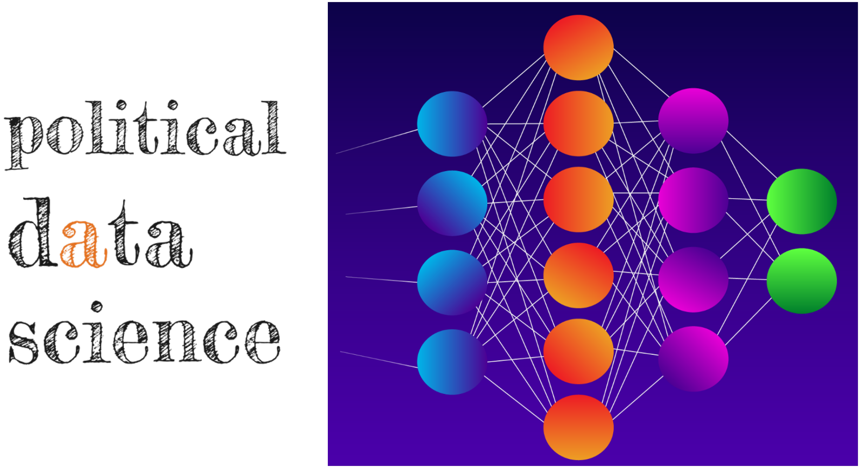
Applied Deep Learning for NLP

Week 6 - IE, Chatbots and QA

Juan Carlos Medina Serrano

Technische Universität München
Hochschule für Politik
Political Data Science

Munich, 26. November 2020



MetaTask: Information Extraction

Extract relevant information from text documents.

Information can mean multiple things, such as events, people, relationships, places ...

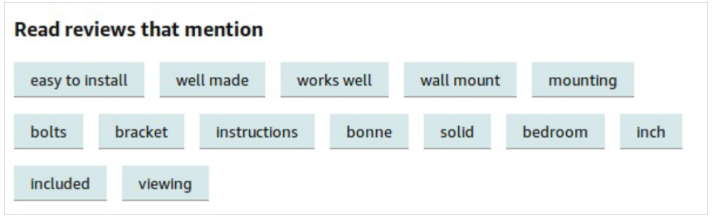
Extracting information that have a fixed pattern (for example, emails) is straightforward with regular expressions

Information extraction is a meta-task with several subtasks:

- Key phrase extraction
- Named entity recognition (NER)
- Named entity guation
- Named entity linking
- Relationship Extraction

Task: Key Phrase Extraction

Extracts important words and phrases that best describe the subject of a document



Two ways to solve this:

- Supervised learning. Requires large datasets of labeled text with corresponding keyphrases. Similar to text classification. Not the best approach
- Unsupervised learning. Text is divided in n-grams and these are represented as nodes in a weighted graph. Keyphrases are identified based on how connected they are to the rest of the graph. Select top-N important nodes. (Think of page rank)

Note: There may be overlapping keyphrases. A solution is to use a similarity measure and choose the ones that are most dissimilar to one another.

Task: Named Entity Recognition

Identify **entities** in a document.
Entities are typically names of personas, locations, organizations, and other specialized strings, such as money symbols and dates.

How to build a NER system?

- Mantain a large collection (database) of person/organizazion/names that are relevant and match the text.

- Rule-based: Based on patterns of POS tags.
  Example: The pattern *NNP was born*, indicates that the NNP is a person.

- Supervised learning. Use a labeled dataset. Similar to text classification BUT the context plays a big role. The classification task with context are often referred as sequence labeling problems.
  Example: *Is Washington the capital of the US?* Does Washington refer to the state or the person?

