“On what topic was the lecture on 29th June conducted, during the SHALA 2020 online course?”

A: Question Answering and Multilingual NLP

Diptesh Kanojia
Prashant K. Sharma

We Start at 9.05 PM

Slides Images and Some Titles from Jurafsky’s book

Jurafsky, Dan. *Speech & language processing*. Pearson Education India, 2000.

Agenda

Motivation for QA

Problem Understanding

Paradigms for Question Answering

Multilinguality Problems and Cross-lingual NLP

Indic NLP Library

Cross-lingual Word Representations

Such **Motivation**, Much Wow!

“No one is dumb who is curious. The people who don't ask questions remain clueless throughout their lives.”

-Neil deGrasse Tyson

One of the major challenges for Artificial Intelligence as an area is to create machines which are responsive in a natural conversational setting.

This brings us to the task where a machine should be able to answer any questions which we ask it, based on prior knowledge or web-search.

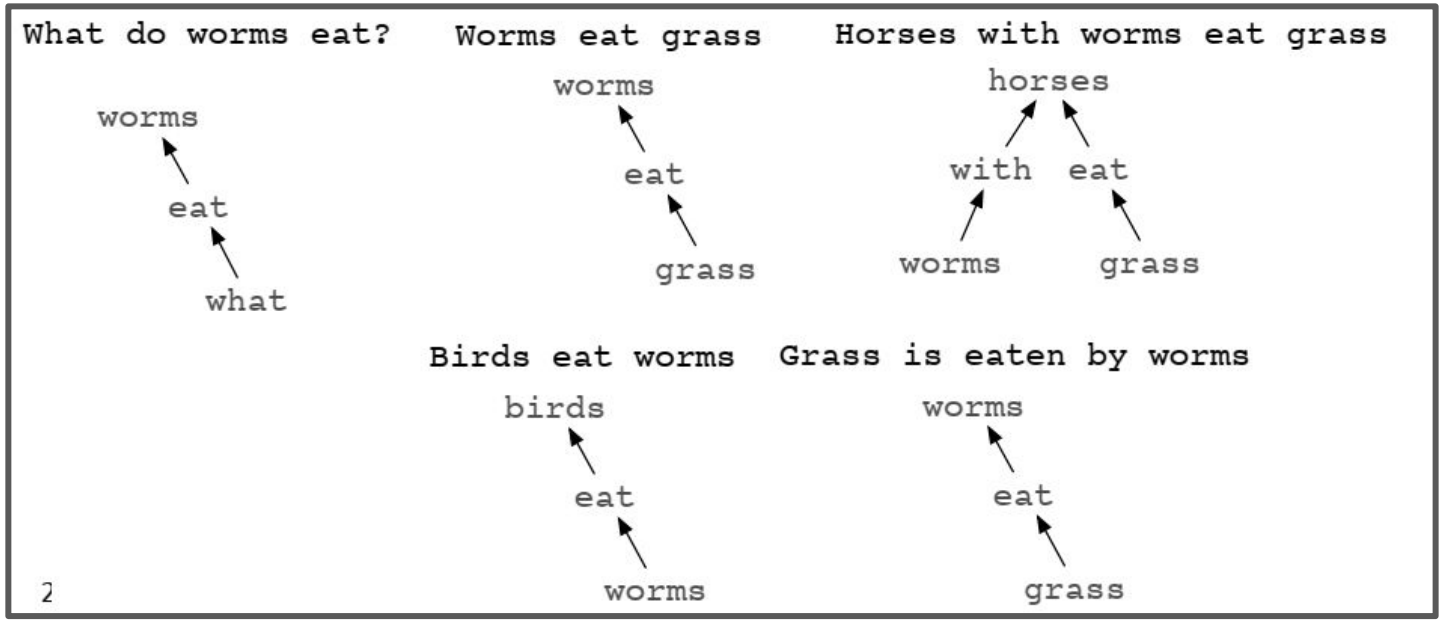
(Spoken) Question Answering is the very response to the question, “What if machines could talk to us?”

(Textual) Question Answering is something we shall discuss today. Mostly, how machines comprehend and answer.

A little history lesson

This is actually one of the oldest tasks in NLP (punched card systems came in 1961)

Simmons, Robert F., Sheldon Klein, and Keren McConlogue. "Indexing and dependency logic for answering English questions." *American Documentation* 15, no. 3 (1964): 196-204.



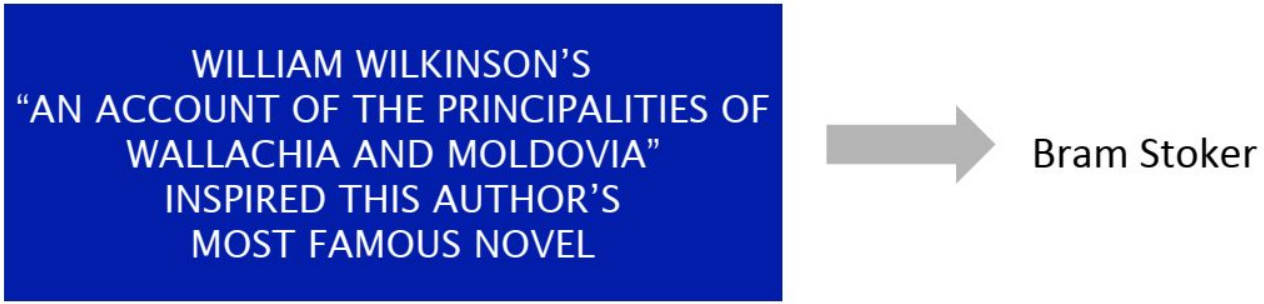
IBM's Watson

Watson is a QA System.

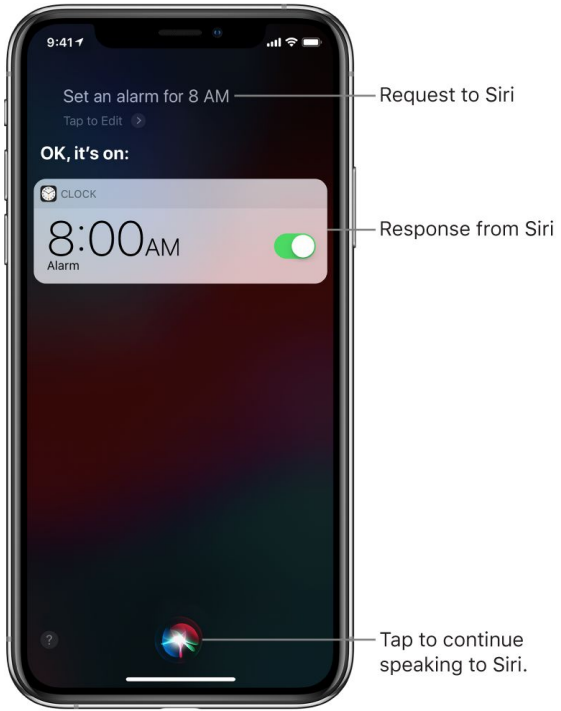
It won the show Jeopardy against the best jeopardy champions. Watch this!

