! This parrot is no more! It has ceased to be! It's gone to meet its maker! It's a late parrot! It's a stiff!



Why Should we Care?

- Recognizing and generating paraphrases is an important component in many natural language processing applications:
 - Question Answering
 - MT Evaluation
 - Semantic Parsing
 - Simplification
 - Summarization
 - Grammar correction
 - Plagiarism detection
 - Evaluation of NLG

buy echo dot



I'd like to buy echo dot

buy me that mini echo thing



hey, can I purchase echo dot

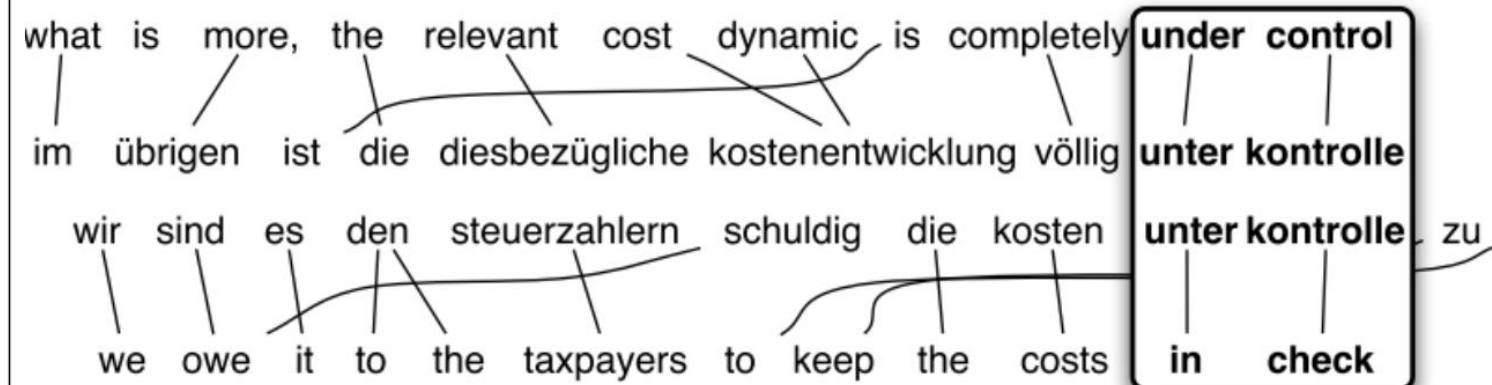


Traditional Methods

- Multiple sequence alignment:
 - Generate paraphrases through the usage of monolingual parallel corpora
 - News articles covering the same event on the same day
 - Training consists of using multi-sequence alignment to generate sentence-level paraphrases from an unannotated corpus.
 - Clustering similar sentences together.
 - Recurring patterns are found within clusters by using multi-sequence alignment.
- Phrase-based Machine Translation:
 - Aligning phrases in a **pivot language** to produce potential paraphrases in the original language
 - Finding sentences which have the same translation in the second language



Traditional Methods



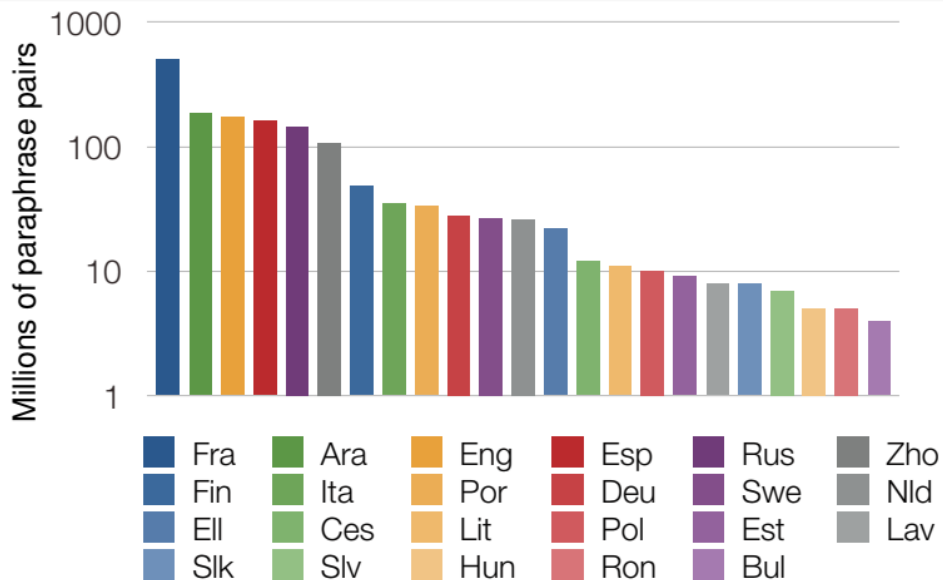
Source: Bannard and Callison-Burch (2005)

- No direct path from en → en; indirect path from en → de and de → en
- If two English phrases are aligned with the **same phrase** in another language they are likely to be a paraphrase.



PPDB: The Paraphrase Database

- PPDB is an automatically extracted database containing millions of paraphrases in 16 different languages.



Source: Ganitkevitch and Callison-Burch, 2014



NMT and Multiple Pivots

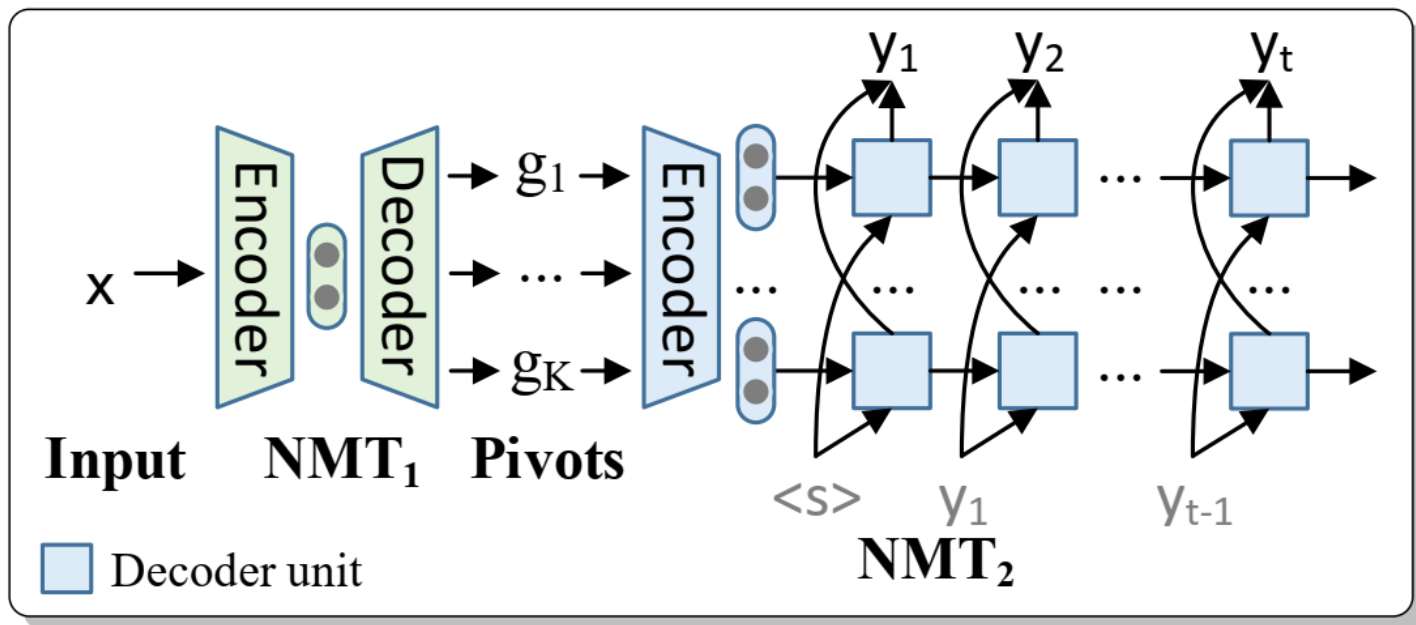
- Neural machine translation
 - **Context-aware**, places greater emphasis on semantics;
 - Allows for **multiple pivots** with multiple models;
 - Assigns a probability to **arbitrary sequences**.
- Relevance to paraphrases
 - Represent paraphrases in **continuous** space;
 - NMT **learns** embeddings for words and phrases **naturally**;
 - Estimate **semantic relatedness** between text segments of arbitrary length;
 - **Generate** candidate paraphrases for any source input.