  Can be solved with classic NLP approaches (Hidden Markov Models) or DL approaches (RNNs, Transformers)

Task: Named Entity Disambiguation

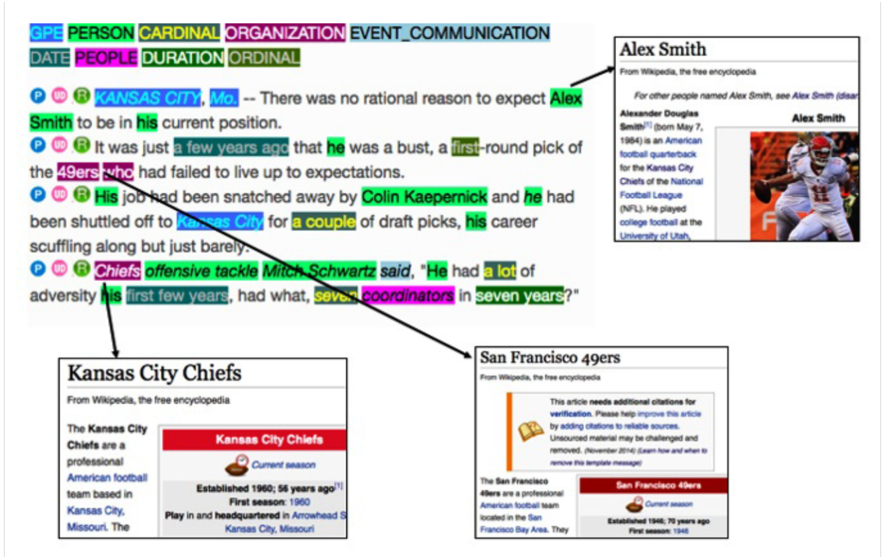
Assigns a unique identity to entities mentioned in the text.

Example: *Angela Merkel is the chancellor of Germany. Next year, Merkel will leave the office...The chancellor is meeting next week with the corona cabinet.*

After NER, NED can be performed using the POS tagging, dependency parsing and coreference resolution to figure out which entities are the same. Normally, pattern heuristics work well.

Task: Named Entity Linking

It links outside real-world information to the entities found by NER and NED. For example, from Wikipedia.



Uses large knowledge bases of connected entities, such as the Google Knowledge Graph. It is possible to build such a graph from Wikipedia. Needs advanced techniques of **Information Retrieval.**

MetaTask: Information Retrieval

It is the activity of obtaining information system resources that are relevant to an information need from a collection of those resources.

Different metatask than Information Extraction! Here you retrieve documents according to a search term (query).

Most common example: Google Search!

IR is outside of the scope of this seminar, but still many Deep Learning NLP techniques we learn here are used for search.

Task: Relationship Extraction

Find the relationships between entities.

Hardest subtask from Information Extraction since we needs NER, NED and NEL as preprocessing step + POS, dependency parsers and correference resolution.

Relations are specific to a given domain. For example, text about movies vs. texts about the human anatomy.

It requires two classification steps:

- Whether two entities are related (binary classification)
- If they are related, what is the relationship? (multiclass classification)

A database of existing relationship helps the training (for example, using Wikipedia infoboxes)

Note on Information Extraction

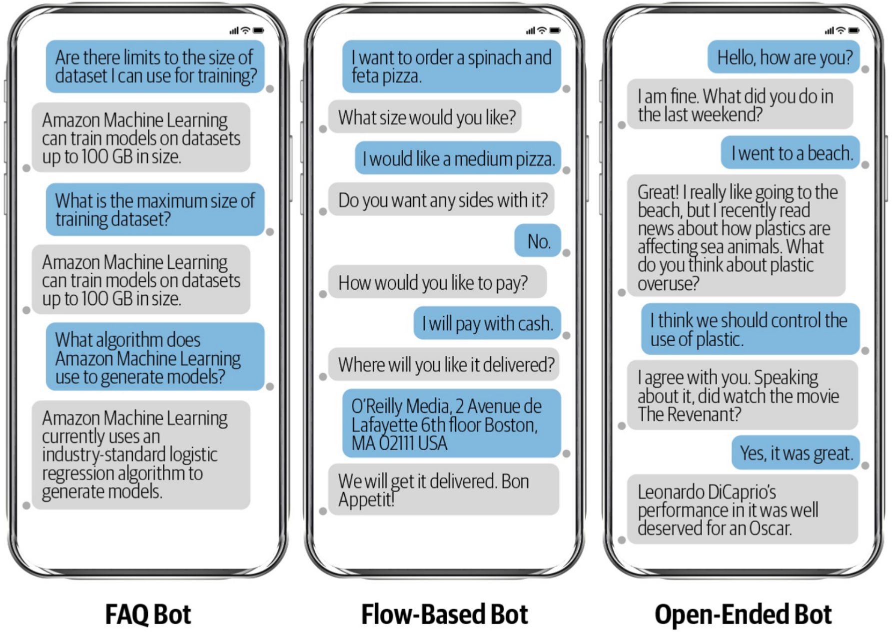
These tasks rely on resources beyond large annotated corpora and also require domain knowledge. Therefore, it is more common to use pre-trained libraries and solutions from large service providers (Cloud APIs).