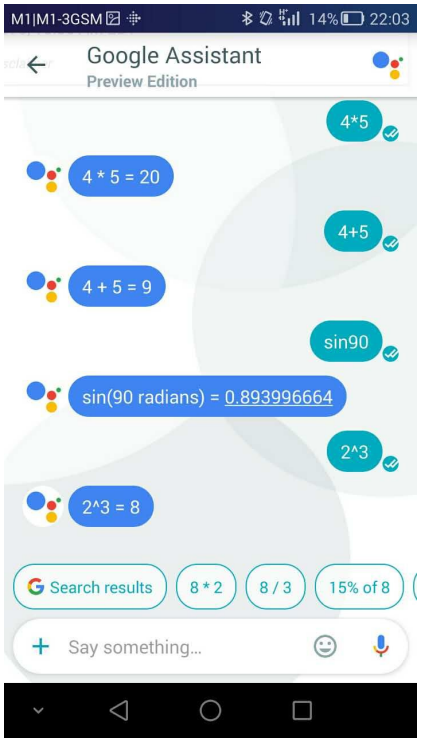
> With the answer: "You just need a nap. You don't have this sleep disorder that can make sufferers nod off while standing up," Watson replied, "What is narcolepsy?"

The Winning reply:



Apple’s Siri, Google’s Assistant, Amazon’s Alexa





Types of Questions in Modern Systems

- Factoid questions
  - *Who wrote "The Universal Declaration of Human Rights"?*
  - *How many calories are there in two slices of apple pie?*
  - *What is the average age of the onset of autism?*
  - *Where is Apple Computer based?*
- Complex (narrative) questions:
  - *In children with an acute febrile illness, what is the efficacy of acetaminophen in reducing fever?*
  - *What do scholars think about Jefferson's position on dealing with pirates?*

Commercial Systems: mostly factoid QA

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 is the telephone number for Stanford University?	650-723-2300

QA Paradigms

IR Based approaches

- TREC, Watson, Google

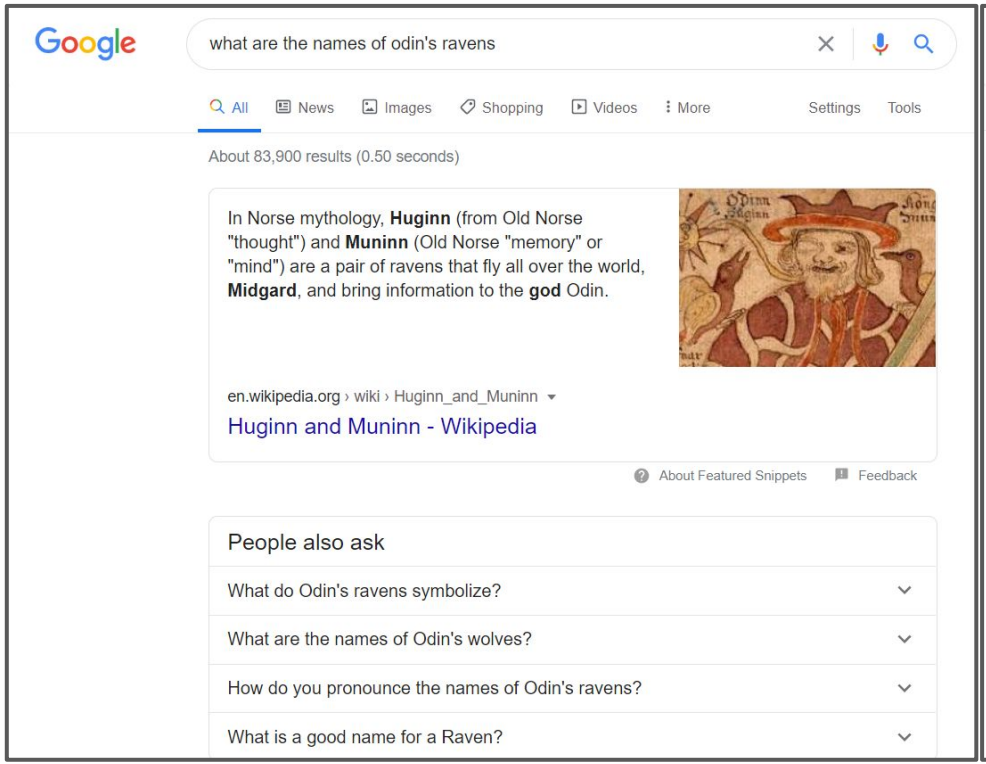
Knowledge-based approaches

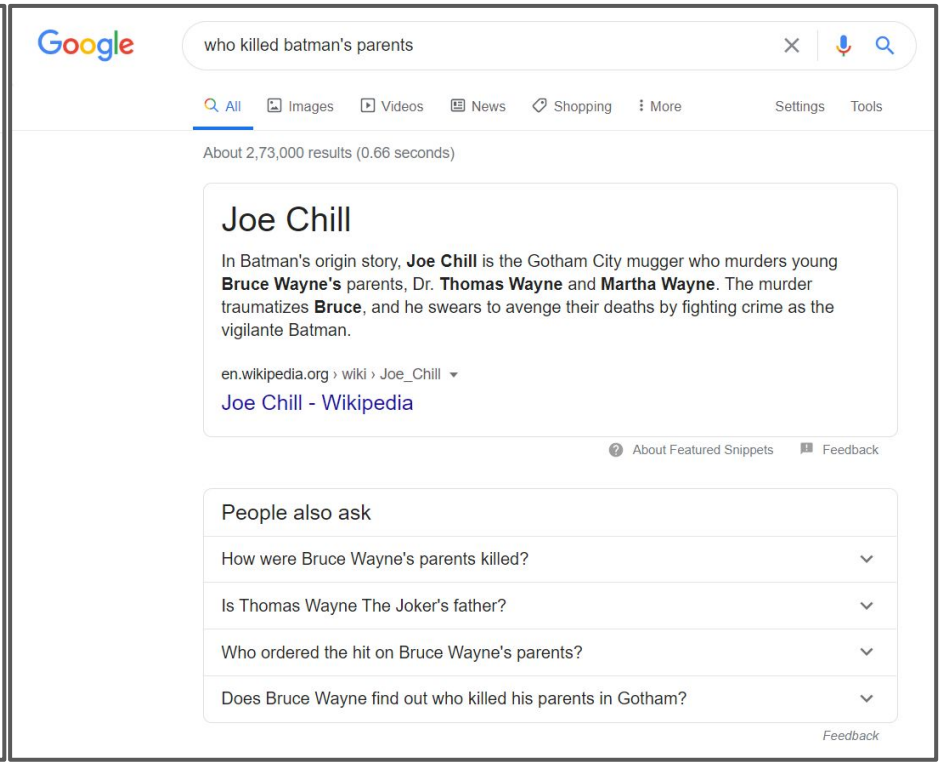
- Apple Siri, Wolfram Alpha

Hybrid approaches

- Watson, True Knowledge Evi

Even Fictional Answers!