Source: Paraphrasing Revisited with Neural Machine Translation

AUT, Language Understanding Course, Fall 2022, Hossein Zeinali



ParaNet



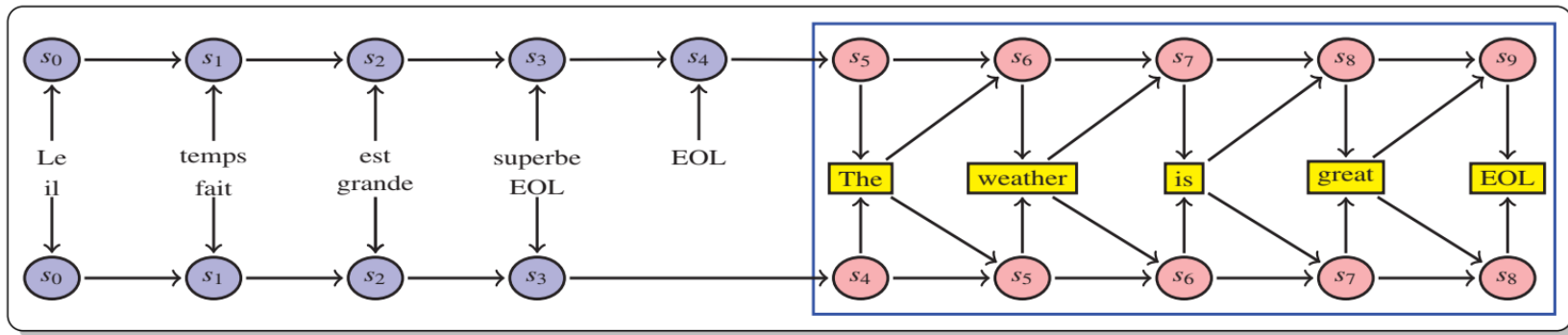
- Source is translated into K -best list of foreign pivots

Source: Paraphrasing Revisited with Neural Machine Translation

AUT, Language Understanding Course, Fall 2022, Hossein Zeinali



ParaNet



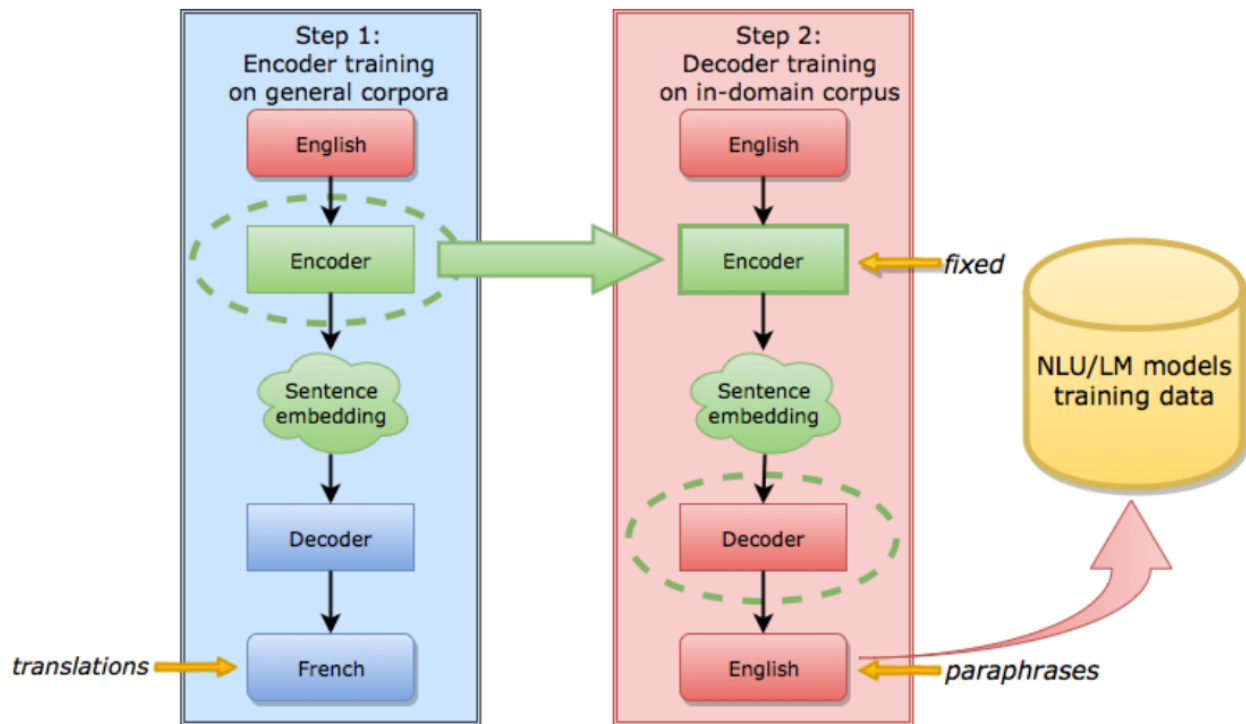
- Late-weighted combination
- Pivot sentences are encoded separately by two encoders
- At each time step, two decoders produce a probability distribution over all words, which are then combined (in the yellow square)
- From this combined distribution a word is chosen, which is then given as input to each decoder.

