(Low-Resource) NLP Tasks

Graham Neubig

@ CMU Low-resource NLP Bootcamp

5/18/2020

Carnegie Mellon University

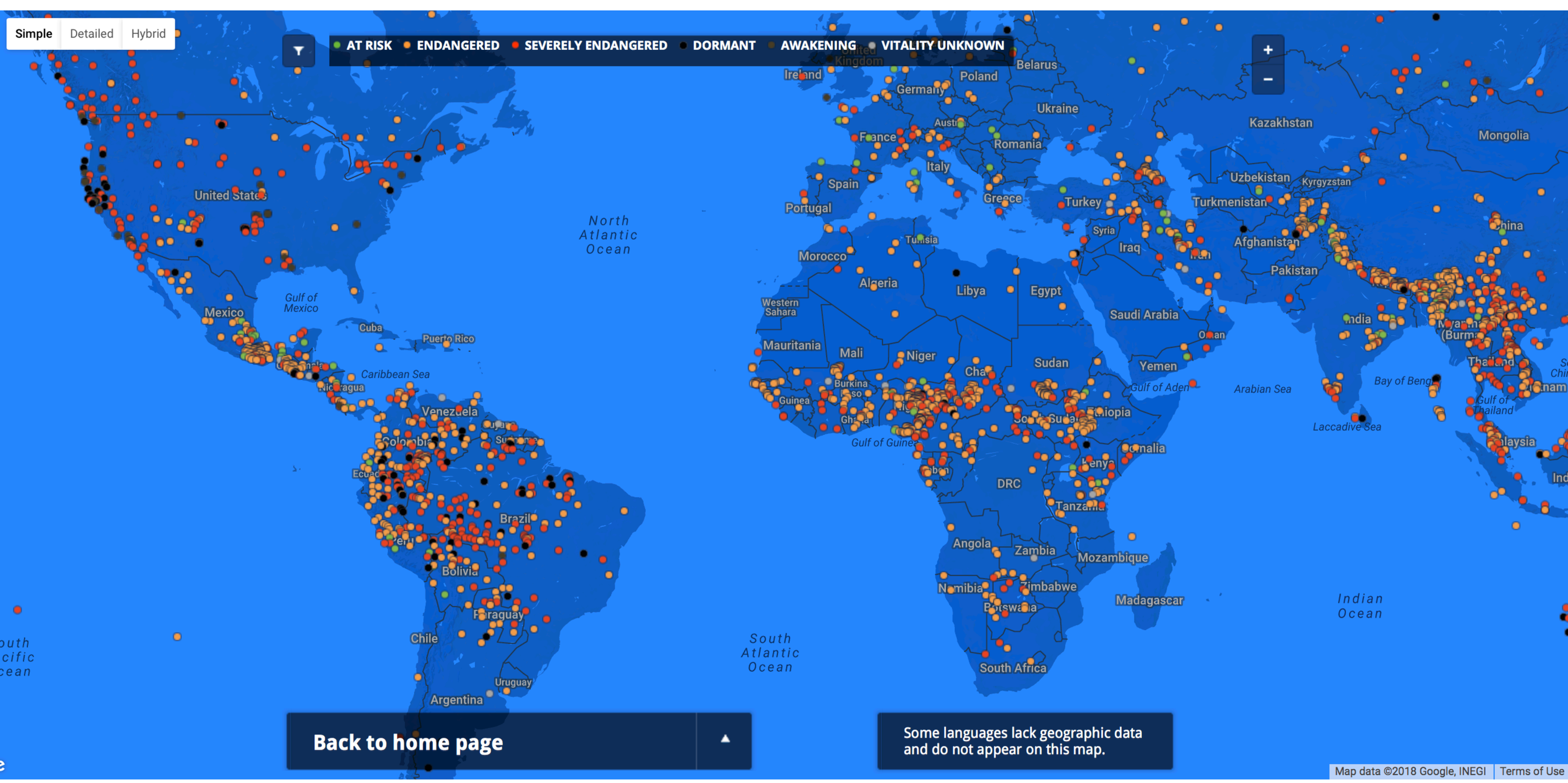
Language Technologies Institute

NEULAB

Most Spoken Languages of the World

1. English (1.132 B)
2. 中文(普通话) (1.116 B)
3. हिन्दी (615.4 M)
4. Español (534.4 M)
5. Français (279.8 M)
6. اَلْعَرَبِيَّةُ (273.9 M)
7. বাংলা (265.0 M)
8. Росси́я (258.2 M)
9. Português (234.1 M)
10. Bahasa Indonesia (279.8 M)

Source: Ethnologue 2019 via Wikipedia



http://endangeredlanguages.com/

Why NLP for All Languages?

- Aid **human-human communication** (e.g. machine translation)
- Aid **human-machine communication** (e.g. speech recognition/synthesis, question answering, dialog)
- **Analyze/understand language** (syntactic analysis, text classification, entity/relation recognition/linking)

Rule-based NLP Systems

- Develop rules, from simple scripts to more complicated rule systems
- Generally must be developed for each language by a linguists
- Appropriate for some simple tasks, e.g. pronunciation prediction in epitran

```
from epitran.backoff import Backoff
>>> backoff = Backoff(['hin-Deva', 'eng-Latn', 'cmn-Hans']
>>> backoff.transliterate('हिन्दी')
'ɦindiː'
>>> backoff.transliterate('English')
'ɪŋglɪʃ'
>>> backoff.transliterate('中文')
'ʈ͡ʂoŋwən'
```

https://github.com/dmort27/epitran

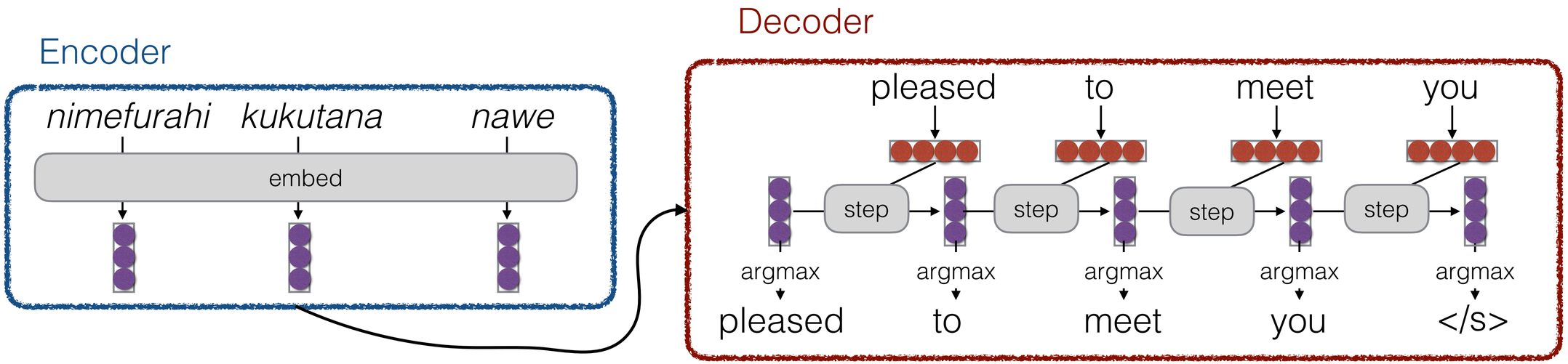
Machine Learning NLP Systems

- Formally, learn a model to map an input *X* into an output *Y*. Examples:

Input *X*	Output *Y*	Task
Text	Text in Other Language	Translation
Text	Response	Dialog
Speech	Transcript	Speech Recognition
Text	Linguistic Structure	Language Analysis

- To learn, we can use
  - Paired data <*X*, *Y*>, source data *X*, target data *Y*
  - Paired/source/target data in *similar* languages

Example Model: Sequence-to-sequence Model with Attention



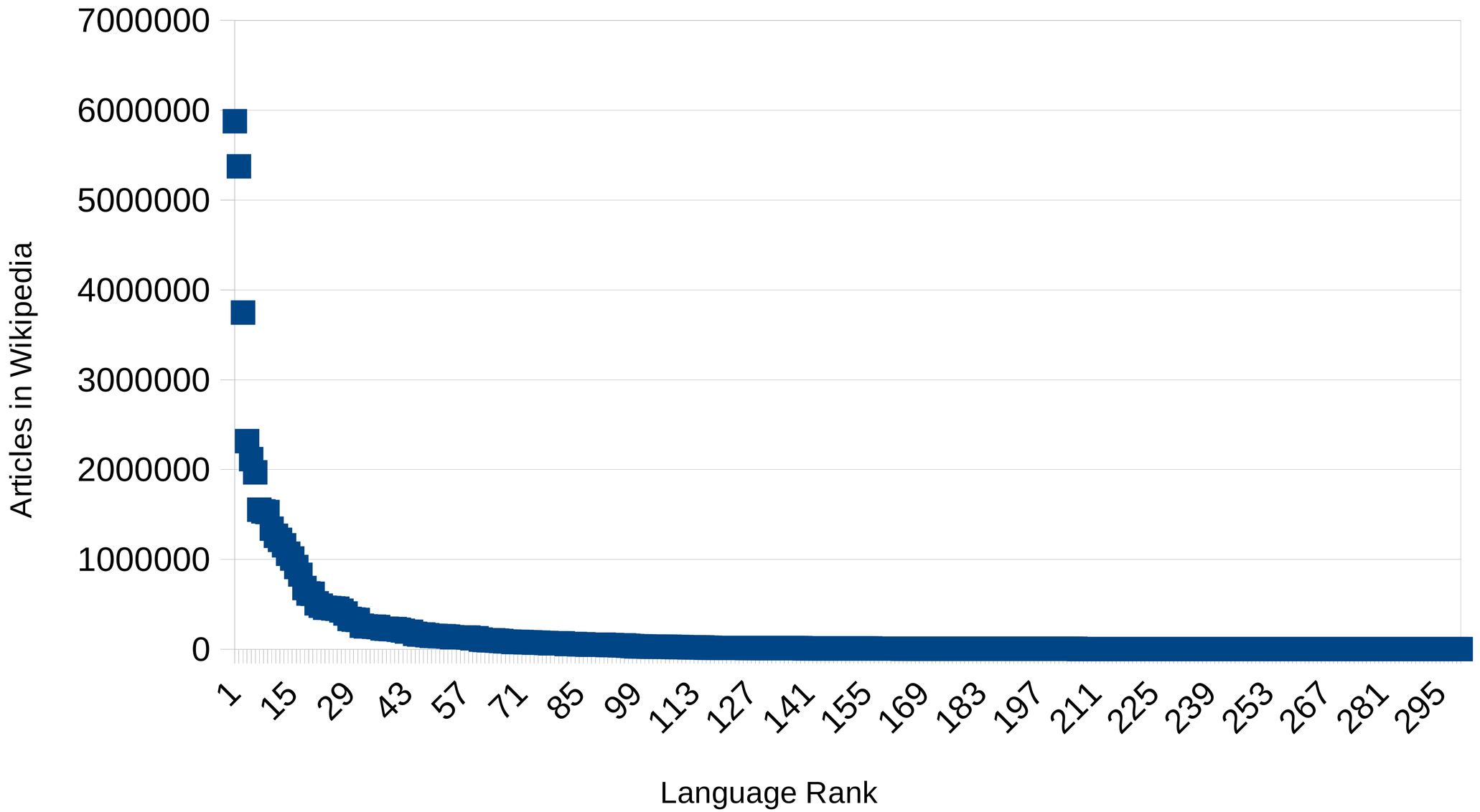
- **Various tasks:** Translation, speech recognition, dialog, summarization, language analysis
- **Various models:** LSTM, transformer
- Generally trained using **supervised learning**: maximize likelihood of <*X*, *Y*>