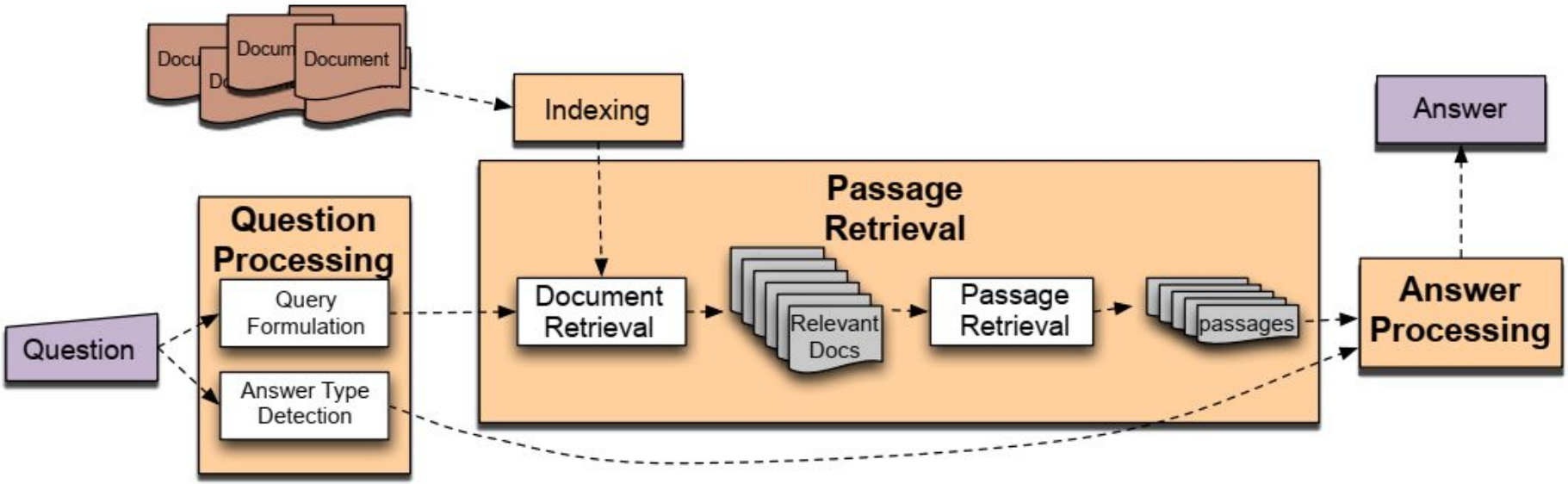
Google’s QA

- Google was a pure IR-based QA, but in 2012 **Knowledge Graph** was added to Google's search engine.

- The Knowledge Graph is a **knowledge base** used by Google to enhance its search engine's search results with semantic-search information gathered from a wide variety of sources.

- Wikipedia: The goal of KGraph is that users would be able to use this information to resolve their query without having to navigate to other sites and assemble the information themselves. [...] According to some news websites, the implementation of Google's Knowledge Graph has played a role in the page view decline of various language versions of Wikipedia.

IR Based Approach



IR based Factoid QA

- QUESTION PROCESSING
  - Detect question type, answer type, focus, relations
  - Formulate queries to send to a search engine
- PASSAGE RETRIEVAL
  - Retrieve ranked documents
  - Break into suitable passages and rerank
- ANSWER PROCESSING
  - Extract candidate answers
  - Rank candidates
    - using evidence from the text and external sources

Knowledge based approaches

- Build a semantic representation of the query
  - Times, dates, locations, entities, numeric quantities
- Map from this semantics to query structured data or resources
  - Geospatial databases
  - Ontologies (Wikipedia infoboxes, dbPedia, WordNet, Yago)
  - Restaurant review sources and reservation services
  - Scientific databases

SIRI at a high level

- Using ASR (Automatic speech recognition) to transcribe human speech (in this case, short utterances of commands, questions, or dictations) into text.
- Using natural language processing (part of speech tagging, noun-phrase chunking, dependency & constituent parsing) to translate transcribed text into "parsed text".
- Using question & intent analysis to analyze parsed text, detecting user commands and actions. ("Schedule a meeting", "Set my alarm", ...)
- Using data technologies to interface with 3rd-party web services such as OpenTable, WolframAlpha, to perform actions, search operations, and question answering.
- **Utterances SIRI has identified as a question, that it cannot directly answer, it will forward to more general question-answering services such as WolframAlpha**
- Transforming output of 3rd party web services back into natural language text (eg, Today's weather report -> "The weather will be sunny")
- Using TTS (text-to-speech) technologies to transform the natural language text from step 5 above into synthesized speech.

Hybrid Approaches

- Build a shallow semantic representation of the query
- Generate answer candidates using IR methods
  - Augmented with ontologies and semi-structured data
- Score each candidate using richer knowledge sources
  - Geospatial databases
  - Temporal reasoning
  - Taxonomical classification

Things to extract from the question

- Answer Type Detection
  - Decide the **named entity type** (person, place) of the answer
- Query Formulation
  - Choose **query keywords** for the IR system
- Question Type classification
  - Is this a definition question, a math question, a list question?
- Focus Detection
  - Find the question words that are replaced by the answer
- Relation Extraction
  - Find relations between entities in the question

Question Processing

Which two states do you enter when you exit the state of Uttar Pradesh from the southern border?

Answer Type: Indian State

Query: two states, border, Uttar Pradesh, south

Focus: two states

Relations: borders(Uttar Pradesh, ?x, south)

Answer type detection: Names Entities

Q: Who found Vistara Airlines?

- PERSON

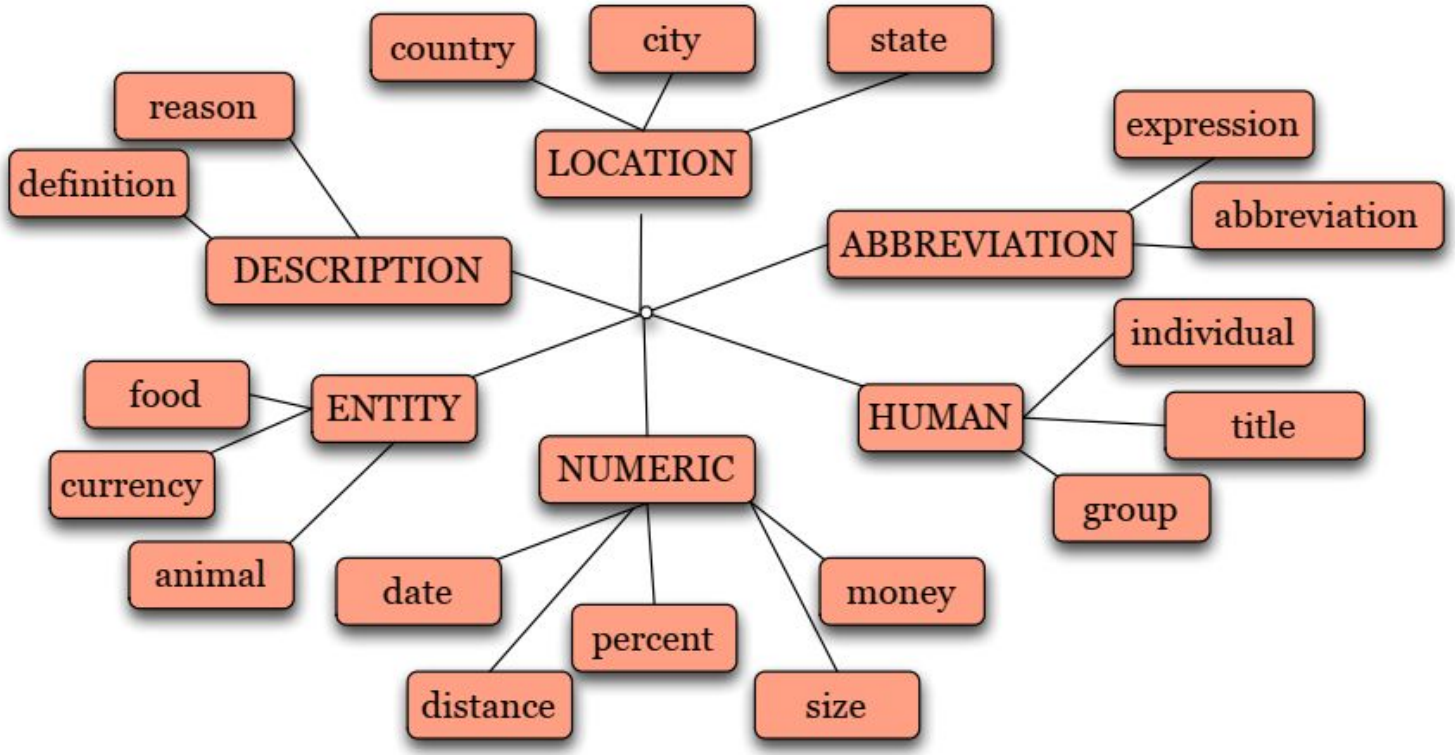
Q: What Indian city has the largest population?