Only build them from scratch if you are working on a super super specialized domain that needs custom solutions!

Chatbots

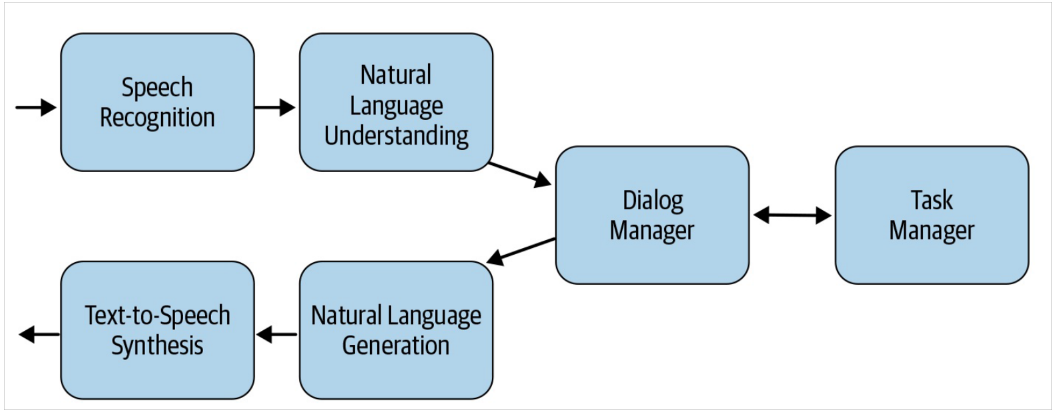
Interactive systems that allow users to interact in natural language.

Types of chatbots:



Chatbots

Pipeline for a Dialogue System

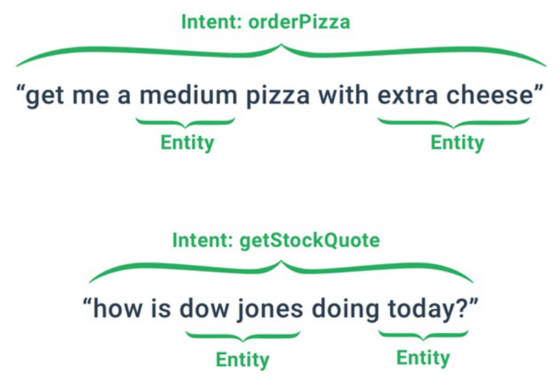


Natural Language Understanding needs all the tools we learned from information extraction.

Chatbots

Intents and Slots Intent: This is the aim of a user command. They are normally pre-defined based on the chatbot's domain

Slot or entity: Fixed construct that holds information regarding specific entities related to the intent.



Intents and Slots Detection

Amazon Alexa and other commercial chatbots have an interface to detect pre-programmed intents and slots. However, automatic detection of intents and slots can also be done with DL:

Intent Detection: A classification problem (again), given a dialog utterance (user sentence), classify it into dialog acts or labels

Slot Detection: This is exactly what we learned in information extraction (KPE, NER, NED, NEL).