Source: Paraphrasing Revisited with Neural Machine Translation

AUT, Language Understanding Course, Fall 2022, Hossein Zeinali



Paraphrase Generation



Source: Neural Machine Translation For Paraphrase Generation

AUT, Language Understanding Course, Fall 2022, Hossein Zeinali



Paraphrase Generation

- Training of the paraphraser model is done in two stages
 - Training encoder on large out-of-domain parallel corpora
 - Then fixing the encoder and retraining the decoder on smaller in-domain data.
- Once the model is trained, multiple paraphrases can be generated for each given input example by using n-best decoding.
- It is important for model training that the out-of-domain corpora are multilingual parallel datasets.
 - It has much better chance to learn the sentence true semantic representation

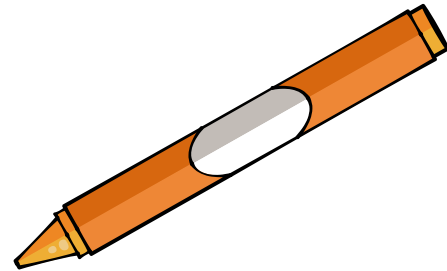
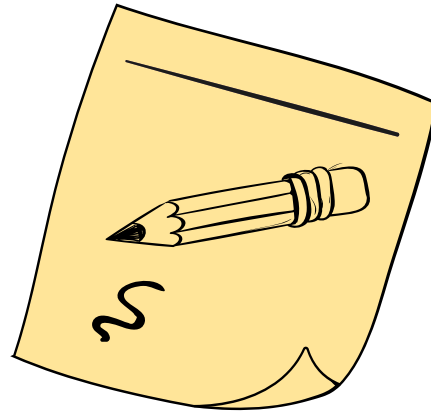
Source: Neural Machine Translation For Paraphrase Generation

AUT, Language Understanding Course, Fall 2022, Hossein Zeinali





Question Answering



Question Answering (QA)

- Earlier systems used the two major paradigms of question answering
 - **Information retrieval-based**
 - Relies on the vast quantities of textual information on the web or in other collections.
 - **Knowledge-based**
 - A system builds a semantic representation (e.g. logical form) of the query, then it is used to query databases of facts.
- In 2011, IBM's Watson question-answering system won the TV game-show *Jeopardy!*
- Most question answering systems focus on **factoid questions**
 - Questions that can be answered with simple facts expressed in short texts.



People Want to Ask Questions ...

- Examples from Ask.com query log
 - how much should I weight
 - what does my name mean
 - how to get pregnant
 - where can I find pictures of hairstyles
 - who is the richest man in the world
 - what is the meaning of life
 - why is the sky blue
 - what is the difference between white eggs and brown eggs
 - can you drink milk after the expiration date
 - what is true love
 - what is the Jonas brothers address
- Around 10-20% of query logs



Why QA?

- QA engines attempt to let you ask your question the way you'd normally ask it.
- More specific than short keyword queries
 - Orange chicken
 - What is orange chicken
 - How to make orange chicken
- Inexperienced search users
- Difficult questions:
 - Question: *How much money did IBM spend on advertising in 2006?*
 - No engine can answer



General Search Engines

- Include question words etc. in stop-list with standard IR
- Sometime it works. Sometime it requires users to do more investigation.
 - Question: *Who was the prime minister of Australia during the Great Depression?*
 - Answer: *James Scullin (Labor) 1929–31.*
 - *Ask.com gives an explicit answer.*
 - *Google's top 1-2 results are also good.*
 - *What is phone number for united airlines*
 - *Ask.com gives a direct answer*
 - *Google gives no direct answers in top 10.*





IR-based Factoid Question Answering



IR-based Factoid Question Answering

- Answer a user's question by finding short text segments on the web or some other collection of documents.

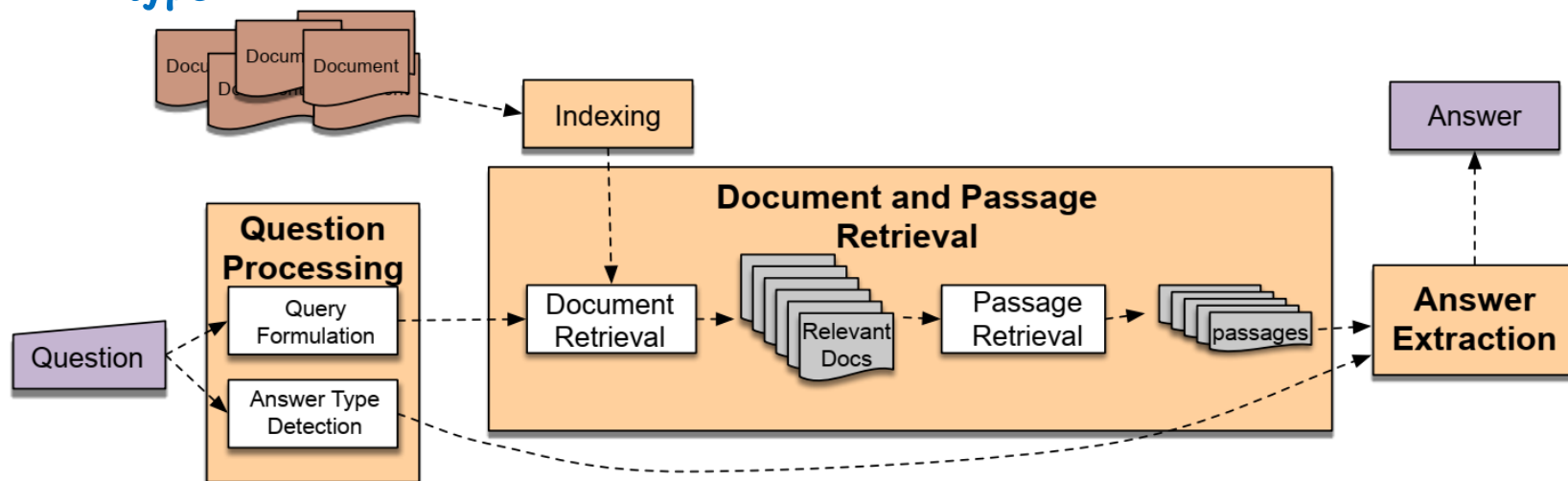
Question	Answer
Where is the Louvre Museum located?	in Paris, France
What's the abbreviation for limited partnership?	L.P.
What are the names of Odin's ravens?	Huginn and Muninn
What currency is used in China?	the yuan
What kind of nuts are used in marzipan?	almonds
What instrument does Max Roach play?	drums
What's the official language of Algeria?	Arabic
How many pounds are there in a stone?	14



IR-based Factoid Question Answering

- **Question Processing**

- Extract the **query**: the keywords passed to the IR system to match potential documents.
- Also extract further information such as: **answer type**, **focus**, **question type**



IR-based Factoid Question Answering

- **Query Formulation:**
 - Creating a query to send to an information retrieval system
- **Query reformulation:**
 - Applying some rules on the query to rephrase the question to make it look like a substring of possible declarative answers.
 - *“when was the laser invented?” => “the laser was invented”*
- **Document and Passage Retrieval:**
 - Sending the query to an IR engine, and resulting in a set of documents ranked by their relevance to the query.
 - **Passage retrieval** is then used to filtering the passages/documents
- **Answer Extraction:**
 - Extract a specific answer from the passage to a question