- CITY

Answer Type Taxonomy (Xin Li, Dan Roth. 2002)

- 6 coarse classes
  - ABBEVIATION, ENTITY, DESCRIPTION, HUMAN, LOCATION, NUMERIC
- 50 finer classes
  - LOCATION: city, country, mountain…
  - HUMAN: group, individual, title, description
  - ENTITY: animal, body, color, currency…

Li and Roth's Taxonomy



Answer Type Detection

- Hand Written
- Machine Learning
- Hybrids

Hand Written

- Regular expression-based rules can get some cases:
  - Who {is|was|are|were} PERSON
  - PERSON (YEAR – YEAR)
- Other rules use the **question headword**:

  (the headword of the first noun phrase after the wh-word)

  - Which **city** in China has the largest number of foreign financial companies?
  - What is the state **flower** of California?

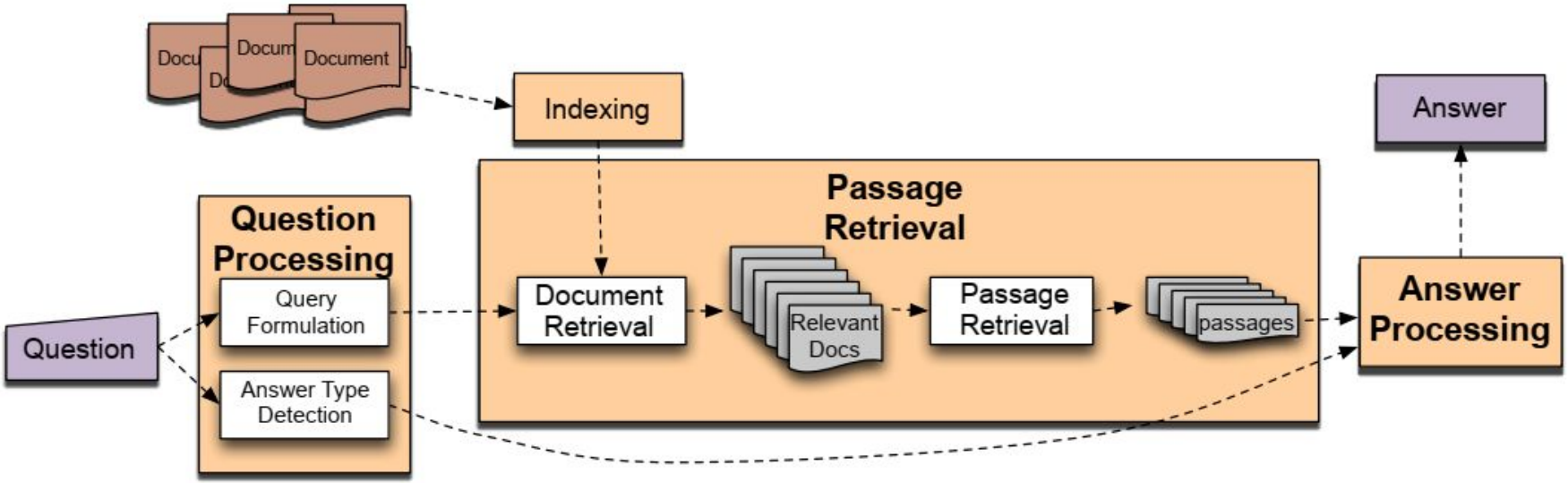
Machine Learning

- Most often, we treat the problem as machine learning classification
  - **Define** a taxonomy of question types
  - **Annotate** training data for each question type
  - **Train** classifiers for each question class using a rich set of features.
    - features include those hand-written rules!

Feature for Answer Type Detection

- Question words and phrases
- Part-of-speech tags
- Parse features (headwords)
- Named Entities
- Semantically related words

Passage Retrieval and Answer Ranking



Passage Retrieval

- Step 1: IR engine retrieves documents using query terms
- Step 2: Segment the documents into shorter units
  - something like paragraphs
- Step 3: Passage ranking
  - Use answer type to help rerank passages

Features for Passage Ranking

- Number of Named Entities of the right type in passage
- Number of query words in passage
- Number of question N-grams also in passage
- Proximity of query keywords to each other in passage
- Longest sequence of question words
- Rank of the document containing passage

Answer Extraction

- Run an answer-type named-entity tagger on the passages
  - Each answer type requires a named-entity tagger that detects it
  - If answer type is CITY, tagger has to tag CITY
    - Can be full NER, simple regular expressions, or hybrid
- Return the string with the right type:
  - `Who is the prime minister of India` (PERSON)

  Narendra Modi, the prime minister of India stated that none of the Chinese Soldiers had entered the Galwan Valley.

  - `How tall is Mt. Everest?` (LENGTH)

  `The official height of Mount Everest is` **`29035 feet`**

Ranking Candidate Answers

Answer type match: Candidate contains a phrase with the correct answer type.

Pattern match: Regular expression pattern matches the candidate.

Question keywords: # of question keywords in the candidate.

Keyword distance: Distance in words between the candidate and query keywords

Novelty factor: A word in the candidate is not in the query.

Apposition features: The candidate is an appositive to question terms

Punctuation location: The candidate is immediately followed by a comma, period, quotation marks, semicolon, or exclamation mark.

Sequences of question terms: The length of the longest sequence of question terms that occurs in the candidate answer.