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. "Neural machine translation by jointly learning to align and translate." *arXiv preprint arXiv:1409.0473* (2014).

Evaluating ML-based NLP Systems

- **Train** on training data
- **Validate** the model on "validation" or "development data
- **Test** the model on unseen data according to a task-specific evaluation metric

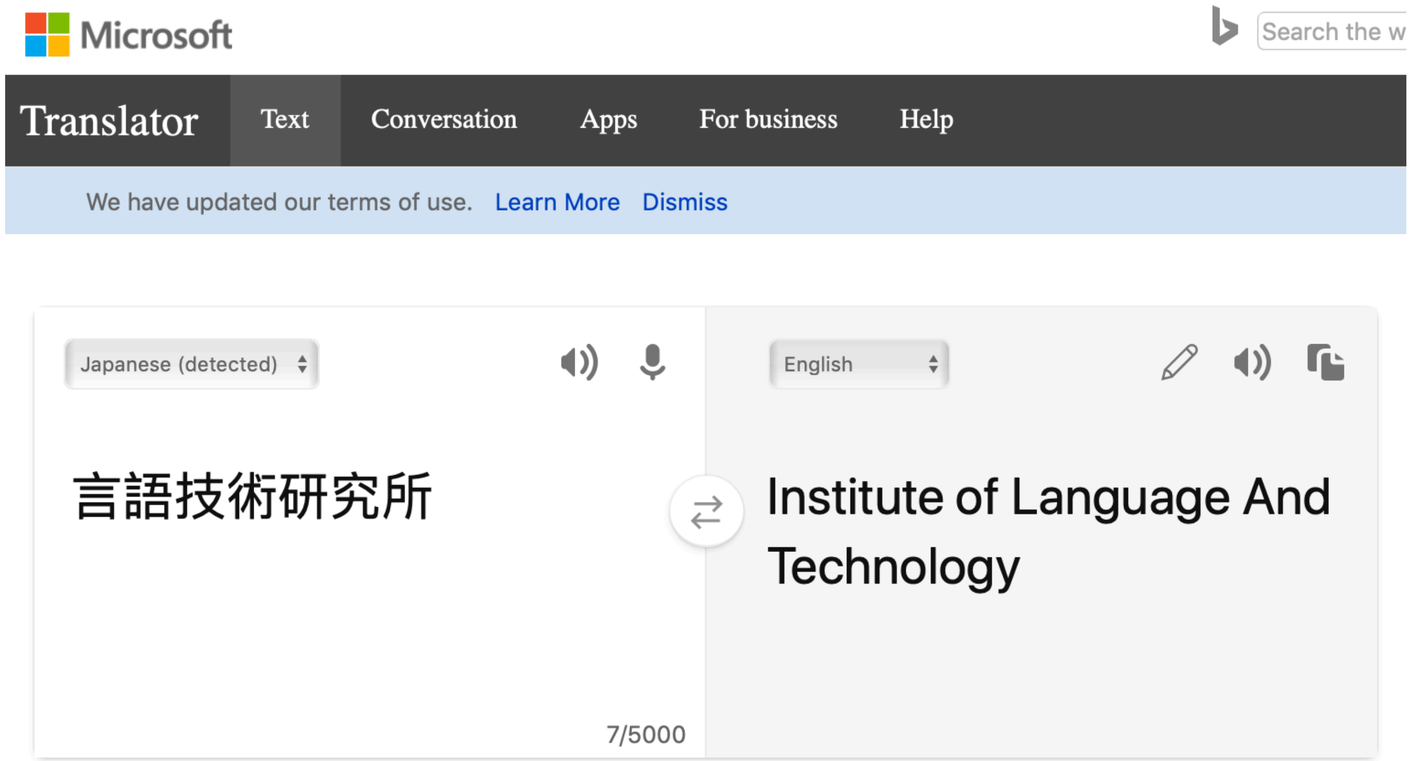
The Long Tail of Data



Aiding Human-Human Communication

Machine Translation

Input *X*	Output *Y*	Task
Text	Text in Other Language	Translation



Machine Translation Data

Last year I showed these two slides so that demonstrate that the arctic ice cap, which for most of the last three million years has been the size of the lower 48 states, has shrunk by 40 percent.
去年 この２つのスライドをお見せして 過去３百万年 アラスカとハワイを除く米国と— 同じ面積があった極域の氷河が 約40%も縮小したことが おわかりいただけたでしょう

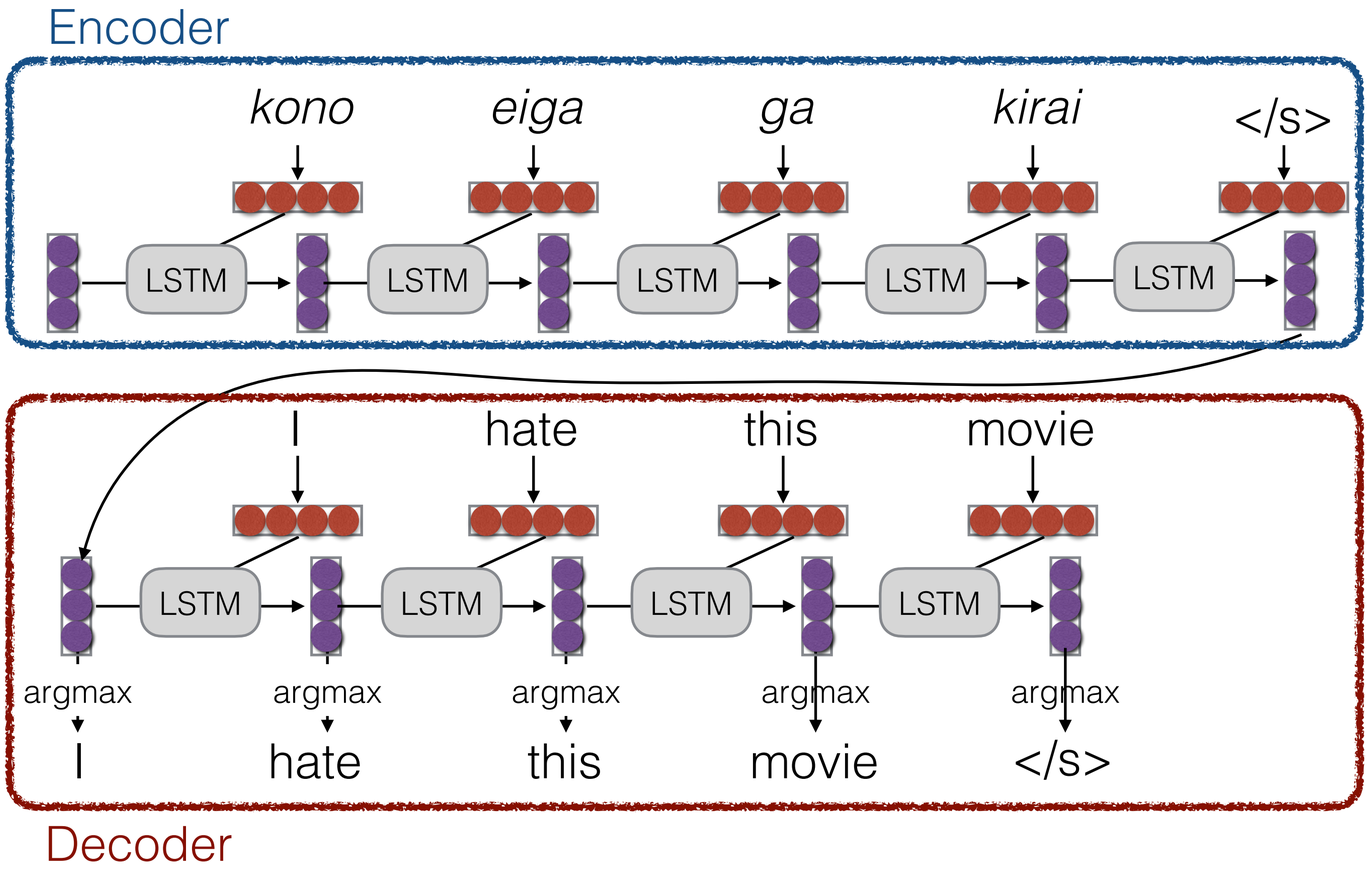
But this understates the seriousness of this particular problem because it doesn't show the thickness of the ice.
しかし もっと深刻な問題というのは 実は氷河の厚さなのです

The arctic ice cap is, in a sense, the beating heart of the global climate system.
極域の氷河は 言うなれば 世界の気候システムの鼓動する心臓で