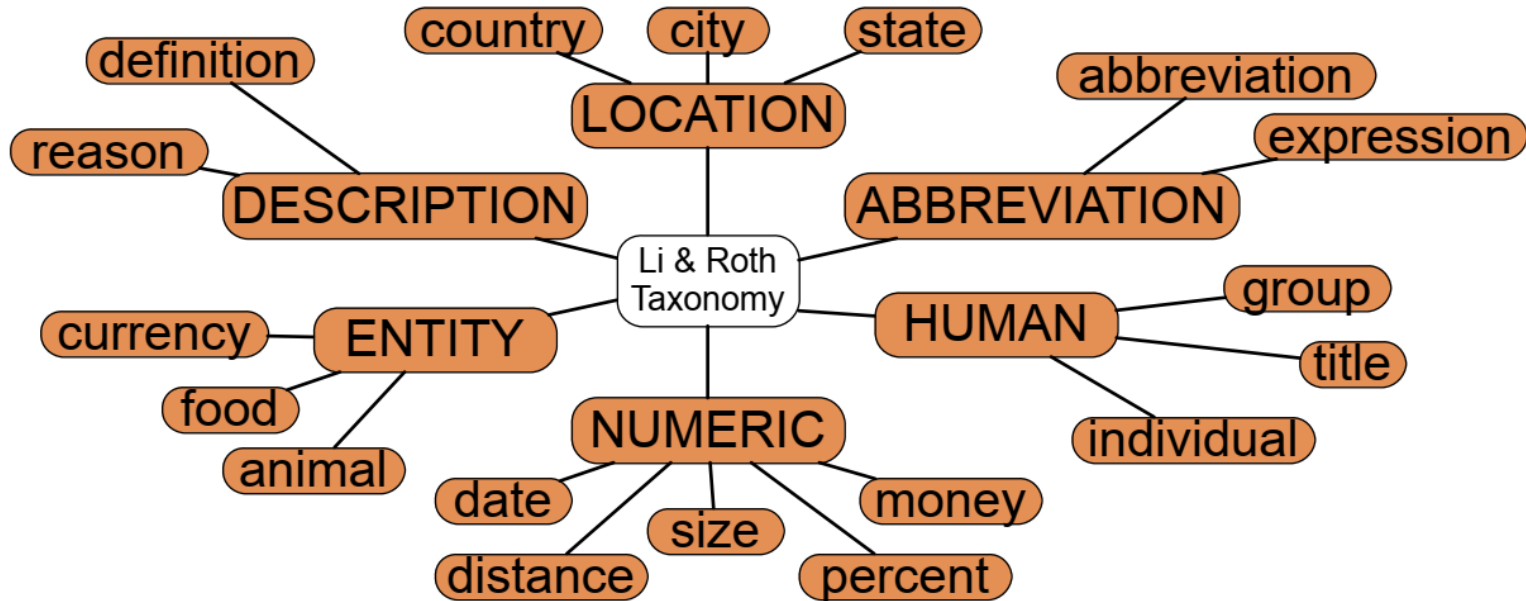


Question/Answer Types

Class 1	Answer: single datum or list of items C: who, when, where, how (old, much, large)
Class 2	A: multi-sentence C: extract from multiple sentences
Class 3	A: across several texts C: comparative/contrastive
Class 4	A: an analysis of retrieved information C: synthesized coherently from several retrieved fragments
Class 5	A: result of reasoning C: word/domain knowledge and common sense reasoning



Question/Answer Types



Neural Answer Extraction

- Motivation: a question and its answer are semantically similar in some appropriate way.
 - Is it possible to computing an embedding for the question and an embedding for each token of the passage, and then selecting passage spans whose embeddings are closest to the question embedding?
- **Reading Comprehension**
 - Given a passage, machine must answer questions about it.
 - Has become a useful way to measure natural language understanding performance.



Reading Comprehension Datasets

- Stanford Question Answering Dataset (SQuAD):
 - Passages from Wikipedia and associated questions whose answers are spans from the passage, as well as some questions that are designed to be unanswerable.
 - Over 150,000 questions.
- NewsQA dataset:
 - Consists of 100,000 question-answer pairs from CNN news articles.
- WikiQA dataset:
 - The span is the entire sentence containing the answer.
 - Can be used for **sentence selection** task
 - Choosing a sentence rather than a smaller answer span.



A bi-LSTM-based Algorithm

- Given a question q of l tokens q_1, \dots, q_l and a passage p of m tokens p_1, \dots, p_m :
 - For each token p_i , compute the probabilities:
 - $p_{start}(i)$ that p_i is the start of the answer span
 - $p_{end}(i)$ that p_i is the end of the answer span
- The question is represented by a single embedding q , which is a weighted sum of representations for each question word q_i .
 - Attention mechanism can be used instead of weighting.

$$q = \sum_i b_i q_i$$
$$b_j = \frac{\exp(\mathbf{w} \cdot \mathbf{q}_i)}{\sum_k \exp(\mathbf{w} \cdot \mathbf{q}_k)}$$