Stanford Questions Answering Datasaet (SQuAD)

100k examples

Answer must be a span in the passage

A.k.a. extractive question answering

SQuAD 1.1:

- Authors collected 3 gold answers
- Systems are scored on two metrics:
- F1 measure is seen as more reliable and taken as primary
  - It's less based on choosing exactly the same span that humans chose, which is susceptible to various effects, including line breaks
- Both metrics ignore punctuation and articles (**a**, **an**, **the** only)

SQuAD 2.0

- A defect of SQuAD 1.0 is that all questions have an answer in the paragraph
- Systems (implicitly) rank candidates and choose the best one
- You don't have to judge whether a span answers the question
- In SQuAD 2.0, 1/3 of the training questions have no answer, and about 1/2 of the dev/test questions have no answer
  - For NoAnswer examples, NoAnswer receives a score of 1, and any other response gets 0, for both exact match and F1
- Simplest system approach to SQuAD 2.0:
  - Have a threshold score for whether a span answers a question
- Or you could have a second component that confirms answering
  - Like Natural Language Inference (NLI) or "Answer validation"

Reference Reads

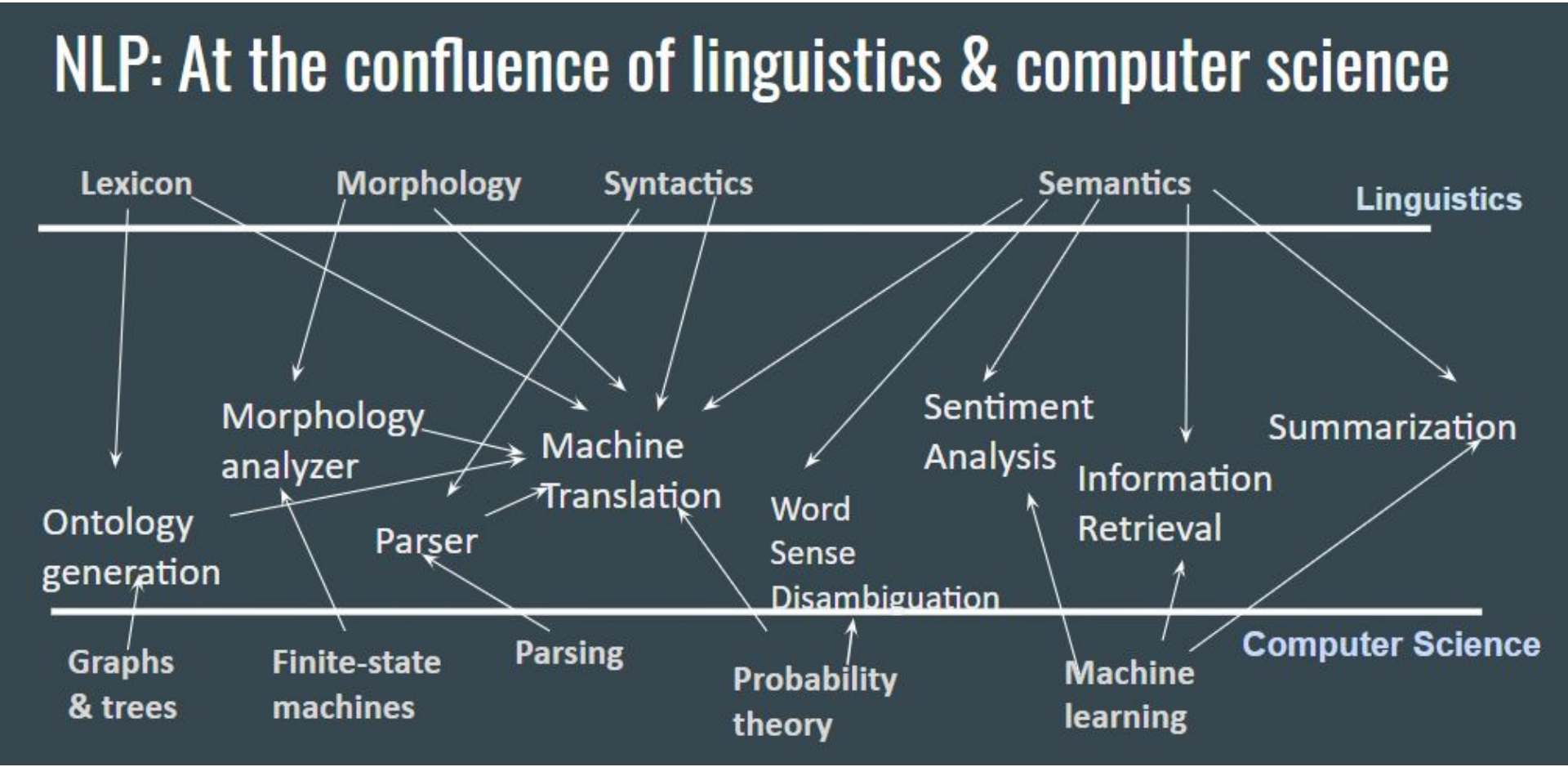
Stanford Attentive Reader

Stanford Attentive Reader++

BiDAF

Please go through their architectures and post any doubts

Multilinguality in NLP

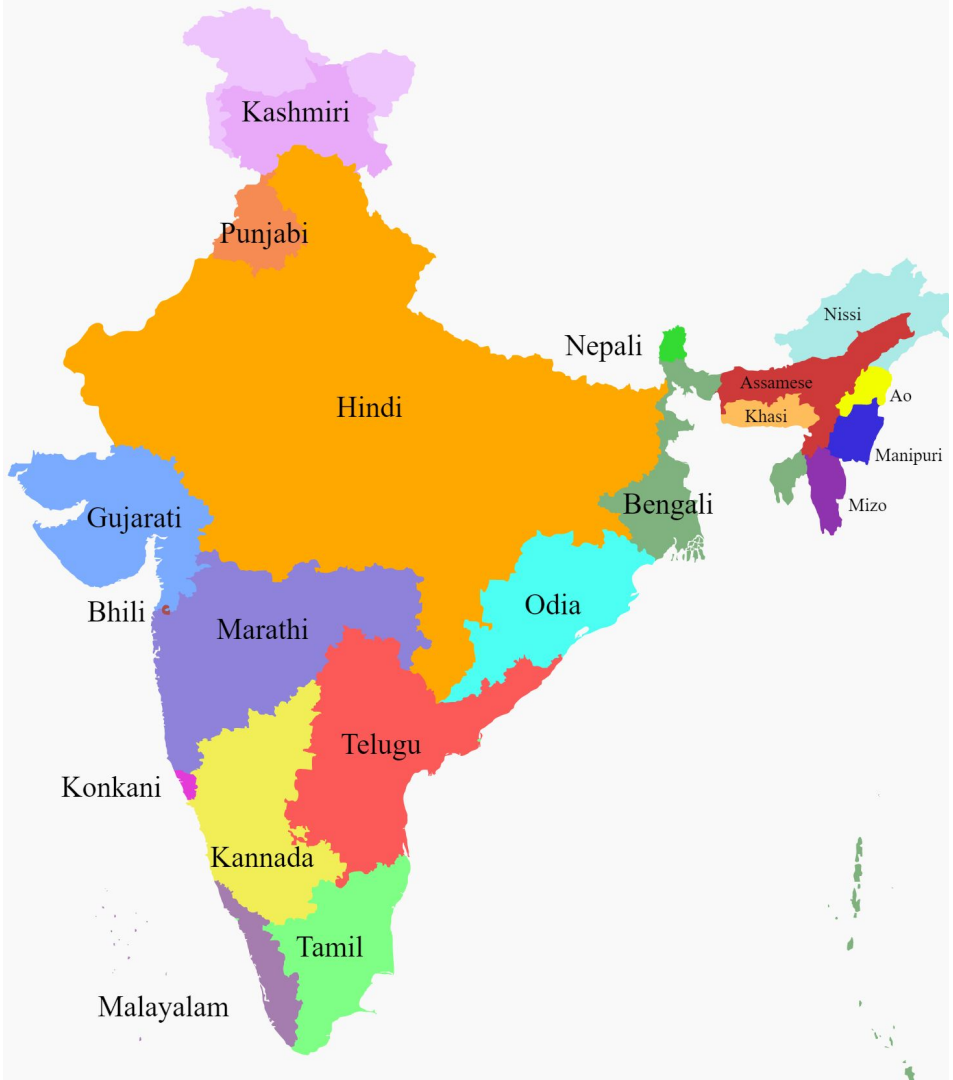


Slide Credit: Prof. Pushpak Bhattacharyya

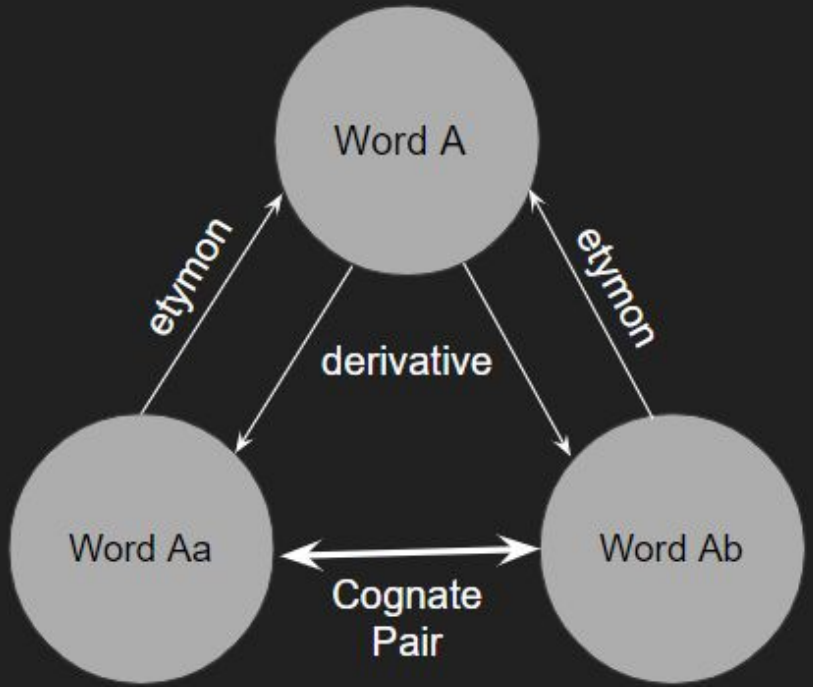
“Do you speak Indian?”

- India has a total of 22 scheduled languages which primarily belong to the *Indo-Aryan* and the *Dravidian* language families.
- I can speak in Hindi, Punjabi, Marathi, and English, but I can only write in Hindi, and English.
- Similarly, many natives of India can speak and understand multiple Indic languages but can only write a subset of those.

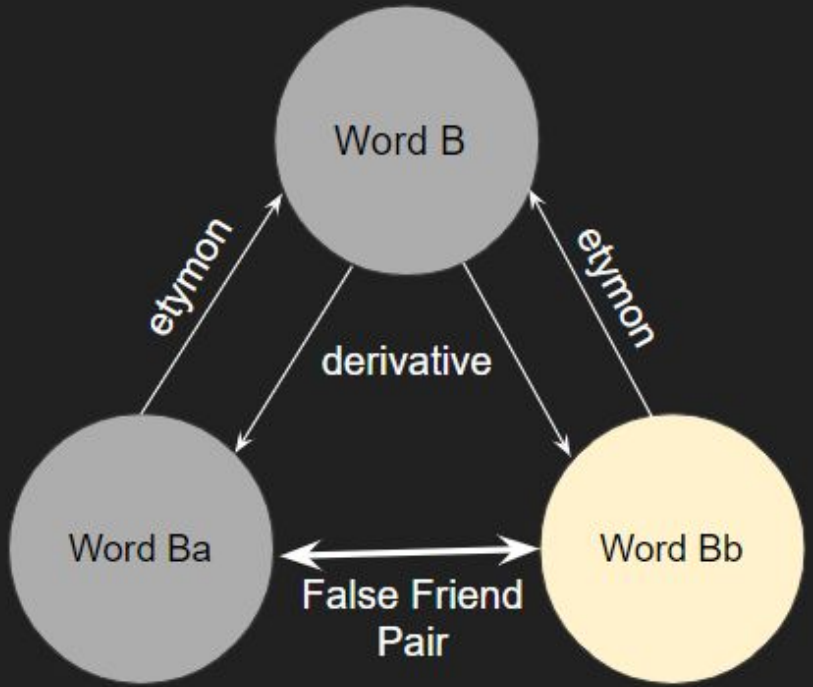
- **I used to wonder why?**


Kashmiri
Punjabi
Hindi
Nepali
Nissi
Assamese
Ao
Khasi
Manipuri
Mizo
Bengali
Gujarati
Bhili
Marathi
Odia
Telugu
Konkani
Kannada
Tamil
Malayalam

See what you (do not) mean!