It expands in winter and contracts in summer.
冬は膨張し夏は縮小します

The next slide I show you will be a rapid fast-forward of what's happened over the last 25 years.
では 次のスライドで 過去25年の動きを早送りにして見てみましょう

MT Modeling Pipeline



FAIRSEQ

https://github.com/pytorch/fairseq

Joey NMT

https://github.com/joeynmt/joeynmt

Naturally Occuring Sources of MT Data

- Compared to other NLP tasks, data relatively easy to find!
  - **News:** Local news, BBC world service, Voice of America
  - **Government Documents:** Governments often mandate translation
  - **Wikipedia:** Some Wikipedia articles are translated into many languages, identify and
  - **Subtitles:** Subtitles of movies and TED talks
  - **Religious Documents:** Bible, Jehova's Witness publications

http://opus.nlpl.eu/

MT Evaluation Metrics

- Two varieties of evaluation:

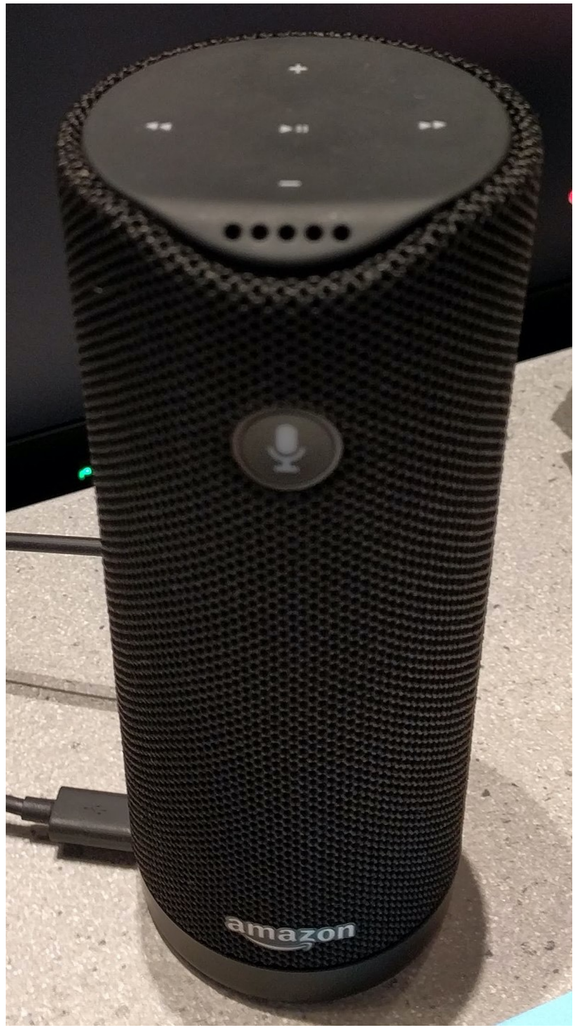
- **Manual Evaluation:** Ask a human annotator how good they think the translation is, including fluency (how natural is the grammar), adequacy (how well does it convey meaning)

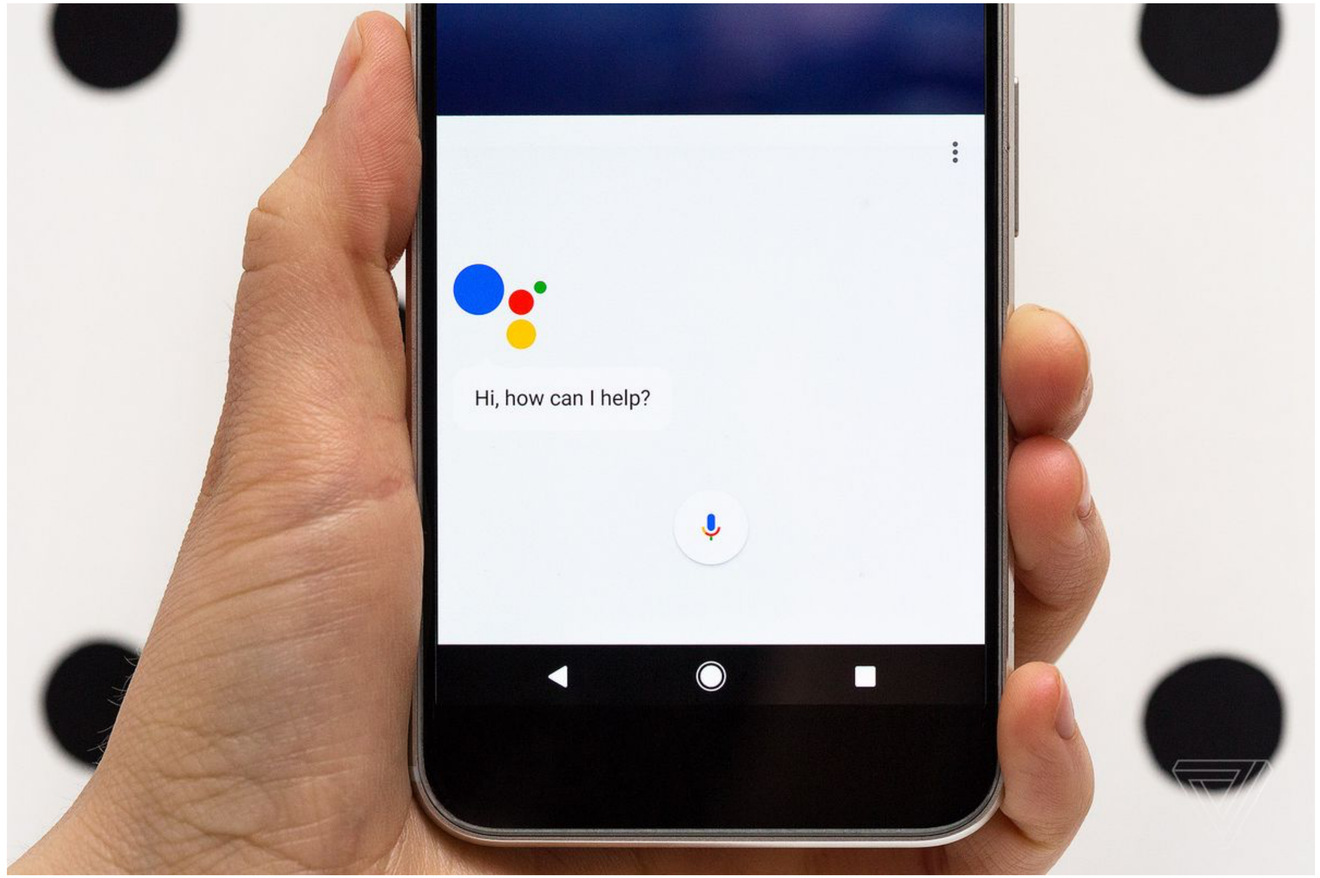
- **Automatic Evaluation:** Compare the output to a reference output for lexical overlap (BLEU, METEOR), or attempt to match semantics (MEANT, BERTscore)

Translation	Fluency	Adequacy	Overlap
please send this package to Pittsburgh	high	high	perfect
send my box, Pitsburgh	low	medium	low
please send this package to Tokyo	high	low	high
I'd like to deliver this parcel, destination Pittsburgh	high	high	low

Aiding Human-Machine Communication

Personal Assistants





Personal Assistant Pipeline

Speech Recognition

what is the weather in Pittsburgh now?

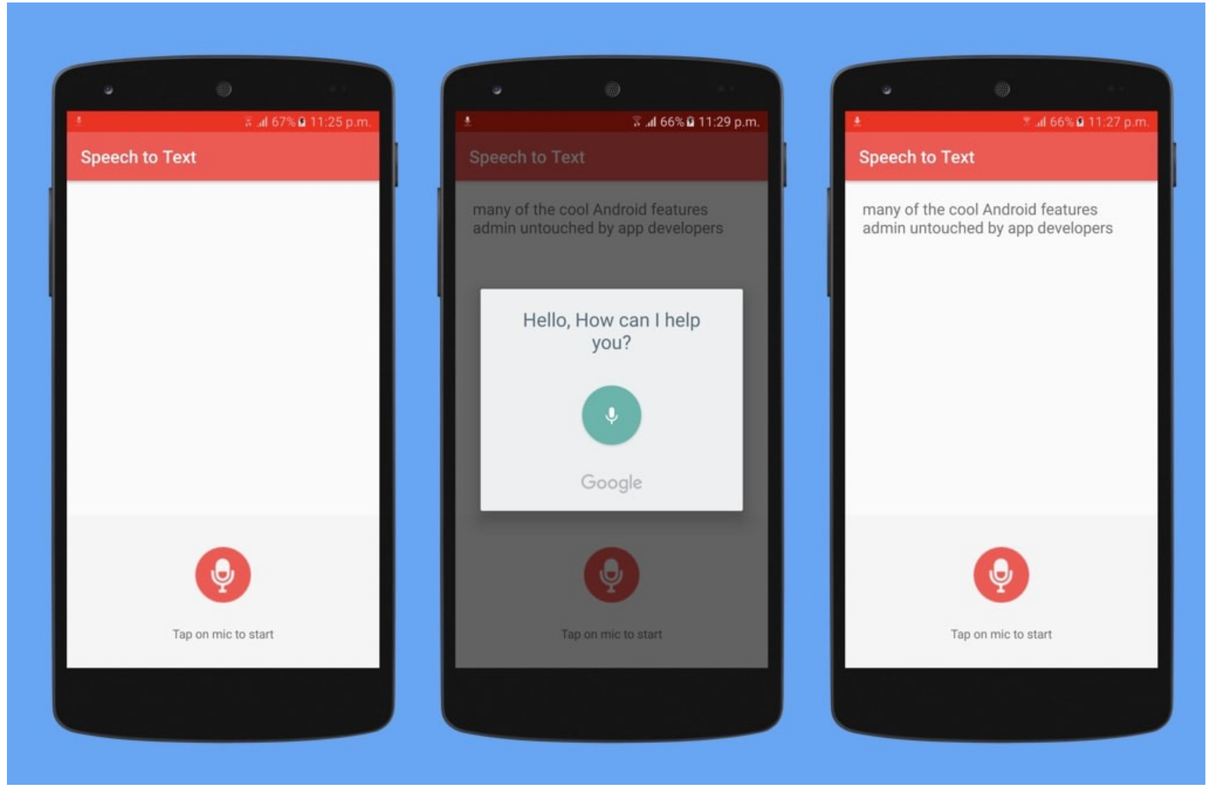
Question Answering,
etc.