A bi-LSTM-based Algorithm

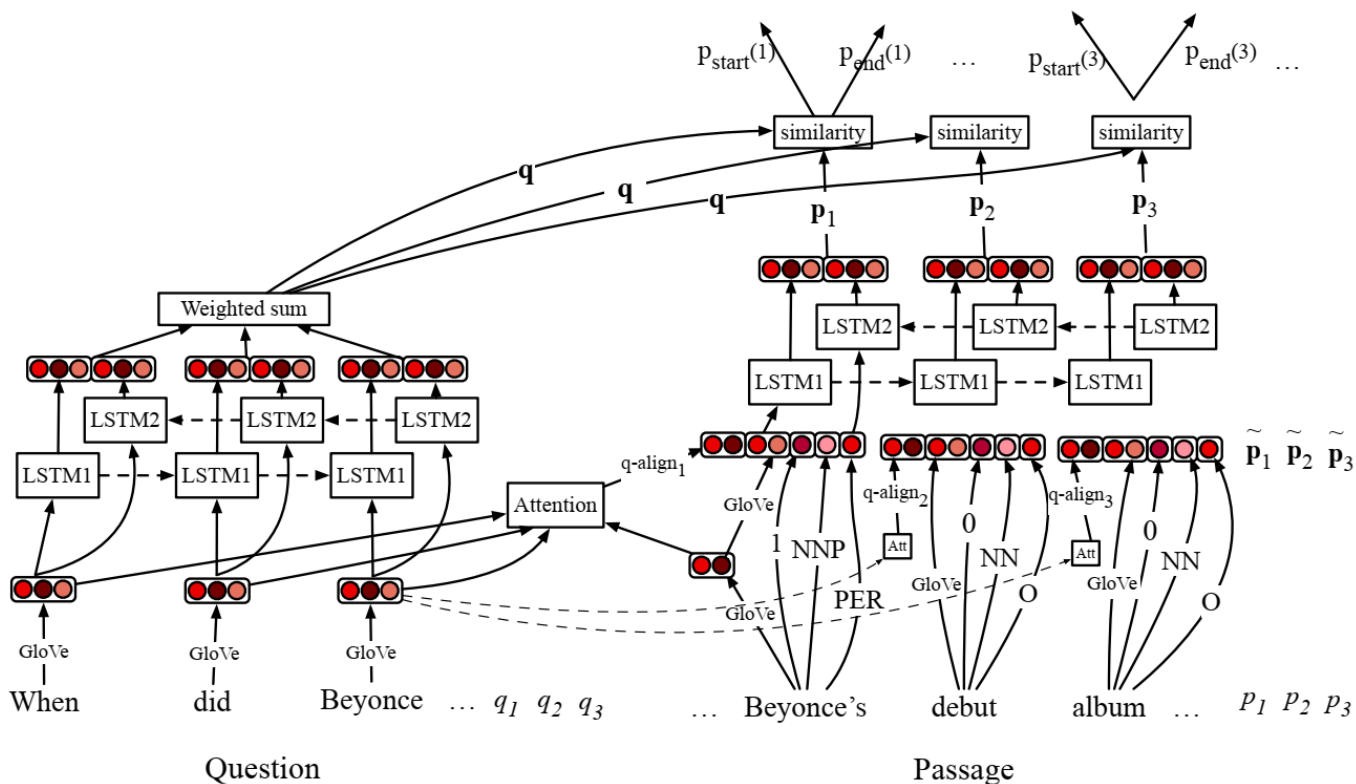
- The input representation \tilde{p} is concatenation of four components:
 - An embedding for each word $E(p_i)$
 - Token features like the part of speech of p_i , or the named entity tag of p_i
 - Exact match features representing whether the passage word p_i occurred in the question: $1(p_i \in q)$.
 - Aligned question embedding: In addition to the exact match features, many QA systems use an attention mechanism to give a more sophisticated model of similarity between the passage and question words.
- Then the input representation \tilde{p} is passed through a bi-LSTM
- Finally two separate classifiers are trained to compute $p_{start}(i)$ and $p_{end}(i)$ separately.

$$p_{start}(i) \propto \exp(\mathbf{p}_i \mathbf{W}_s \mathbf{q})$$

$$p_{end}(i) \propto \exp(\mathbf{p}_i \mathbf{W}_e \mathbf{q})$$



A bi-LSTM-based Algorithm



BERT-based Question Answering

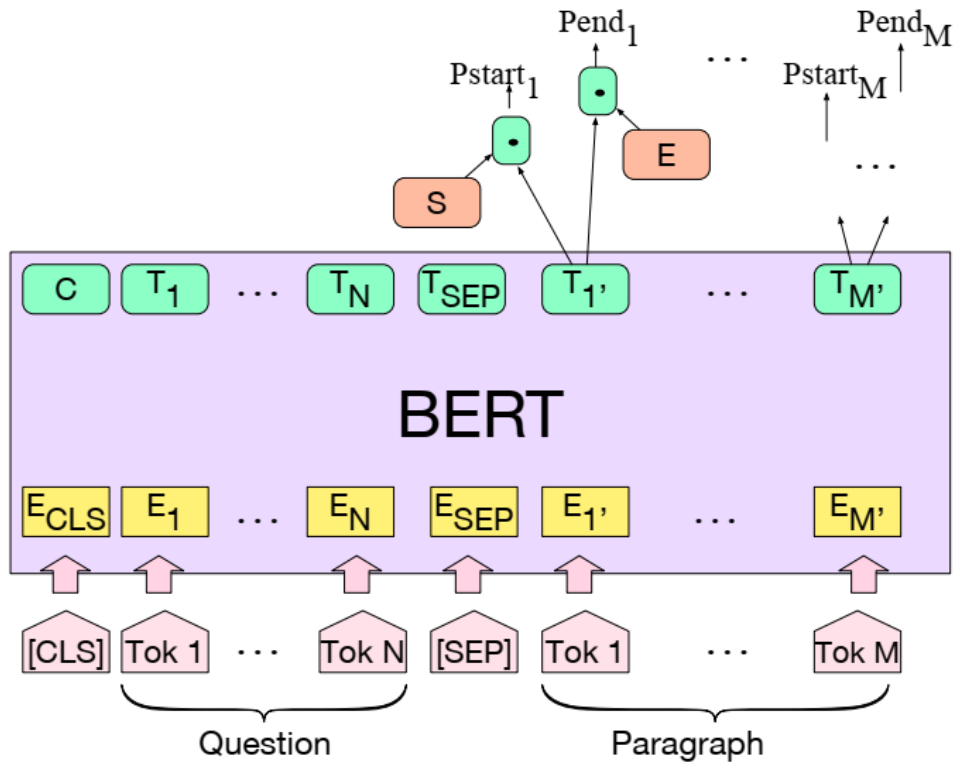
- For span-based question answering, the question is represented as the first sequence and, the paragraph as the second sequence.
- Two new embeddings should be added to the output head: a span-start embedding S and a span-end embedding E .
- The span-start and span-end probabilities are calculated using:

$$p_{start_i} = \frac{\exp(S \cdot T_i)}{\sum_k \exp(S \cdot T_k)} \quad \text{and} \quad p_{end_i} = \frac{\exp(E \cdot T_i)}{\sum_k \exp(E \cdot T_k)}$$

- The score of a candidate span from position i to j is $S \cdot T_i + E \cdot T_j$, and the highest scoring span in which $j \geq i$ is chosen as the model prediction.
- The training objective for fine-tuning is the sum of the log-likelihoods of the correct start and end positions for each observation.