User: I'm looking for a cheaper restaurant
`inform(price=cheap)`
System: Sure. What kind - and where?
User: Thai food, somewhere downtown
`inform(price=cheap, food=Thai, area=centre)`
System: The House serves cheap Thai food
User: Where is it?
`inform(price=cheap, food=Thai, area=centre); request(address)`
System: The House is at 106 Regent Street

Response Generation

- **Fixed responses** FAQ bots mainly use them. There is a pool of responses and the bot has to choose which one is the best for the user input
- **Templates** Slot values are used to come up with a follow-up question.
  Example: *<restaurant name> serves <food dish> at only <food price> Euros!*
- **Automatic Generation** A generative model takes the dialog input and generates the next response based on trained language models (DL)

End-to-End Approach

We can use **Seq2Seq** models to train the dialog system!

Hard for open domain chatbots. However, much advance in the last few years on solving question and answering systems with DL.

Task: Question Answering

Answer questions! We divide the task in two major parts:

1. Finding relevant documents that contain the answer This is a task of **Information Retrieval**
2. Finding the answer in the retrieved document (paragraph). This problem is referred to as **Reading Comprehension**. We focus on this subtask.

Stanford Question Answering Dataset (SQuAD)

Private schools, also known as independent schools, non-governmental, or nonstate schools, are not administered by local, state or national governments; thus, they retain the right to select their students and are funded in whole or in part by charging their students tuition, rather than relying on mandatory taxation through public (government) funding; at some private schools students may be able to get a scholarship, which makes the cost cheaper, depending on a talent the student may have (e.g. sport scholarship, art scholarship, academic scholarship), financial need, or tax credit scholarships that might be available.

Along with non-governmental and nonstate schools, what is another name for private schools?

Gold answers: ① independent ② independent schools ③ independent schools

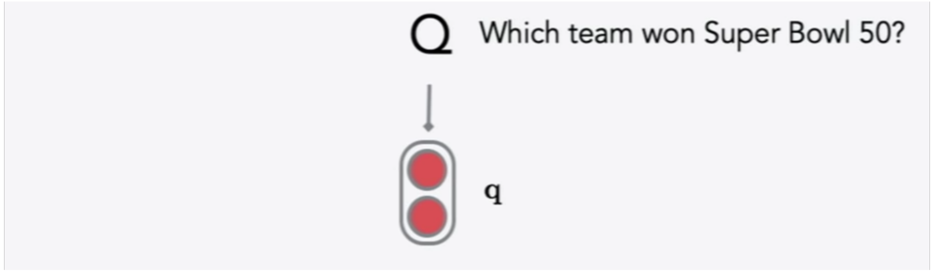
Along with sport and art, what is a type of talent scholarship?

Gold answers: ① academic ② academic ③ academic

Rather than taxation, what are private schools largely funded by?

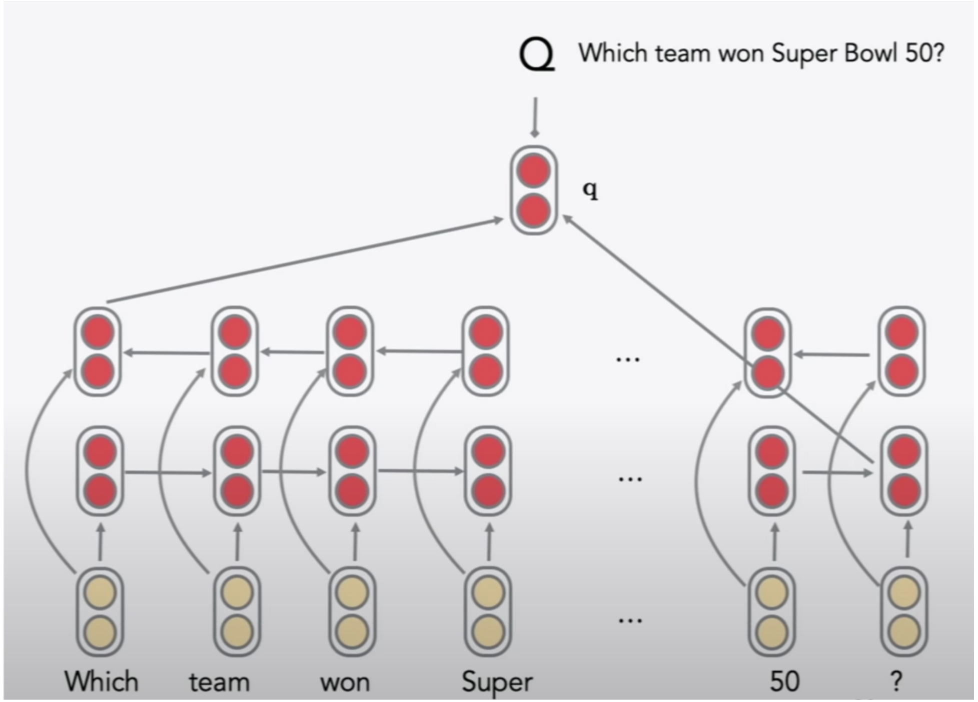
Gold answers: ① tuition ② charging their students tuition ③ tuition