75 degrees and sunny

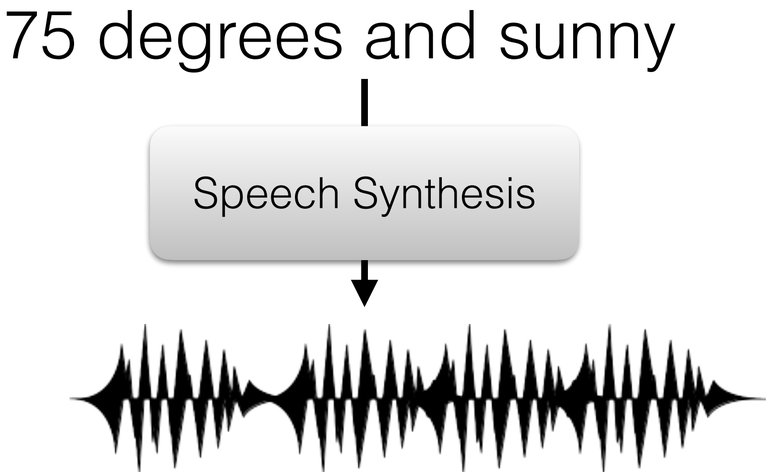
Speech Synthesis

Speech

Input *X*	Output *Y*	Task
Speech	Text	Speech Recognition



Input *X*	Output *Y*	Task
Text	Speech	Speech Synthesis



Speech Data

Last year I showed these two slides so that demonstrate that the arctic ice cap, which for most of the last three million years has been the size of the lower 48 states, has shrunk by 40 percent.

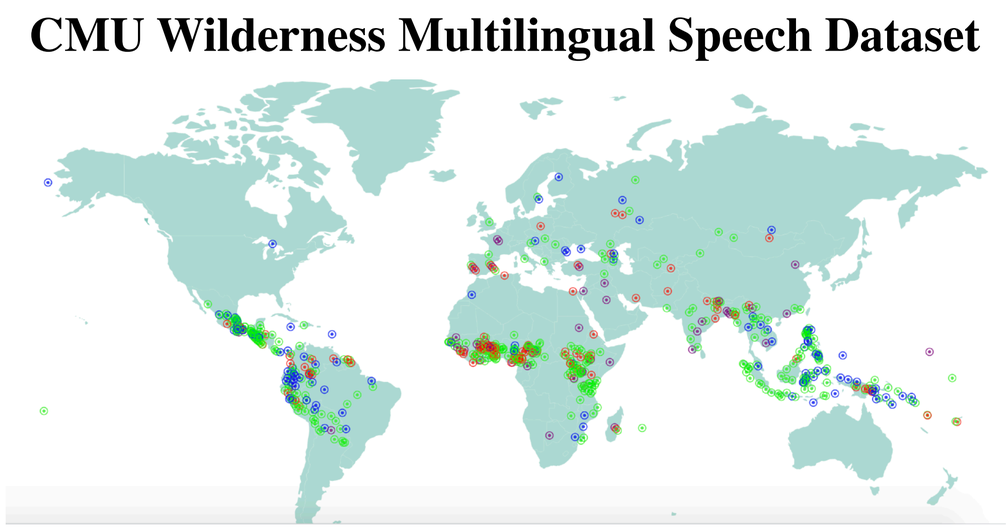
But this understates the seriousness of this particular problem because it doesn't show the thickness of the ice.

The arctic ice cap is, in a sense, the beating heart of the global climate system.

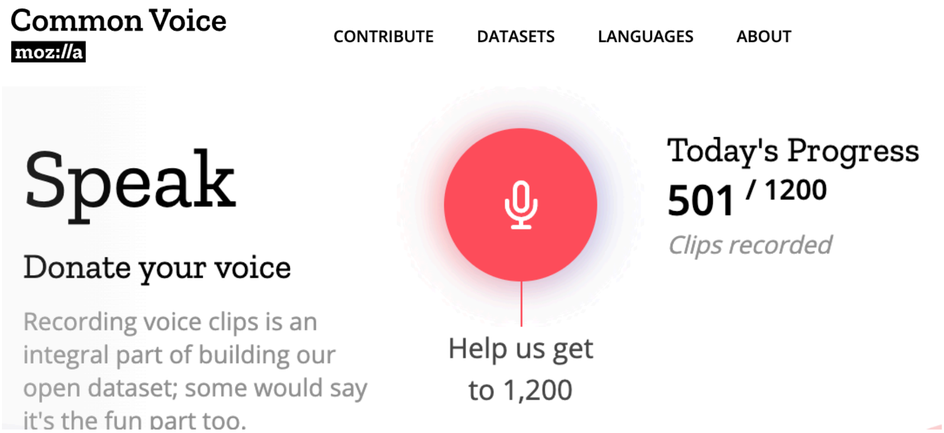
- **Speech Recognition:** Multi-speaker, noisy, conversational best for robustness
- **Speech Synthesis:** Single-speaker, clean, clearly spoken best for clarity

Naturally Occurring Sources of Speech Data

- **Transcribed News:** Sometimes spoken radio news also has transcriptions
- **Audio Books:** Regular audio books or religious books
- **Subtitled Talks/Videos:** TED(x) talks or YouTube videos often have transcriptions
- **Manually Transcribed Datasets:** Record speech you want and manually transcribe yourself (e.g. CallHome)



https://github.com/festvox/datasets-CMU_Wilderness



https://voice.mozilla.org/en

Speech Recognition Modeling Pipeline

- **Feature Extraction:** Convert raw wave forms to features, such as frequency features
- **Speech Encoder:** Run through an encoder (often reduce the number of frames)
- **Text Decoder:** Decode using sequence-to-sequence model or special-purpose decoder such as CTC

https://github.com/espnet/espnet

ASR Evaluation Metrics

- **Automatic evaluation:** word error rate

C=correct, S=substitution, D=deletion, I=insertion

correct:	this	is	some	recognized				speech
recognized:	this		some	wreck		a	nice	speech
type:	C	D	C	S		I	I	C

WER = (S + D + I) / reference length

(2+1+1)/5 = 80%

Speech Synthesis Modeling Pipeline

- **Text Encoder:** Encode text into representations for downstream use
- **Speech Decoder:** Predicts features of speech, such as frequency
- **Vocoder:** Turns spoken features into a waveform

ASR Evaluation Metrics

- **Automatic evaluation:** word error rate

C=correct, S=substitution, D=deletion, I=insertion

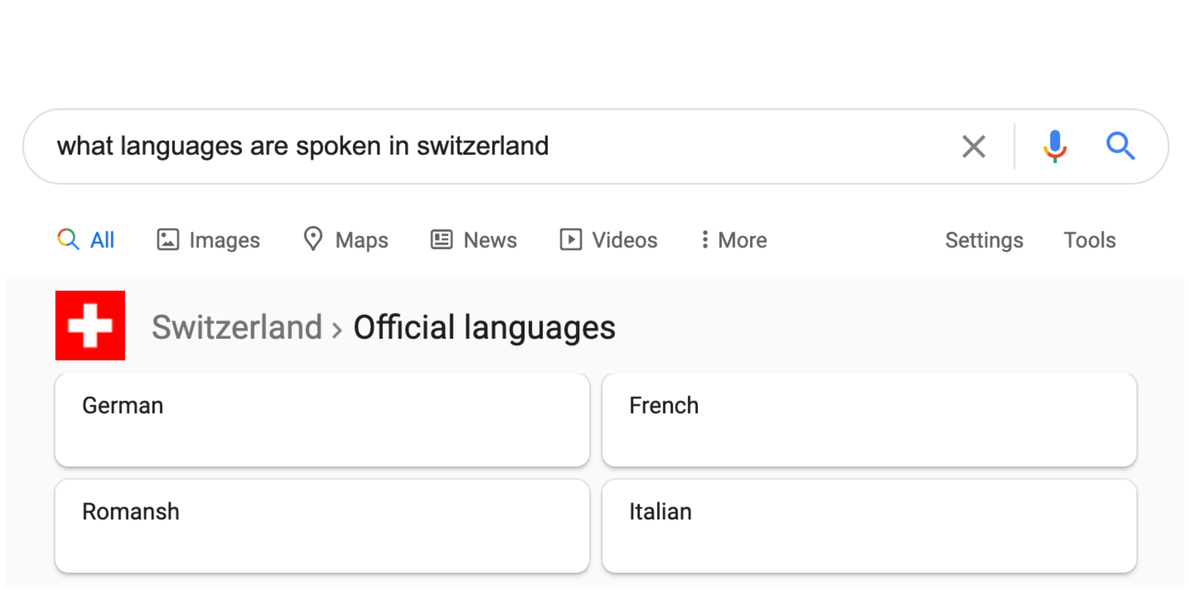
correct:	this	is	some	recognized				speech
recognized:	this		some	wreck		a	nice	speech
type:	C	D	C	S		I	I	C