BERT-based Question Answering





Knowledge-based Question Answering



Knowledge-based Question Answering

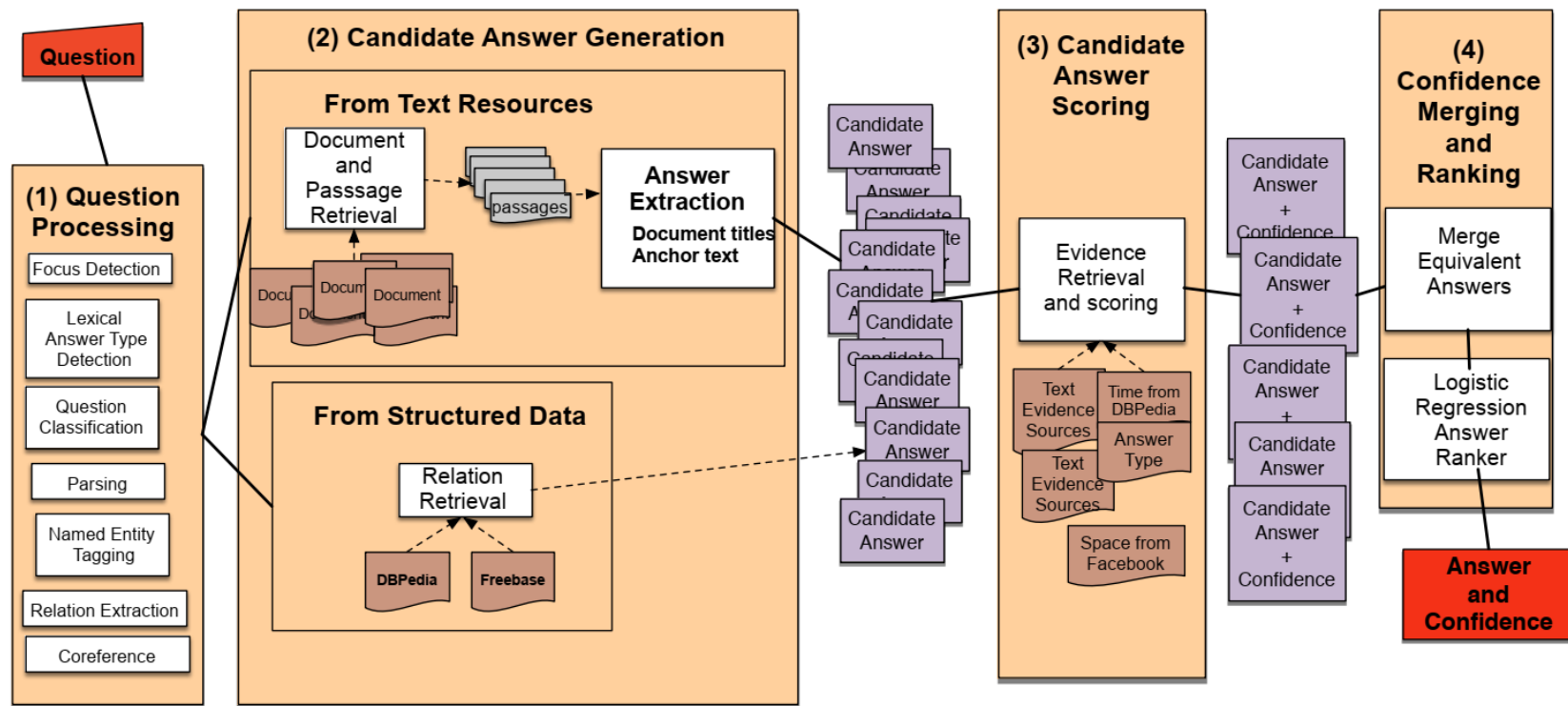
- Answering a natural language question by mapping it to a query over a structured database.
- We saw that the systems for mapping from a text string to any logical form are **semantic parsers**.
- The logical form of the question is thus either in the form of a query or can easily be converted into one.
 - E.g. SQL query.
- **RDF triples** is a 3-tuple, a predicate with two arguments, expressing some simple relation or proposition.

- E.g. Freebase or DBpedia

subject	predicate	object
Ada Lovelace	birth-year	1815



IBM's Watson



IBM's Watson

- Won the Jeopardy challenge in 2011
- There are 4 stages in the DeepQA system:
 - **Question processing:**
 - Runs parsing, named entity tagging, and relation extraction (coreference) on the question.
 - Extracts the **focus**, the **answer type** and performs **question classification** and **question sectioning**.
 - **Candidate answer generation:**
 - Combine the processed question with external documents and other knowledge sources to suggest many candidate answers.
 - **Candidate answer scoring:**
 - uses many sources of evidence such as the lexical answer type to score the candidates.
 - **Answer merging and scoring:**
 - First merges candidate answers that are equivalent and then assigns a confidence value to each candidate answer.



Evaluation of Factoid Answers

- A common evaluation metric for factoid question answering is mean reciprocal rank (MRR)
 - Assumes a test set of MRR questions that have been human-labeled with correct answers.
 - Also assumes that systems are returning a short ranked list of answers or passages containing answers.
 - Each question is then scored according to the reciprocal of the **rank** of the first correct answer.
 - Questions with return sets that do not contain any correct answers are assigned a zero.



$$\text{MRR} = \frac{1}{N} \sum_{i=1 \text{ s.t. } \text{rank}_i \neq 0}^N \frac{1}{\text{rank}_i}$$



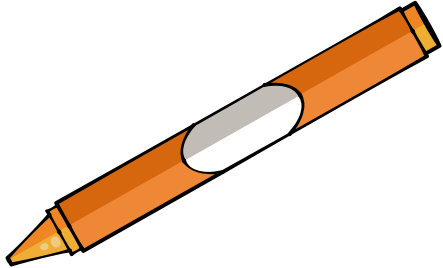
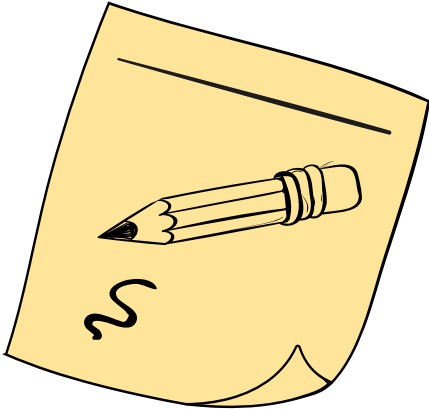

Evaluation of Factoid Answers

- For reading comprehension systems:
 - **Exact match:** The percentage of predicted answers that match the gold answer exactly.
 - **F_1 score:** The average overlap between predicted and gold answers. Treat the prediction and gold as a bag of tokens, and compute F_1 , averaging the F_1 over all questions.
- Multiple choice questions, with the task being to choose among the given answers.
 - Accuracy can be used for evaluation





Spoken Question Answering



Spoken Question Answering

- Spoken question answering implies doing the search in spoken data and/or from spoken questions.
- The structure of spoken language is different from that of written language.
 - Some of the anchor points used in processing such as punctuation must be inferred and are therefore error-prone.
 - It is also necessary to deal with spoken-language phenomena including disfluencies, repetitions, restarts, and corrections.
 - An additional challenge is dealing with the recognition errors.
 - Spoken questions are not necessary well formed.
- Question-answering in spoken data collections means that the answer has to be found in the audio data.