Stanford Attentive Reader

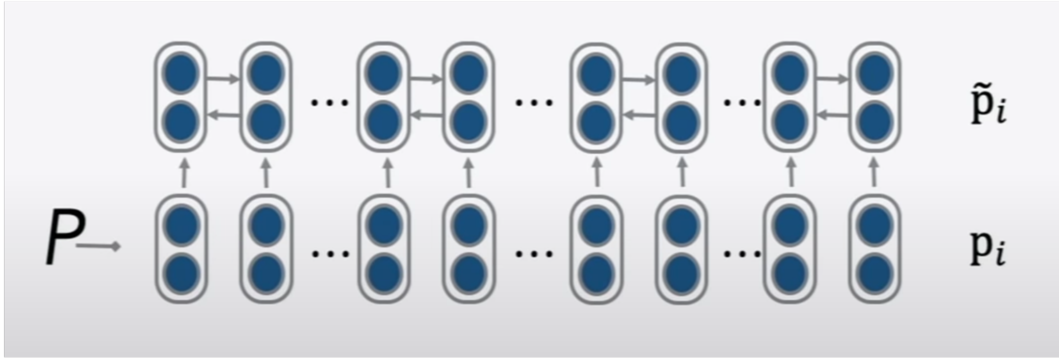


Images from: `https://web.stanford.edu/class/cs224n/`

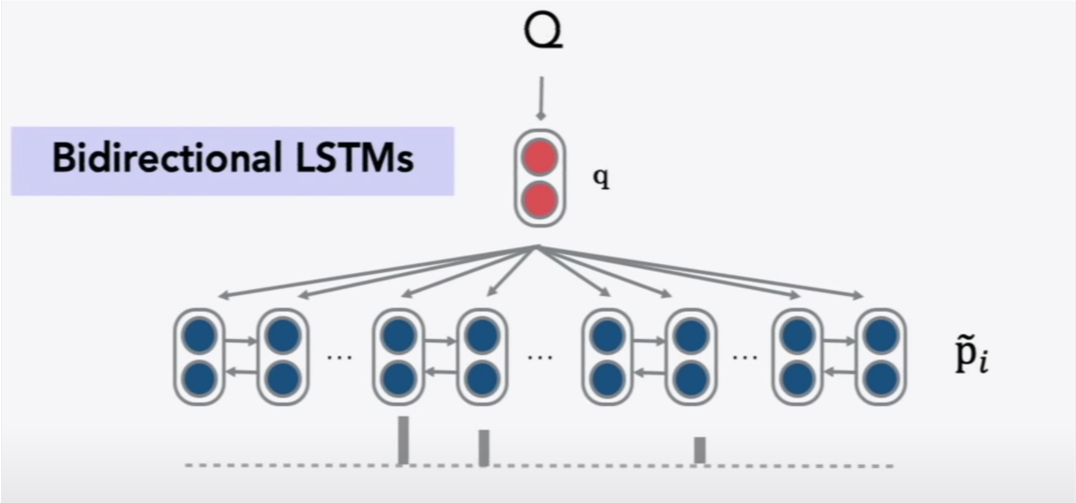
Stanford Attentive Reader



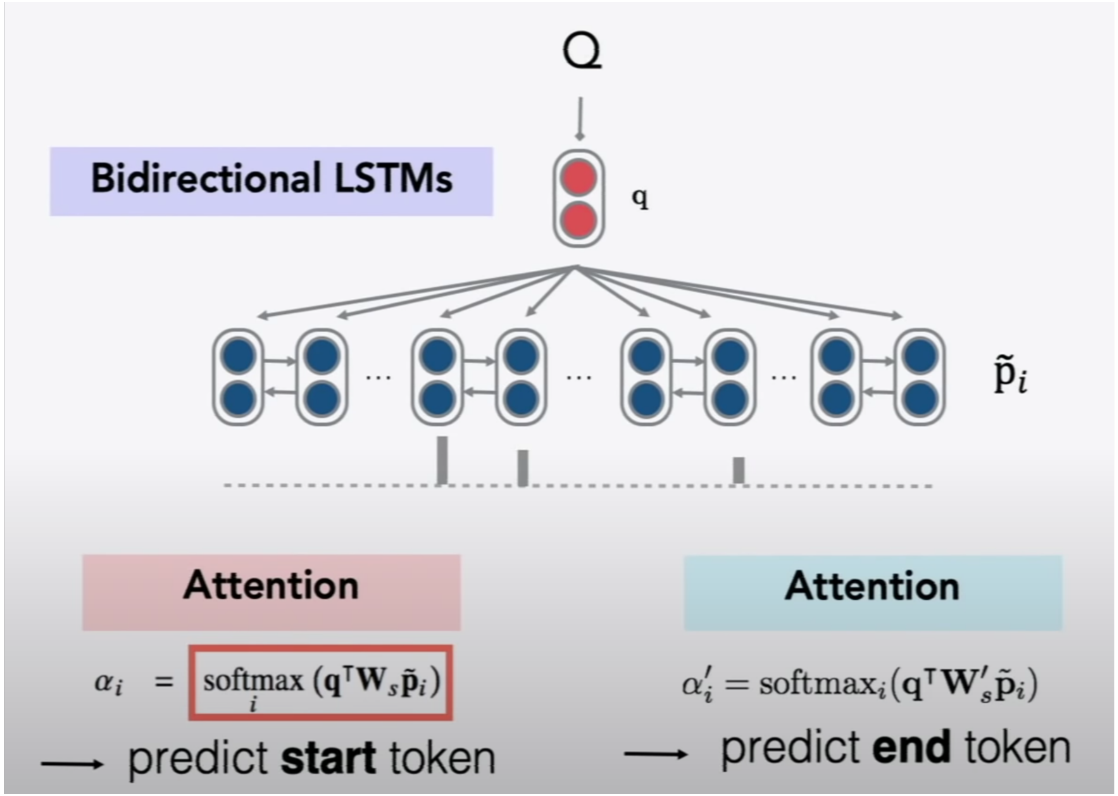
Stanford Attentive Reader



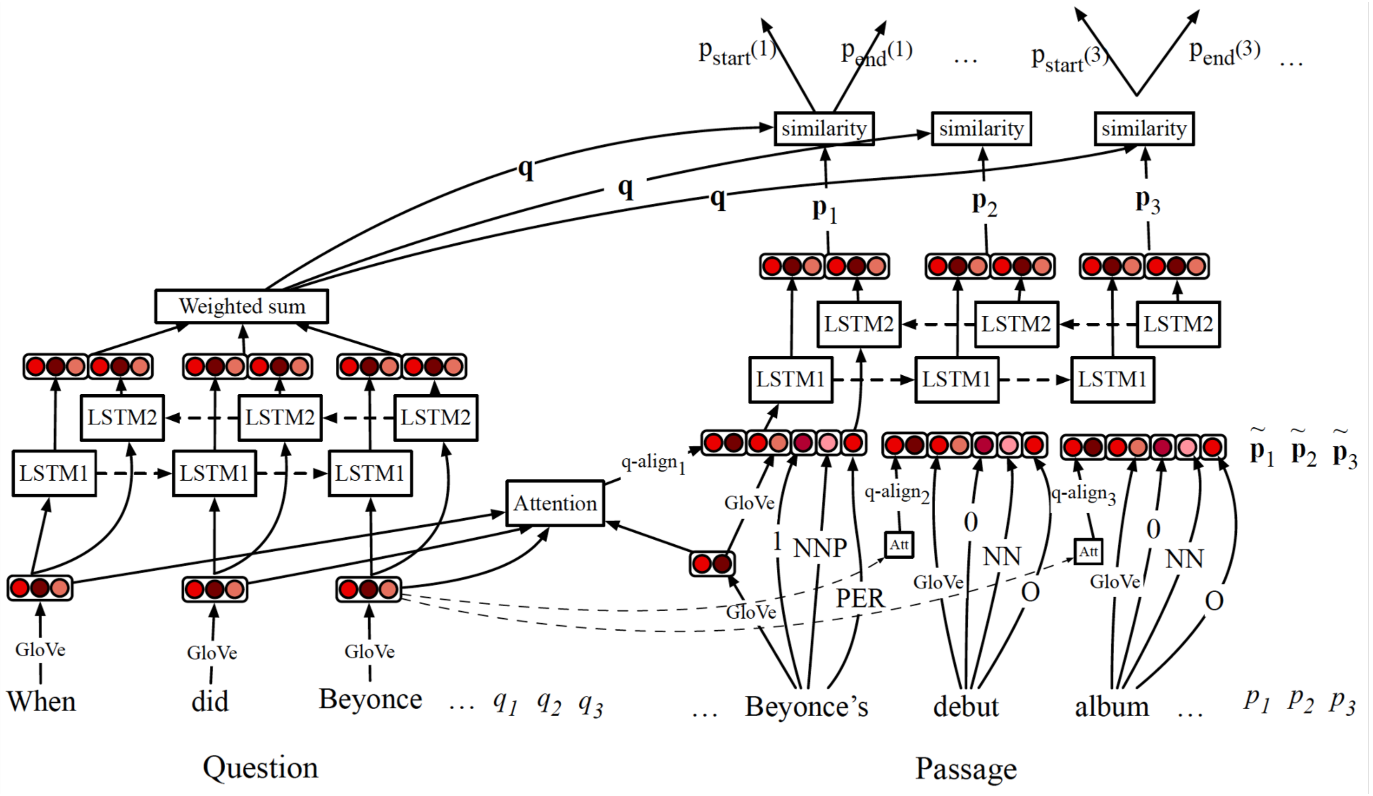
Stanford Attentive Reader