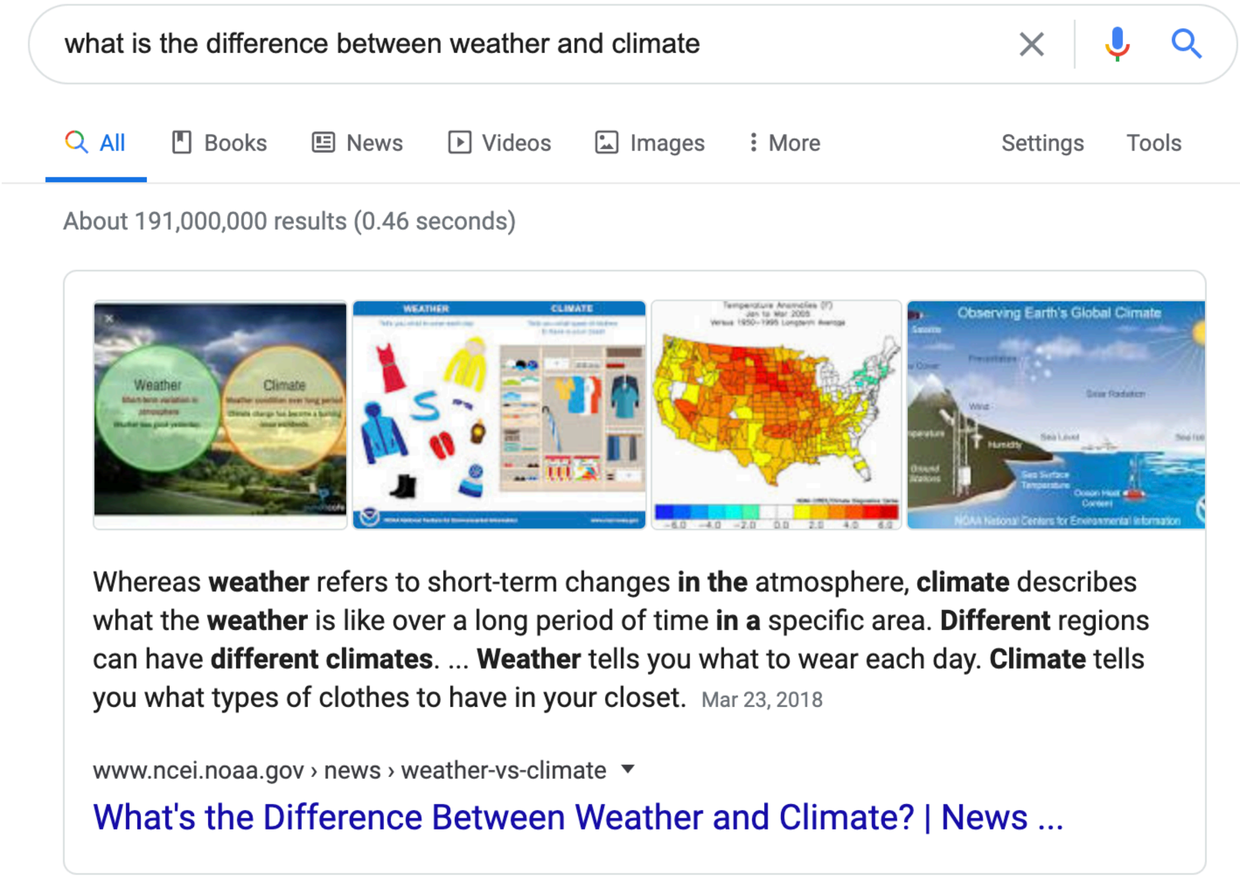
WER = (S + D + I) / reference length

(2+1+1)/5 = 80%

Question Answering

Input *X*	Output *Y*	Task
Textual Question	Answer	Question Answering

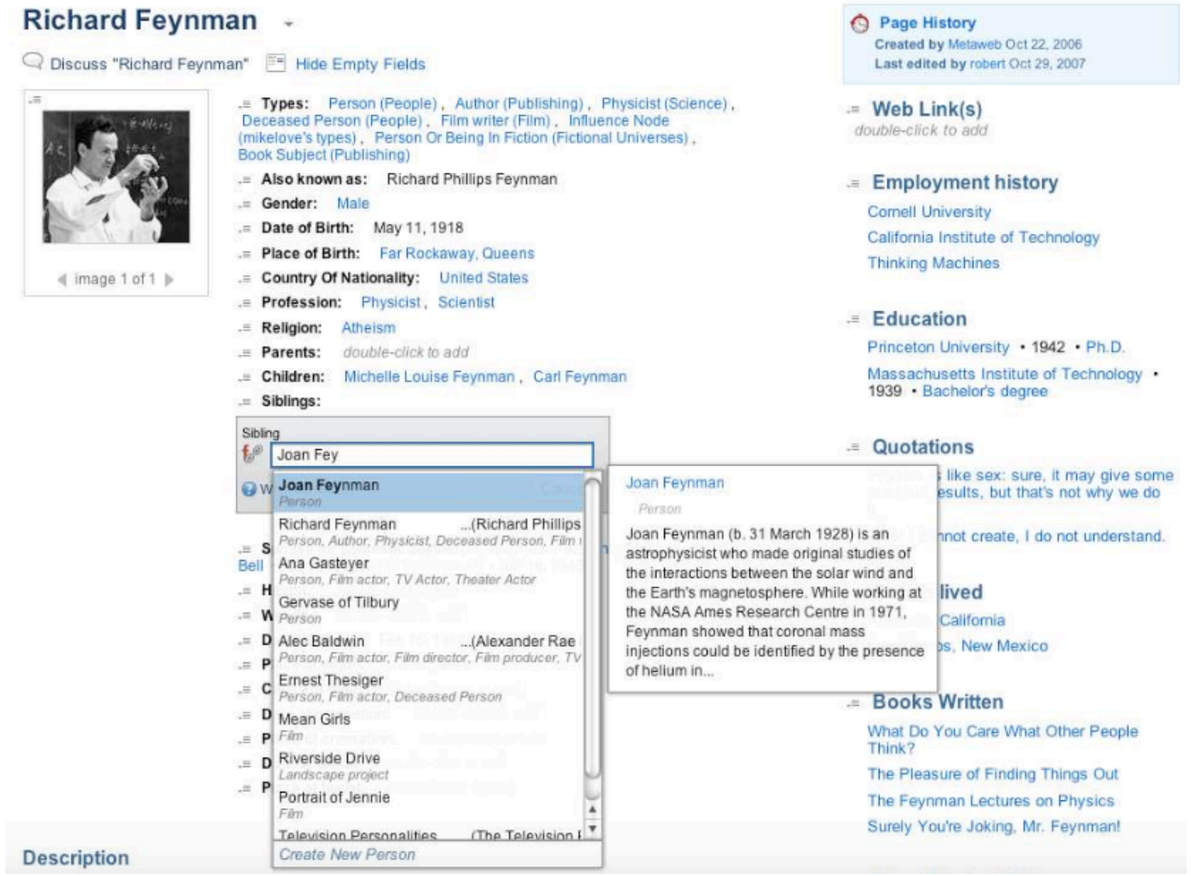




QA over Knowledge Bases

QA over Text

Example Knowledge Base: WikiData

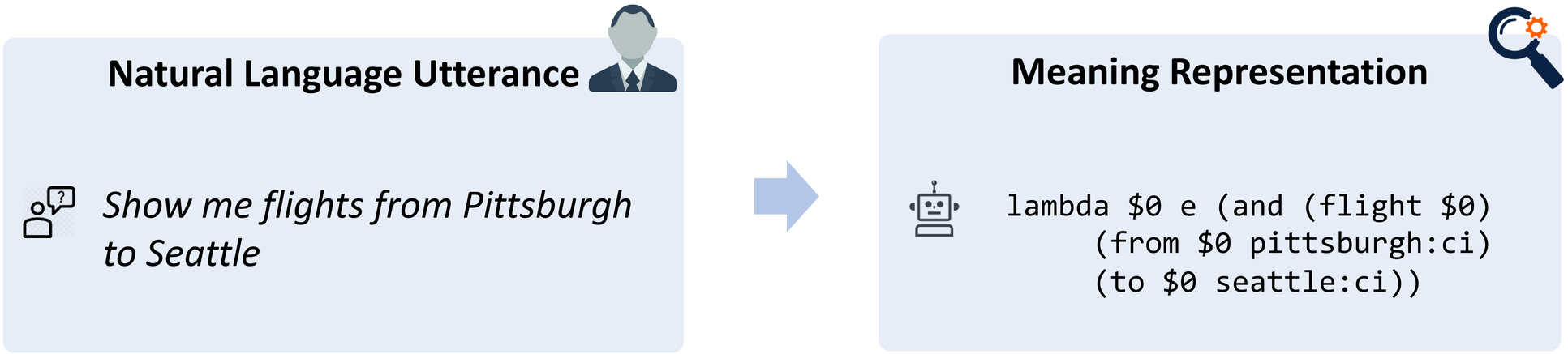


Language	Label	Description	Also known as
English	Richard Feynman	American theoretical physicist	Richard Phillips Feynman Richard P. Feynman Ofey
Japanese	リチャード・P・ファインマン	アメリカ合衆国の物理学者	Feynman ファインマン リチャード・ファインマン
Spanish	Richard Feynman	físico estadounidense y premio Nobel	Richard Phillips Feynman Feynman Richard P. Feynman Richard P Feynman
Traditional Chinese	No label defined	No description defined	
Afrikaans	Richard Feynman	No description defined	
Amharic	ሪቻርድ ፋይንማን	No description defined	ሪቻርድ ፌይንማን
Aragonese	Richard Feynman	No description defined	
Arabic	ريتشارد فاينمان	فيزيائي امريكي	ريتشارد فينمان
Egyptian Arabic	ريتشارد فاينمان	No description defined	
Assamese	ৰিচার্ড ফাইনমেন	No description defined	
Asturian	Richard Feynman	No description defined	
Azerbaijani	Riçard Feynman	No description defined	
South Azerbaijani	ریچارد فاینمن	No description defined	
Belarusian	Рычард Філіпс Фейнман	No description defined	Рычард Фейнман Рычард Файнман
Bulgarian	Ричард Файнман	Американски физик	Ричард Филипс Фейнман Ричард Фейнман Фейнмън Фейнман Ричард Филипс Файнмън Ричард Файнмън Ричард Фейнмън Файнмън

https://www.wikidata.org/

Semantic Parsing

- The process of converting natural language to a more abstract, and often operational semantic representation