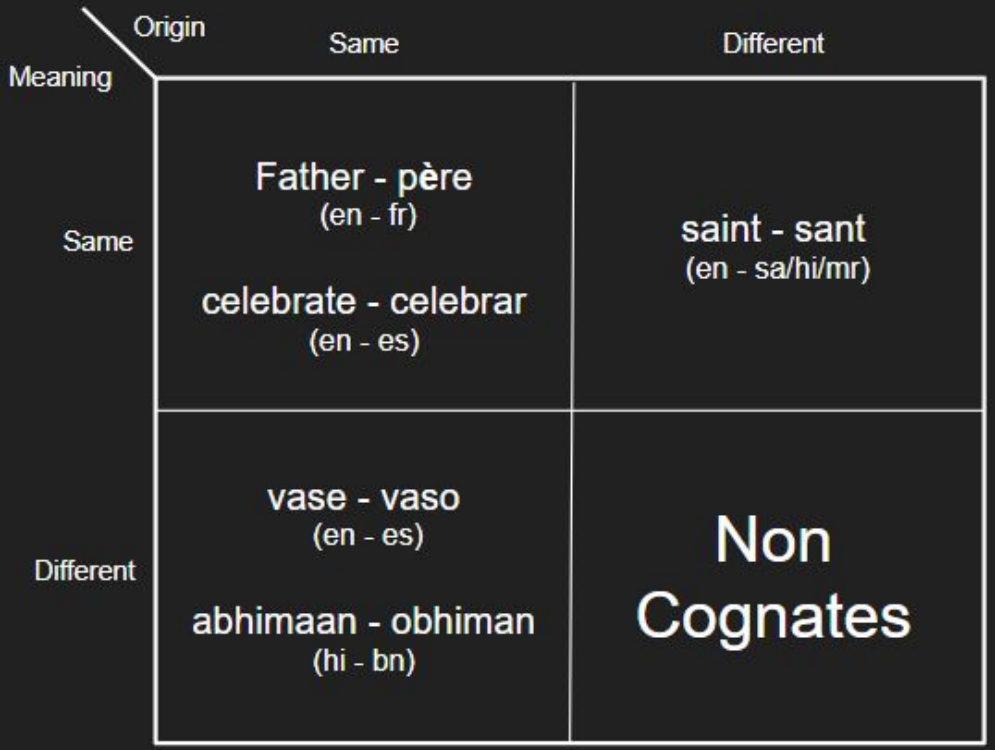


They carry the same meaning



They differ in meaning

The etymological matrix



Indic NLP Library [1 / 3]

Language Support

Indo-Aryan			Dravidian	Others
Assamese (asm)	Marathi (mar)	Sindhi (snd)	Kannada (kan)	English (eng)
Bengali (ben)	Nepali (nep)	Sinhala (sin)	Malayalam (mal)	
Gujarati (guj)	Odia (ori)	Sanskrit (san)	Telugu (tel)	
Hindi/Urdu (hin/urd)	Punjabi (pan)	Konkani (kok)	Tamil (tam)	

1. The IndicNLP Library by Anoop Kunchukuttan
2. https://anoopkunchukuttan.github.io/indic_nlp_library/

Indic NLP Library [2 / 3]

Tasks

Monolingual	Indo-Aryan													Dravidian			
	san	hin	urd	pan	nep	snd	asm	ben	ori	guj	mar	kok	sin	kan	tel	tam	mal
Script Information	✓	✓	✗	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓	✓	✓	✓
Normalization	✓	✓	✗	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓	✓	✓	✓
Tokenization	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Word segmentation	✗	✓	✗	✓	✗	✗	✗	✓	✗	✓	✓	✓	✗	✓	✓	✓	✓
Romanization (ITRANS)	✓	✓	✗	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
ITRANS to Script	✓	✓	✗	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Bilingual

- **Script Conversion:** Amongst the above mentioned languages, except Urdu and English
- **Transliteration:** Amongst the 18 above mentioned languages
- **Translation:** Amongst these 10 languages: (hin, urd, pan, ben, guj, mar, kok, sin, kan, tel, tam, mal) + English

Indic NLP Library [3 / 3]

MUSE

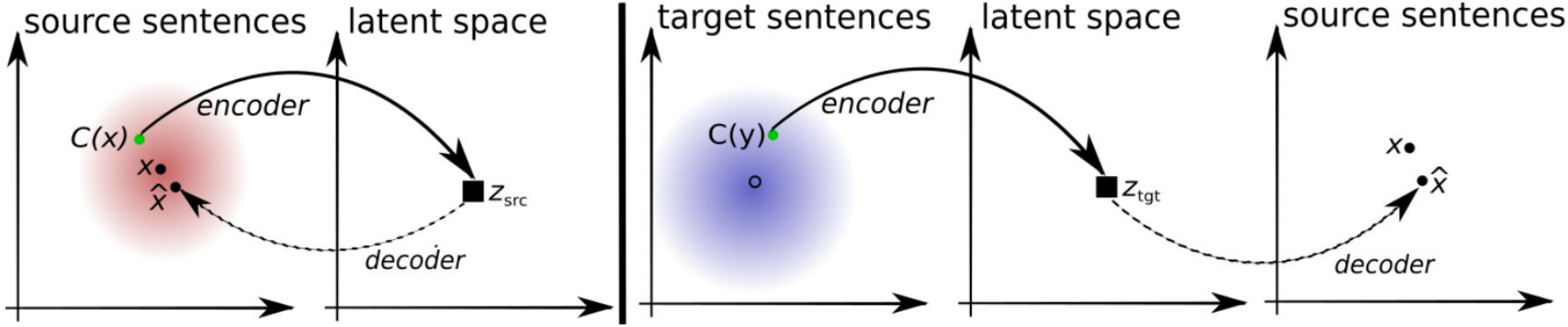


Figure 1: Toy illustration of the principles guiding the design of our objective function. Left (auto-encoding): the model is trained to reconstruct a sentence from a noisy version of it. x is the target, $C(x)$ is the noisy input, $\hat{x}$ is the reconstruction. Right (translation): the model is trained to translate a sentence in the other domain. The input is a noisy translation (in this case, from source-to-target) produced by the model itself, M, at the previous iteration (t), $y = M^{(t)}(x)$. The model is symmetric, and we repeat the same process in the other language. See text for more details.

1. Lample, Guillaume, et al. "Unsupervised machine translation using monolingual corpora only." *arXiv preprint arXiv:1711.00043* (2017)
2. https://ai.facebook.com/tools/muse/