Stanford Attentive Reader

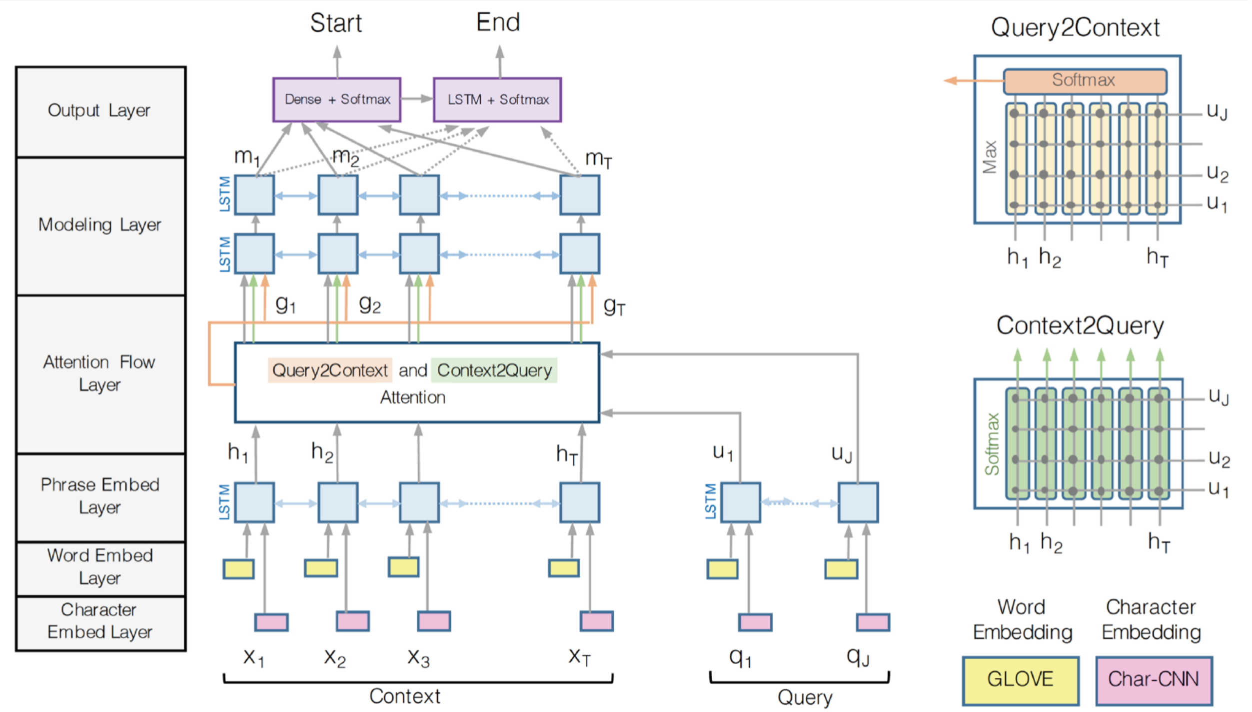


Stanford Attentive Reader++

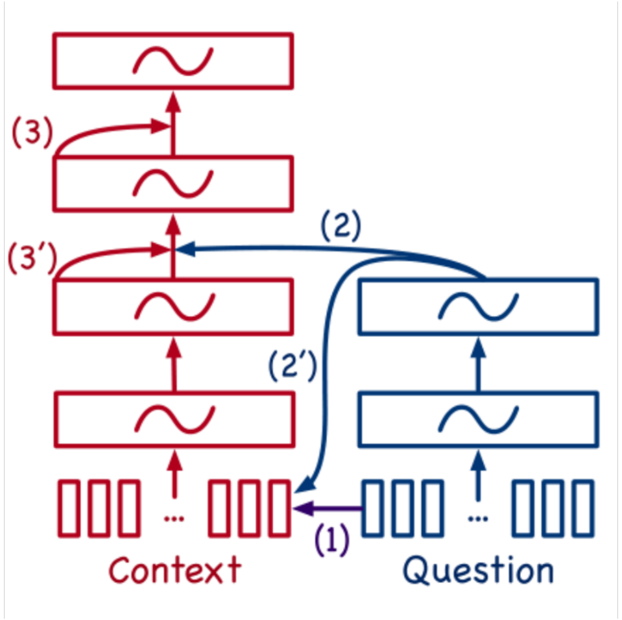


BiDAF

Bi-Directional Attention Flow for Machine Comprehension



General Seq2seq QA models



Stanford Question Answering Dataset 2 (SQuAD2)

Leaderboard

SQuAD2.0 tests the ability of a system to not only answer reading comprehension questions, but also abstain when presented with a question that cannot be answered based on the provided paragraph.

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Apr 06, 2020	SA-Net on Albert (ensemble) *QIANXIN*	**90.724**	**93.011**
2 May 05, 2020	SA-Net-V2 (ensemble) *QIANXIN*	90.679	92.948
2 Apr 05, 2020	Retro-Reader (ensemble) *Shanghai Jiao Tong University* http://arxiv.org/abs/2001.09694	90.578	92.978
3 Jul 31, 2020	ATRLP+PV (ensemble) *Hithink RoyalFlush*	90.442	92.877
3 May 04, 2020	ELECTRA+ALBERT+EntitySpanFocus (ensemble) *SRCB_DML*	90.442	92.839