Aspects of Spoken QA

- The specific aspects of handling speech in QA systems:
 - Developing or adapting methods, algorithms and tools to be efficient on transcribed speech.
 - Enhancing these approaches to handle, or at least be robust to, errors produced by automatic speech recognition systems.
- Working with spoken rather than with written language:
 - Transcribed speech is structurally different from written texts
 - Lack of punctuations
 - Can use prosody or directly convert speech to meaning
 - Handling the output of automatic speech recognition systems
 - Some errors are critical and others not important.
- Handling the difficulties of transcribed speech:
 - When analyzing the documents: indexing time
 - Take into account the structural specificities of speech
 - When matching documents and questions: information retrieval time
 - Handling and compensating for speech recognition errors



QA Evaluation Campaigns

- The top block of this table focuses on question type; the second block concerns the data collection and the last one concerns miscellaneous aspects.
- **Document Types:**
 - Spoken Documents and Spoken Questions
 - Written

	TREC							QA@Clef Main Track					QAsT			NTCIR				
	1	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
	9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	9	0	1	2	3	4	5	6	7	3	4	5	6	7	8	9	7	8	9	2
Factual
Simple definition		
Definitions				
Why																			.	
How										.								.		
Yes/no																				
Open lists			
Closed lists			◇	◇					
Follow-up			◇		◇	.	
Topics																
Information																
Spoken questions													.							
Newspapers
Speech													.	.	.					
Law													.							
Wikipedia												.	.							
Blogs						.														
Question class given					
Multiple answers
Long answers									
Justification																
Translingual													◇	◇
Parallel docs.										.										
Temporal restriction																
Timecodes														.	.					

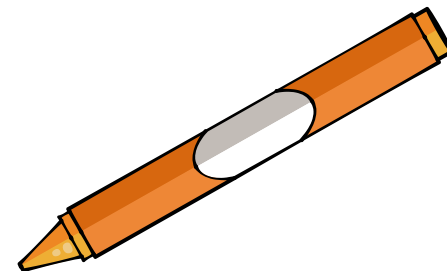
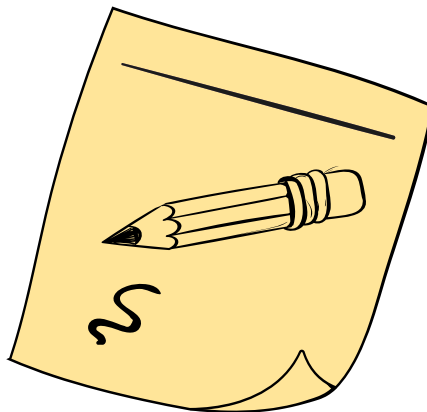
QA Question Types

- **Factual questions:** the simplest and most frequent question type.
 - The answer can be a single word or a multi-word expression, often is a named entity.
 - **List questions:** the factual questions for which the answer is a list of elements.
 - Closed-list questions give the number of expected elements vs open list.
- **Definition questions:** any kind of answer (named entity or not, simple word, phrase, complete sentence, etc.) is possible.
- **Complex questions:** types such as *how*, *why*

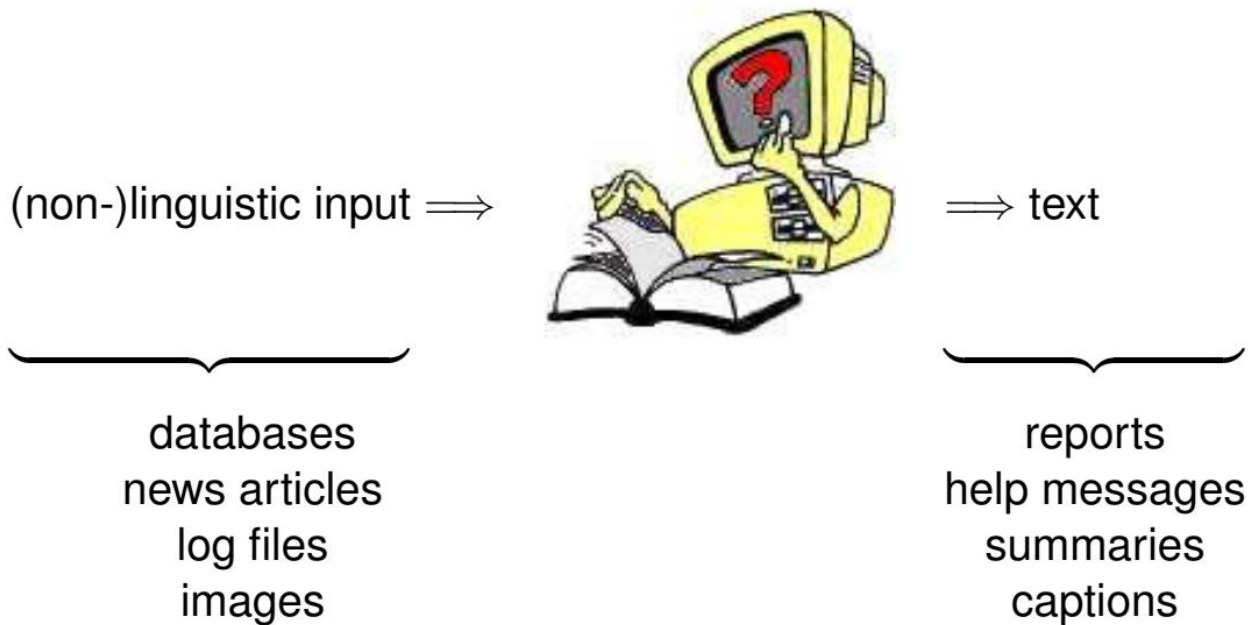




Data-to-Text Generation



What is NLG?



- We will look at how to generate a document from a database.

An Example: Basketball

TEAM	WIN	LOSS	PTS	FG_PCT	RB	AST
Nuggets	8	15	85	41	45	24
Wizards	8	13	92	45	35	21

PLAYER	H/V	AST	RB	PTS	FG	CITY
Bradley Beal	H	5	3	26	8	Washington
John Wall	H	5	7	15	5	Washington
Markieff Morris	H	4	3	15	5	Washington
Nikola Jokic	V	1	11	17	6	Denver
Jusuf Nurkic	V	1	7	13	6	Denver
...

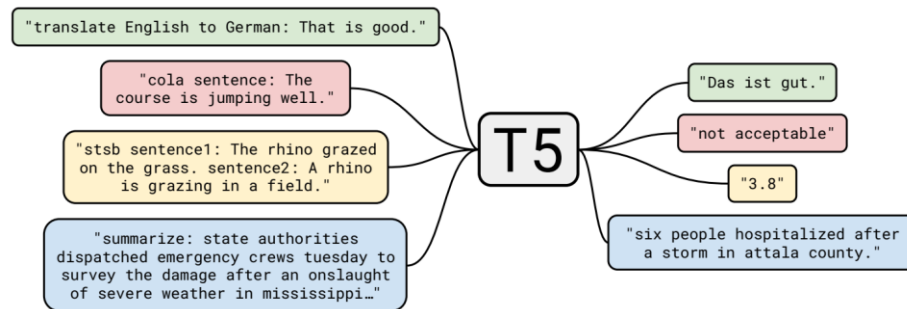
PTS: points, FT_PCT: free throw percentage,
RB: rebounds, AST: assists, H/V: home or vis-
iting, FG: field goals, CITY: player team city.