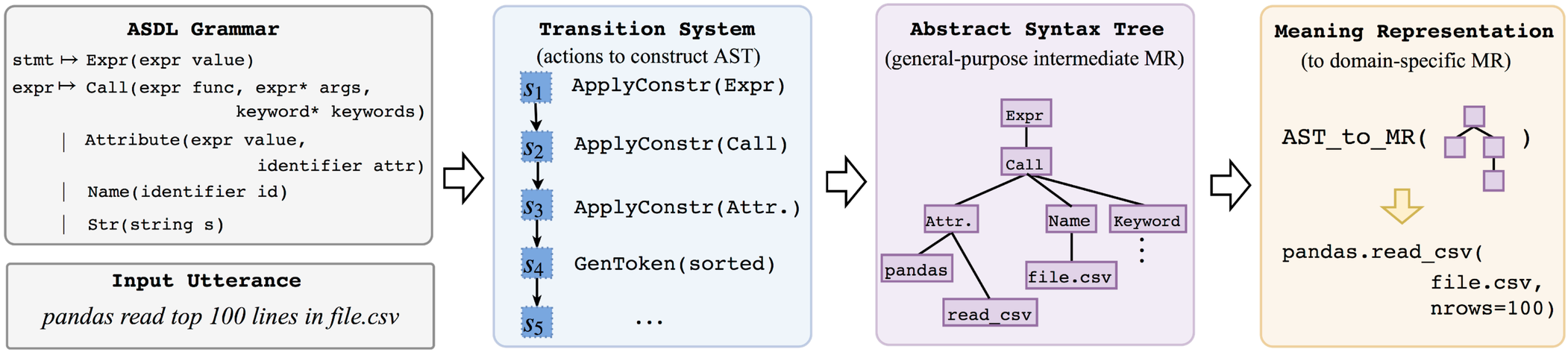


- These can be used to query **databases** (SQL), **knowledge bases** (SPARQL), or even generate **programming code** (Python)

Semantic Parsing Modeling Pipeline

- **Text Encoder:** Encode text into representations for downstream use
- **Tree/Graph Decoder:** Predict a tree or graph structured

TranX

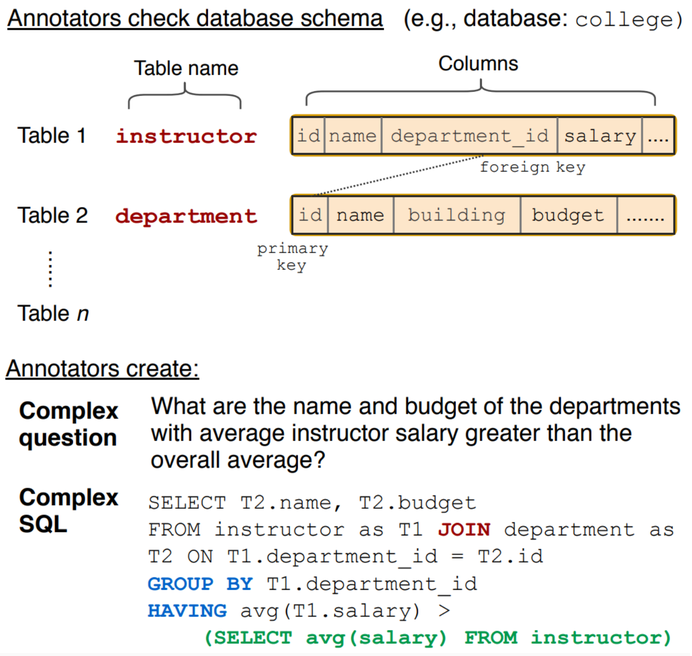


https://github.com/pcyin/tranX

Semantic Parsing Datasets

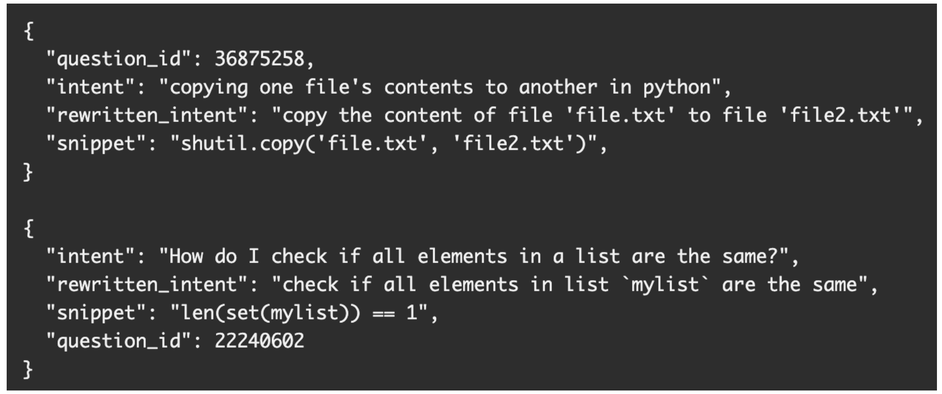
- **Text-to-SQL:** WikiSQL, Spider datasets
- **Text-to-knowledge graph:** WebQuestions, ComplexWebQuestions
- **Text-to-program:** CoNaLa, CONCODE

Spider



https://yale-lily.github.io/spider

CoNaLa



https://conala-corpus.github.io/

Example Tasks/Datasets for QA over Text

Span Selection (SQuAD)

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

Multiple Choice (MCTest)

James the Turtle was always getting in trouble. Sometimes he'd reach into the freezer and empty out all the food. Other times he'd sled on the deck and get a splinter. His aunt Jane tried as hard as she could to keep him out of trouble, but he was sneaky and got into lots of trouble behind her back.
One day, James thought he would go into town and see what kind of trouble he could get into. He went to the grocery store and pulled all the pudding off the shelves and ate two jars. Then he walked to the fast food restaurant and ordered 15 bags of fries. He didn't pay, and instead headed home.
His aunt was waiting for him in his room. She told James that she loved him, but he would have to start acting like a well-behaved turtle.
After about a month, and after getting into lots of trouble, James finally made up his mind to be a better turtle.

1) What is the name of the trouble making turtle?
A) Fries
B) Pudding
C) James
D) Jane

2) What did James pull off of the shelves in the grocery store?
A) pudding
B) fries
C) food
D) splinters

Cloze (CNN Daily Mail)

Original Version

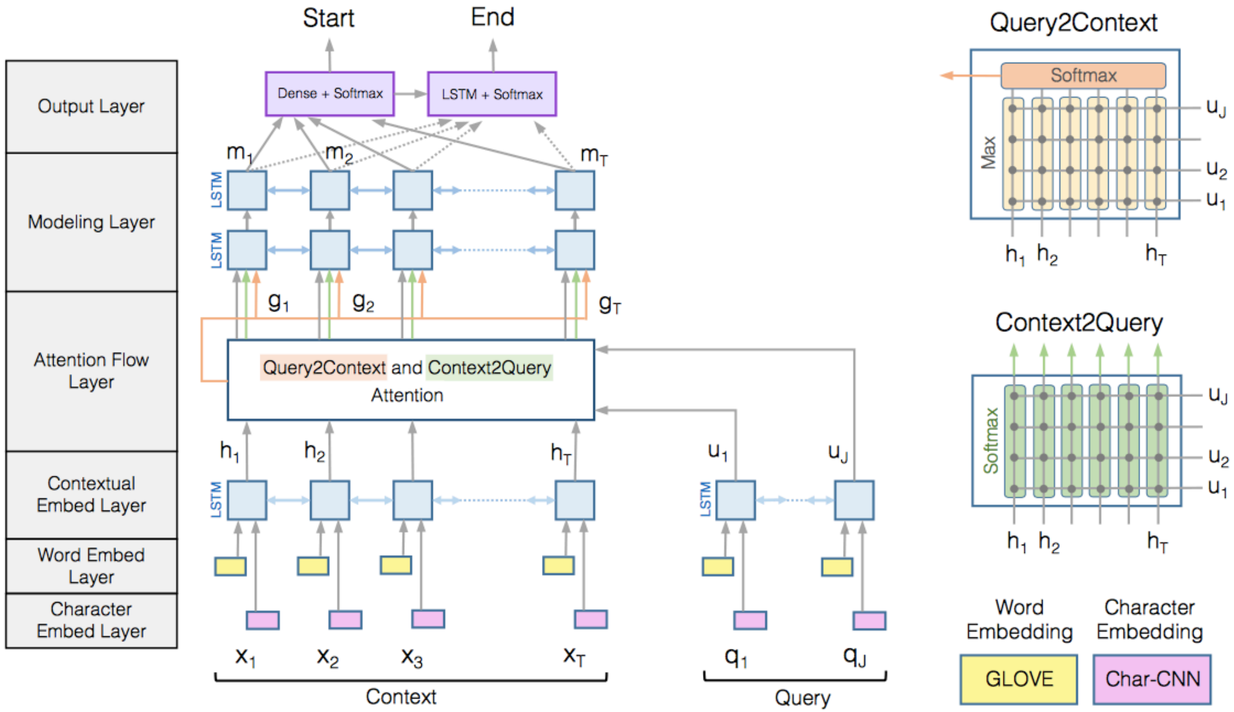
Context
The BBC producer allegedly struck by Jeremy Clarkson will not press charges against the "Top Gear" host, his lawyer said Friday. Clarkson, who hosted one of the most-watched television shows in the world, was dropped by the BBC Wednesday after an internal investigation by the British broadcaster found he had subjected producer Oisin Tymon "to an unprovoked physical and verbal attack." ...

Query
Producer **X** will not press charges against Jeremy Clarkson, his lawyer says.

Answer
Oisin Tymon

Machine Reading Modeling Pipeline

- **Document Encoder:** Encode text into representations for downstream use
- **Question Encoder:** Encode the question into some usable representation
- **Matcher:** Match between the input and output



https://github.com/allenai/bi-att-flow

Multilingual QA

- e.g. TyDiQA

Q: من هو موزارت ؟
mn hw mwzArt ?
Who is Muzart?
A: أماديوس موتسارت
>mAdyws mwtsArt
...Amadeus Mozart ...

Q: Kuka keksi viiko-n-päivä-t ?
who invented week-GEN-day-PL ?
Who invented the days of the week?

A: Seitsen-päivä-inen viikko on
seven-NOM-day-PL.ADJ week-NOM is
todennäköisesti lähtöisin Babylonia-sta...
likely origin Babylonia-ELA
The seven-day week is likely from Babylonia.