VecMap

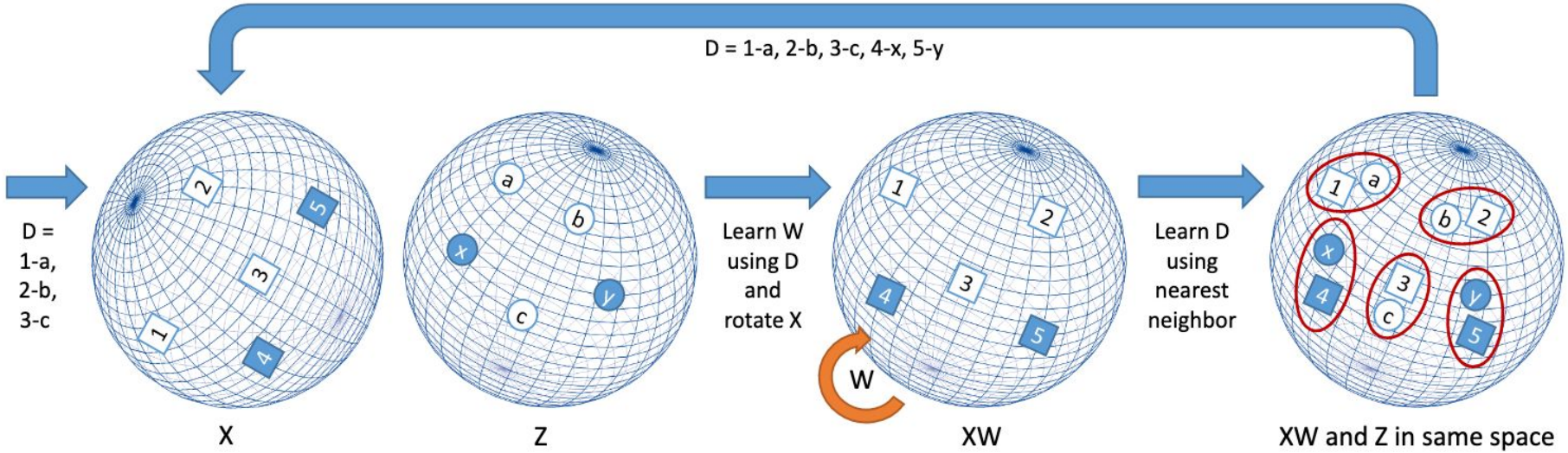
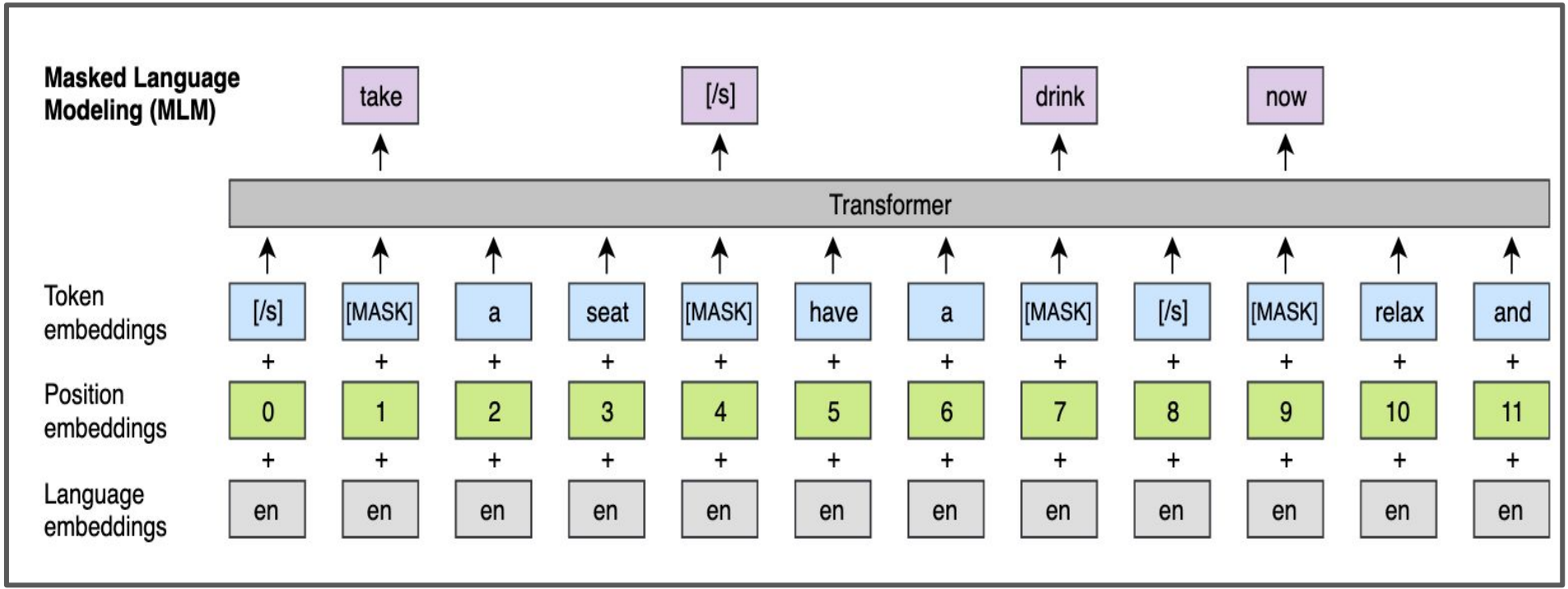


Figure 1: A general schema of the proposed self-learning framework. Previous works learn a mapping W based on the seed dictionary D, which is then used to learn the full dictionary. In our proposal we use the new dictionary to learn a new mapping, iterating until convergence.

1. Lample, Guillaume, and Alexis Conneau. "Cross-lingual language model pretraining." *arXiv preprint arXiv:1901.07291* (2019).
2. https://github.com/facebookresearch/XLM

XLM-R = **XLM** + **R**oBERTa [1 / 2]



1. Lample, Guillaume, and Alexis Conneau. "Cross-lingual language model pretraining." *arXiv preprint arXiv:1901.07291* (2019).
2. https://github.com/facebookresearch/XLM

XLM-R = **XLM** + **R**oBERTa [2 / 2]

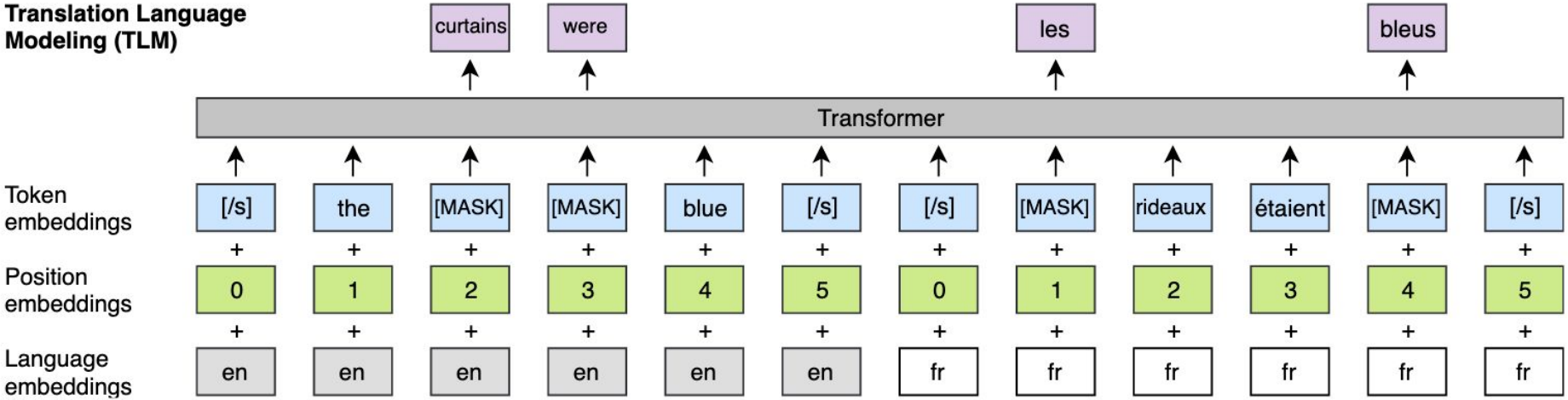


Figure 1: **Cross-lingual language model pretraining.** The MLM objective is similar to the one of Devlin et al. (2018), but with continuous streams of text as opposed to sentence pairs. The TLM objective extends MLM to pairs of parallel sentences. To predict a masked English word, the model can attend to both the English sentence and its French translation, and is encouraged to align English and French representations. Position embeddings of the target sentence are reset to facilitate the alignment.

1. Lample, Guillaume, et al. "Unsupervised machine translation using monolingual corpora only." *arXiv preprint arXiv:1711.00043* (2017)
2. https://ai.facebook.com/tools/muse/

Summary

Question Answering Problem + Motivation

Paradigms in QA

QA Methods (Question Processing + Answer Type + Passage Retrieval)

SQuAD

Multilinguality in NLP

Cognate Detection

Cross-lingual Representations

Thank you!

Questions? 😎