These two teams have really struggled this season and the 48 combined turnovers shows just that. **Denver** was the more inept team though, as they set an NBA season-high with 29 turnovers of their own. The shooting wasn't any better, as they shot **41** percent from the field, **67** percent from the free-throw line and **22** percent from three-point range. **Bradley Beal** was the best player in this game, tallying **26** points, **three** rebounds and **five** assists. **John Wall** went just **5-of-14** from the field, but still finished with **15** points, **seven** rebounds, **five** assists and **three** steals. **Markieff Morris** also scored **15** points, as all five starters played at least 33 minutes. **Nikola Jokic** was the only effective Nugget on offense, finishing with **17** points and **11** rebounds. **Wilson Chandler** had a game to forget, as he went just **2-of-16** from the field, en route to five points.



Text-to-Text Transfer Transformer (T5)

- Treating every problem as a text-to-text task
 - The model is based on the Transformer architecture
- Main differences with the original topology:
 - Removing the Layer Norm bias
 - Placing the layer normalization outside the residual path
 - Using relative position embeddings instead of fixed position embeddings



Source: Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer

AUT, Language Understanding Course, Fall 2022, Hossein Zeinali



Text-to-Text Transfer Transformer (T5)

- Training dataset:
 - Colossal Clean Crawled Corpus (C4)
 - Is reasonably clean and natural English text
 - About 750 GB text
- Downstream Tasks
 - Machine translation, question answering, abstractive summarization, and text classification
- Train a single model on the diverse set of above tasks
 - Cast all of the tasks we consider into a “text-to-text” format
 - A task where the model is fed some text for context or conditioning and is then asked to produce some output text.
 - Provides a consistent training objective both for pre-training and fine-tuning
 - Add a task-specific (text) prefix to the original input sequence before feeding it to the model.

Source: Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer

AUT, Language Understanding Course, Fall 2022, Hossein Zeinali



Data-to-Text Using T5

- Using T5 pre-trained models and fine-tune them to the task
- Fine-tuning
 - The data-to-text task is cast in the text-to-text framework by representing the structured data as a flat string (linearization).
 - Fine-tune T5 on the data-to-text corpus for a small number of steps.

Dataset	Train	Dev	Test
WebNLG	18.1K	2268	4928
ToTTo	120K	7700	7700
Multiwoz	56.8K	7374	7372

Source: Text-to-Text Pre-Training for Data-to-Text Tasks

AUT, Language Understanding Course, Fall 2022, Hossein Zeinali

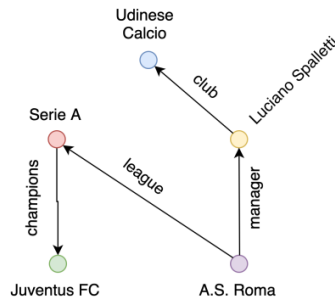


Data-to-Text: Datasets

- **ToTTo** consists of Wikipedia tables paired with natural language descriptions. The input is a table with a subset of cells highlighted. A model must generate text that describes the highlighted cells
- **MultiWoz** is a corpus of 10K human-human dialogs for developing task oriented dialogue systems. For the NLG task, a meaning representation encapsulating system actions must be verbalized into natural language response.
- **WebNLG** where the task is to convert a graph of subject-object-predicate triples into a textual description.



Data-to-Text: Datasets



<S> Serie A <P> champions <O> Juventus F.C.
 <S> Luciano Spalletti <P> club <O> Udinese Calcio
 <S> A.S. Roma <P> manager <O> Luciano Spalletti
 <S> A.S. Roma <P> league <O> Serie A

AS Roma play in the Serie A league where Juventus FC are the champions. Their manager is Luciano Spalletti who has been associated with Udinese Calcio.

Domain	train
Inform	arrive_by : 11:51
Request	num_people

train inform arriveby = 11:51 | train request people = ?

The closest arrival time i can give you is 11:51 , is that ok ? And how many tickets would you like ?

Table Title: Cristhian Stuani				
Section Title: International goals				
No.	Date	Venue	Opponent	Result
2	13 November 2013	Amman International Stadium, Amman, Jordan	Jordan	5-0

<page_title> Cristhian Stuani </page_title>
 <section_title> International goals </section_title>
 <table> <cell> 2. <col_header> No. </col_header> </cell>
 <cell> 13 November 2013 <col_header> Date </col_header>
 </cell> <cell> Amman International Stadium, Amman, Jordan <col_header> Venue </col_header> </cell> <cell> Jordan <col_header> Opponent </col_header> </cell>
 <cell> 5-0 <col_header> Result </col_header> </cell>
 </table>

On 13 November 2013 Cristhian Stuani netted the second in a 5–0 win in Jordan.

Source: Text-to-Text Pre-Training for Data-to-Text Tasks

AUT, Language Understanding Course, Fall 2022, Hossein Zeinali



Results on WebNLG

Model	BLEU			METEOR		
	Overall	Seen	Unseen	Overall	Seen	Unseen
Melbourne [†]	45.1	54.5	33.3	0.37	0.41	0.33
GTR-LSTM [†]	37.1	54.0	29.2	0.31	0.37	0.28
Transformer	51.7	56.4	38.9	0.32	0.41	0.21
Step-by-Step [†]	47.4	53.3	34.4	0.39	0.44	0.34
PlanEnc	52.8	64.4	38.2	0.41	0.45	0.37
T5-Small	52.0	62.6	38.8	0.41	0.45	0.37
T5-Base	55.2	64.7	49.4	0.43	0.46	0.41
T5-Large	57.1	63.9	52.8	0.44	0.46	0.41
T5-3B	54.0	62.8	52.0	0.43	0.45	0.42

Source: Text-to-Text Pre-Training for Data-to-Text Tasks

AUT, Language Understanding Course, Fall 2022, Hossein Zeinali



Results on the ToTTo

Model	Overall		Overall Subset		Nonoverlap Subset	
	BLEU	PARENT	BLEU	PARENT	BLEU	PARENT
Content Planner	19.2	29.2	24.5	32.5	13.9	25.8
Pointer-Generator	41.6	51.6	50.6	58.0	32.2	45.2
BERT-to-BERT	44.0	52.6	52.7	58.4	34.8	46.7
T5-3B	49.5	58.4	57.5	62.6	41.4	54.2

Source: Text-to-Text Pre-Training for Data-to-Text Tasks

AUT, Language Understanding Course, Fall 2022, Hossein Zeinali



Results on Multiwoz

Model	BLEU	SER
HDSA	26.5	12.14
SC-GPT2	30.8	0.53
T5-Small	34.6	1.27
T5-Base	35.1	0.99
T5-Large	34.7	0.92
T5-3B	34.8	0.86

Source: Text-to-Text Pre-Training for Data-to-Text Tasks

AUT, Language Understanding Course, Fall 2022, Hossein Zeinali





Thanks for your attention



References and IP Notice

- Daniel Jurafsky and James H. Martin, “Speech and Language Processing”, 3rd ed., 2019
- Tur, Gokhan, and Renato De Mori. “Spoken language understanding”, 2011
- Some of the slides were selected from Mirella Lapata's slides.
- Some slides from Tao Yang's slides on QA.
- Some graphics were selected from [Slidesgo](#) template