Q: Как далеко Уран от
how far Uranus-SG.NOM from
Земл-и?
Earth-SG.GEN?
How far is Uranus from Earth?

A: Расстояние между Уран-ом
distance between Uranus-SG.INSTR
и Земл-ёй меняется от 2,6
and Earth-SG.INSTR varies from 2,6
до 3,15 млрд км...
to 3,15 bln km...
The distance between Uranus and Earth fluctuates from 2.6 to 3.15 bln km...

LANGUAGE	LATIN SCRIPT[a]	WHITE SPACE TOKENS	SENTENCE BOUNDARIES	WORD FORMATION[b]	GENDER[c]	PRODROP
ENGLISH	+	+	+	+	+[d]	—
ARABIC	—	+	+	++	+	+
BENGALI	—	+	+	+	+	+
FINNISH	+	+	+	+++	—	—
INDONESIAN	+	+	+	+	—	+
JAPANESE	—	—	+	+	—	+
KISWAHILI	+	+	+	+++	—[e]	+
KOREAN	—	+[f]	+	+++	—	+
RUSSIAN	+	+	+	++	+	+
TELUGU	—	+	+	+++	+	+
THAI	—	—	—	+	+	+

https://github.com/google-research-datasets/tydiqa

Dialog Systems

Input *X*	Output *Y*	Task
Utterance	Response	Response Generation

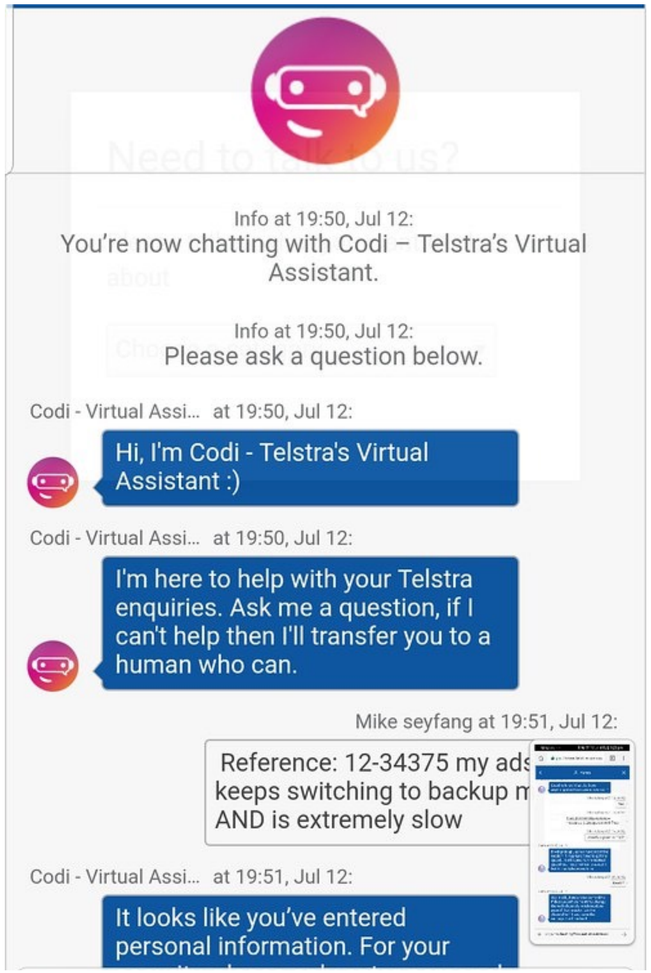


Image: Mike Seyfang @ flickr

Dialog System Data

Task-oriented Dialog

ASSISTANT:	How can I help you?
USER:	Hi, could you help me with booking movie tickets for tonight?
ASSISTANT:	What movie are you interested in?
USER:	The Upside.
ASSISTANT:	Did you have a theater in mind?
USER:	Could you check if the Regal Neshaminy... No, AMC Neshaminy in Neshaminy, PA is playing it?
ASSISTANT:	Could you spell that?
USER:	Sure, n e s h a m i n y.
ASSISTANT:	I have a showtime at 7:30 and at 10:30, is that okay?
USER:	Yes, could you get two tickets for the 7:30?
ASSISTANT:	One moment. Okay so that's 2 tickets for 7:30 at the AMC Neshaminy 24?
USER:	Yes.
ASSISTANT:	It'll be twenty-four ninety-nine for your tickets.
USER:	That sounds great.
ASSISTANT:	I've confirmed your tickets, they'll arrive via text shortly. Did you need any other information?
USER:	No, that was it. Thank you so much for your help.
ASSISTANT:	Great, no problem. I hope you have fun.
USER:	I hope so, too. Thank you so much.

Example from Taskmaster-1 Corpus (Byrne et al. 2019)

Chat

[PERSON 1:] Hi
[PERSON 2:] Hello ! How are you today ?
[PERSON 1:] I am good thank you , how are you.
[PERSON 2:] Great, thanks ! My children and I were just about to watch Game of Thrones.
[PERSON 1:] Nice ! How old are your children?
[PERSON 2:] I have four that range in age from 10 to 21. You?
[PERSON 1:] I do not have children at the moment.
[PERSON 2:] That just means you get to keep all the popcorn for yourself.
[PERSON 1:] And Cheetos at the moment!
[PERSON 2:] Good choice. Do you watch Game of Thrones?
[PERSON 1:] No, I do not have much time for TV.
[PERSON 2:] I usually spend my time painting: but, I love the show.

Example from Persona-Chat Corpus (Zhang et al. 2019)

Naturally Occurring Sources of Conversation Data

- **Human-Machine Dialog:** Let's Go!, CMU Communicator
- **Human-Human Constrained Dialog:** Map Task, Debates
- **Human-Human Spontaneous Dialog:** Switchboard, AMI Meeting Corpus
- **Human-Human Scripted Dialog:** Movie Dialog
- **Human-Human Written Dialog:** Twitter, Reddit, Ubuntu Chat

https://breakend.github.io/DialogDatasets/

Dialogue Modeling Pipeline

- **Context Encoder:** Encode the entire previous context
- (optionally) **Explicit Belief Tracking, Database Acccess**
- **Utterance Decoder:** Generate the output utterance

Dialogue Evaluation

- **Task Oriented Dialog:**
  - Task completion rate
  - Time to task completion
  - User satisfaction
- **Free-form Dialog**
  - Attempts to use overlap with reference utterances not successful
  - Largely resort to human evaluation

Language Analysis

Text Classification

Input *X*	Output *Y*	Task
Text	Label	Text Classification

Input *X*	Output *Y*	Task
Text Pair	Label	Text Pair Classification

Sentiment Analysis

I hate this movie → very good / good / neutral / bad / very bad

I love this movie → very good / good / neutral / bad / very bad

Textual Entailment

The woman bought a sandwich for lunch
→ The woman bought lunch
⟶ Entailment

The woman bought a sandwich for lunch
→ The woman did not buy a sandwich
⟶ Contradiction

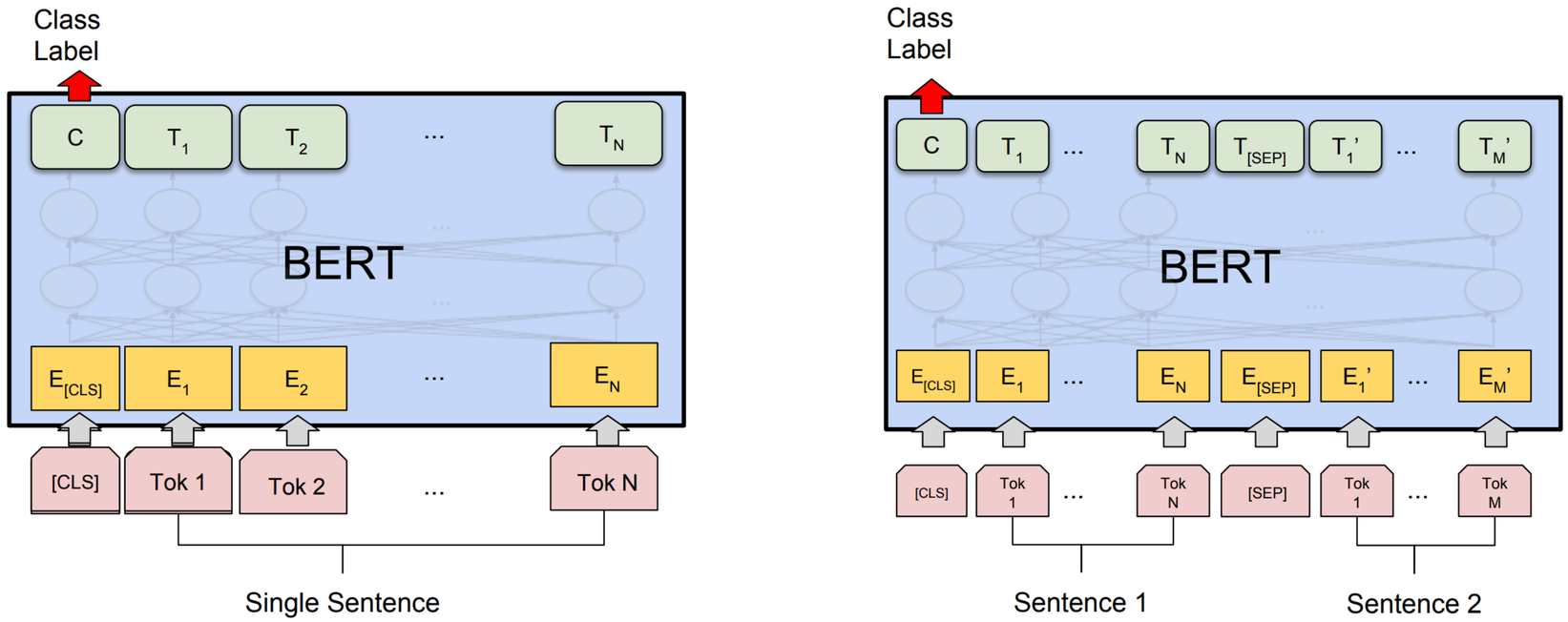
Text Classification Datasets

- **Sentiment Analysis:** Stanford sentiment treebank, Amazon reviews
- **Topic Classification:** 20 newsgroups, Wiki-500k
- **Paraphrase Identification:** Microsoft Paraphrase Corpus
- **Textual Entailment:** Stanford/Multi Natural Language Inference
- many many others!

Text Classification Pipeline

- **Text Encoder:** Encode the text you want to analyze
- **Predictor:** Predict using a label

Example: BERT



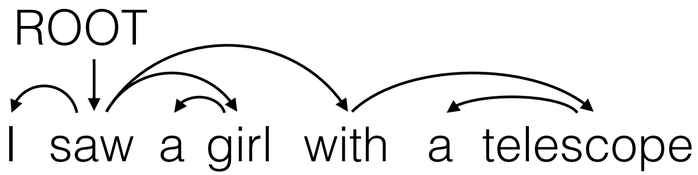
Text Analysis Tasks

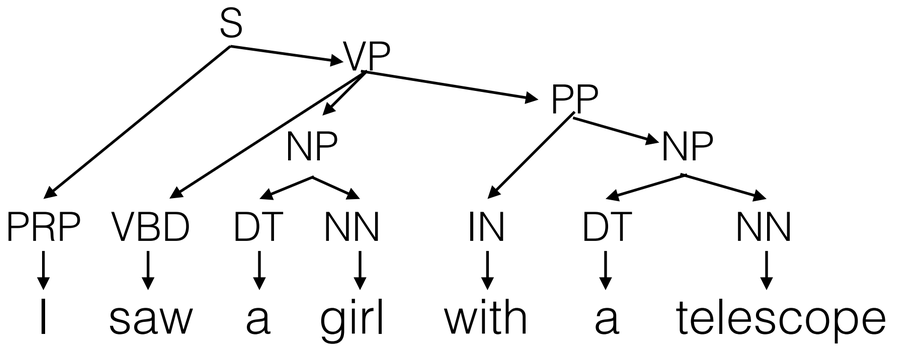
Input *X*	Output *Y*	Task
Text	Per-word Tags	Sequence Labeling

I	watched	the	movie
PRN	VBD	DET	NN

CMU	is	in	Pittsburgh
ORG	X	X	LOC

Input *X*	Output *Y*	Task
Text	Syntax Trees	Syntactic Parsing





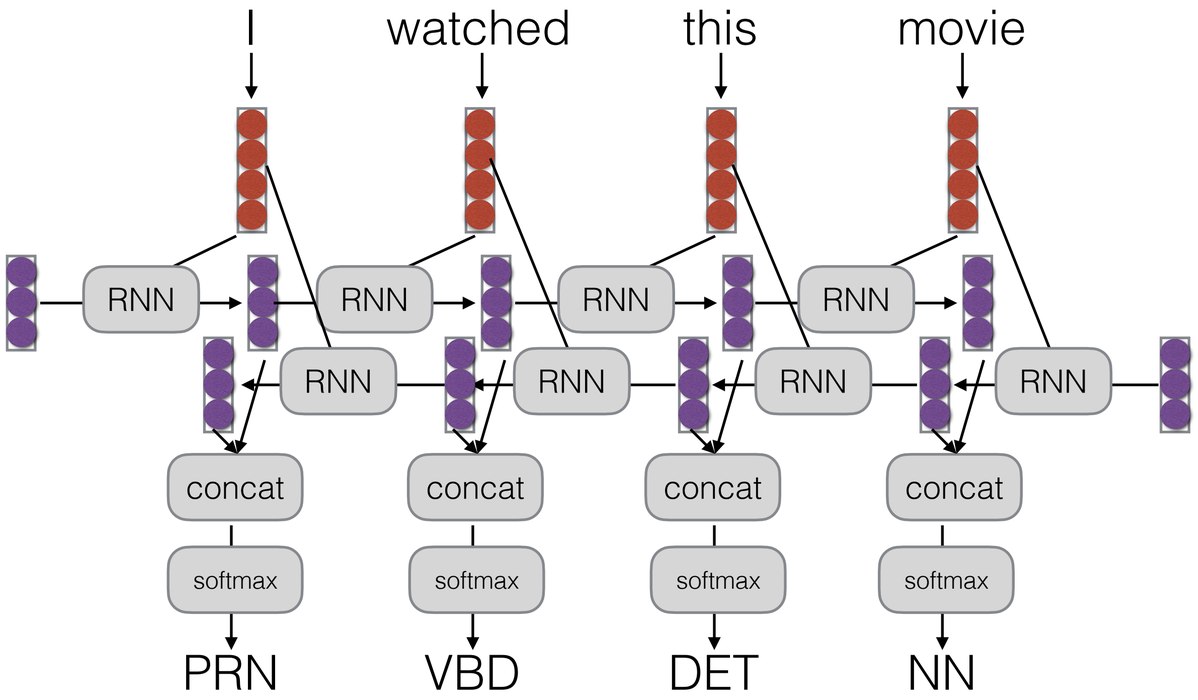
Text Analysis Data

- **OntoNotes:** A large corpus with many different annotations in English.
- **Universal Dependencies Treebank:** Includes parts-of-speech and dependency trees for many languages

	Abaza	1	3K		Northwest Caucasian
	Afrikaans	1	49K		IE, Germanic
	Akkadian	1	1K		Afro-Asiatic, Semitic
	Albanian	1	<1K		IE, Albanian
	Amharic	1	10K		Afro-Asiatic, Semitic
	Ancient Greek	2	416K		IE, Greek
	Arabic	3	1,042K		Afro-Asiatic, Semitic
	Armenian	1	52K		IE, Armenian
	Assyrian	1	<1K		Afro-Asiatic, Semitic
	Bambara	1	13K		Mande
	Basque	1	121K		Basque
	Belarusian	1	13K		IE, Slavic
	Bhojpuri	2	6K		IE, Indic
	Breton	1	10K		IE, Celtic
	Bulgarian	1	156K		IE, Slavic
	Buryat	1	10K		Mongolic
	Cantonese	1	13K		Sino-Tibetan

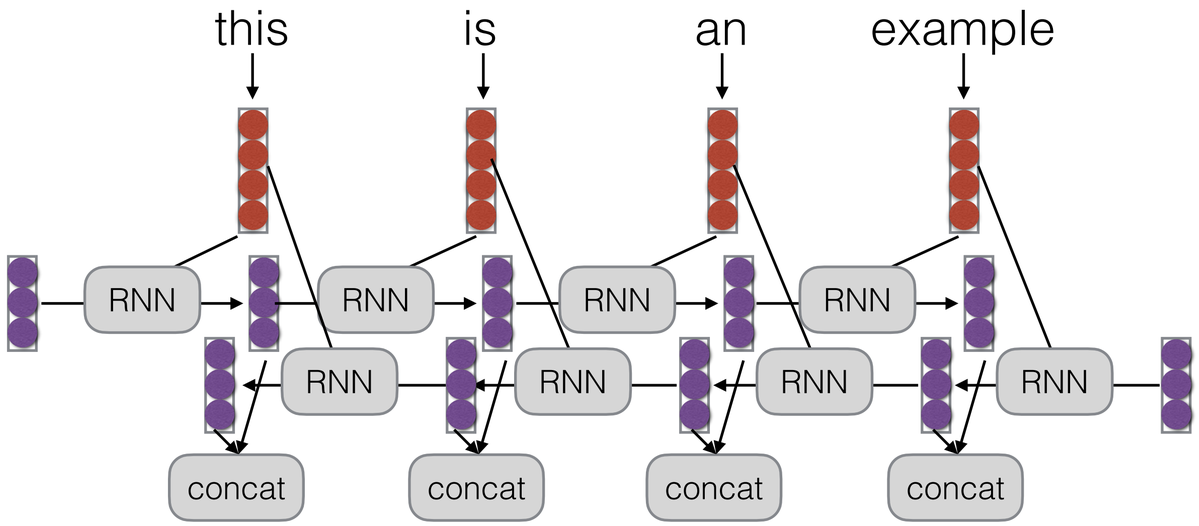
Sequence Labeling Pipeline

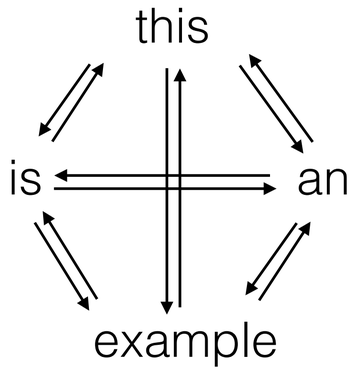
- **Text Encoder:** Encode the text you want to analyze
- **Sequence Labeler:** Predict each label independently, or in a joint fashion

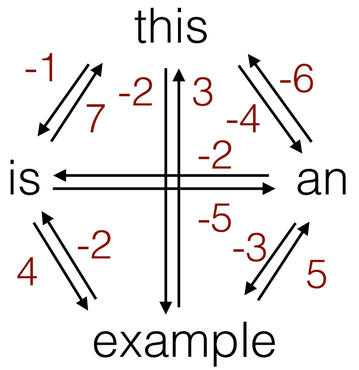


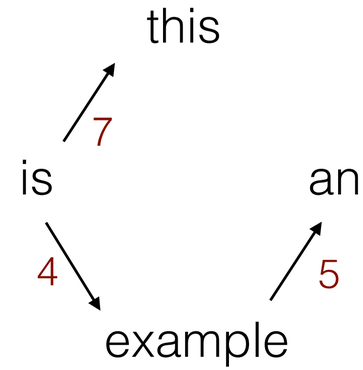
Syntactic Parsing Pipeline

- **Text Encoder:** Encode the text you want to analyze
- **Tree Generatior:** Generate dependency trees or constituency trees









Evaluation

- **POS Tagging:** Word-by-word accuracy
- **Entity Recognition:** Entity F-measure
- **Dependency Parsing:** Labeled or unlabeled attachment score
- **Constituency Parsing:** Phrase F-measure

Conclusion

Tackling NLP Tasks

- Data collection
- Modeling
- Train/test evaluation

Questions?