Introduction to Natural Language Processing

01a Introduction

Marc von Wyl
marc.von-wyl@epita.fr

About the course

- 2 modules of 4 sessions
- Half-theory, half-practice (outside the first class)
- Prerequisites
  - Python (3)
  - Basic linear algebra
  - Basic probabilities
  - Basic differential calculus
  - Basic machine learning
- If possible
  - Basic deep learning (2nd module)
  - Basic PyTorch (2nd module)

About the class

Graded on projects + MCQs

- 3 points on MCQs + 13 points on projects -> 16
  - 1 point per MCQ, 4 MCQs
  - More than 4 extra points available as bonus questions and going further than asked.
- Understanding of the methods used
- Data analysis
- Quality of code
- Results discussion and analysis
- Bonus on a language presentation
  - 5-10min, neither English nor French
  - Check the topic with me before

Your teacher

Marc von Wyl

- Engineer degree from the Engineering School of Geneva
  - Geneva (CH), 2003
- Master in Computer Science from the University of Geneva
  - Geneva (CH), 2009
- Research assistant at the University of Geneva
  - Geneva (CH), 2009-2012
  - Multimodal search based on the user's behaviour
- Speech Research Engineer à Autonomy HP
  - Cambridge (UK) 2012-2016
  - Language modelling applied to speech-to-text
- NLP Engineer > Lead NLP Engineer at Factmata
  - London (UK), 2019
  - Extract insight from large quantities of text data
- Senior Machine Learning Engineer at Algolia
  - Paris 2020-2022
  - Decompounding of Germanic languages (and Finnish)
  - Working on our "AI" features including
    - Dynamic synonyms suggestions
    - Dynamic re-ranking
    - Query categorization
- Senior NLP Engineer at Sightengine
  - France (remote) 2022-now
  - Multimodal moderation

Course content

Course content

First module

1. Introduction to linguistics
2. Text processing and sentiment analysis
3. Vector semantics and projections
4. Language modelling and applications

Second module

1. Encoder decoder and the transformer
2. Benchmarks, transfer learning 1.
   a. Contextual embeddings
3. Transfer learning 2.
4. NLP in real life.

What is Natural Language Processing?

What’s NLP?

Natural Language Processing, also called

- Computational linguistics
- Speech and language processing
- Human language technology

What is it for?

- Communication machine-human
- Improve human to human communication
- Analyse human generated text
- Support and improve people’s life and work
- Automatize tasks
- ...

What’s NLP?

Examples

- Information analysis
  - Information retrieval (search), text summarization, Q&A, sentiment analysis, …
- Better human-to-human communication
  - Machine translation, spell checking, E-mail rewriting, speech-to-text, text-to-speech

Multimodal applications

- Image/video description and search
- Image and video generation
- Speech-to-text, text-to-speech

Exemples: sentiment analysis

- “I love this movie”
  - **=> {'label': 'POSITIVE', 'score': 0.999...}**
- “This restaurant is disgusting”
  - **=> {'label': 'NEGATIVE', 'score': 0.999...}**
- “If you are looking for the best place to lose your faith in humanity and get a lovely gastro, I warmly recommend this restaurant”
  - **=> {'label': 'POSITIVE', 'score': 0.999...}**

Exemple: text generation

A unicorn was knocking at my door,

Exemple: text generation

A unicorn was knocking at my door, *his face was ashen with shock.*

The unicorn was not really an ordinary unicorn. He was a bit of an intimidating one, only standing about three meters and with slightly more than fifty strong horns.

Exemples: text summarization

“A major drawback of statistical methods is that they require elaborate feature engineering. Since 2015,[16] the field has thus largely abandoned statistical methods and shifted to neural networks for machine learning. Popular techniques include the use of word embeddings to capture semantic properties of words, and an increase in end-to-end learning of a higher-level task (e.g., question answering) instead of relying on a pipeline of separate intermediate tasks (e.g., part-of-speech tagging and dependency parsing). In some areas, this shift has entailed substantial changes in how NLP systems are designed, such that deep neural network-based approaches may be viewed as a new paradigm distinct from statistical natural language processing. For instance, the term neural machine translation (NMT) emphasizes the fact that deep learning-based approaches to machine translation directly learn sequence-to-sequence transformations, obviating the need for intermediate steps such as word alignment and language modeling that was used in statistical machine translation (SMT). Latest works tend to use non-technical structure of a given task to build proper neural network.[17]”

Exemples: text summarization

'The field of natural language processing (NLP) has undergone a major shift in recent years.'

Exemple: traduction automatique

- “How was your day?” > “तिम्रो दिन कस्तो भयो?”
- “Your book is on my desk” > “შენი წიგნი ჩემს მაგიდაზეა”
- “The nurse's name was Mike.” > “Le nom de l'infirmière était Mike.”

- “तिम्रो” is informal and rather rude for a foreigner (should be “तपाइको”).
- “შენი” is informal (should be “თქვენი”).
- As Mike is a man’s name, it should be “infirmier”.

A brief history of NLP

Symbolic NLP 1950 - early 1990s

- Alan Turing: Computing Machinery and Intelligence (Mind 49)
  - Can a machine think?
  - The Turing test
- 1954 The Georgetown experiment: machine translation of 60 sentences from Russian to English
  - Mainly based on lexical methods.
  - *" within three or five years, machine translation would be a solved problem"*
- 1966 ALPAC report
  - Evaluate progress in computational linguistics and especially in machine translation.
  - Caused a "dramatic" reduction of the allowed budget by the US government.

Symbolic NLP 1950 - early 1990s

- 1966: ELIZA a Rogerian psychotherapist simulation

User_1:	You are like my father in some ways.
ELIZA_1:	WHAT RESEMBLANCE DO YOU SEE
User_2:	You are not very aggressive but I think you don't want me to notice that.
ELIZA_2:	WHAT MAKES YOU THINK I AM NOT AGGRESSIVE
User_3:	You don't argue with me.
ELIZA_3:	WHY DO YOU THINK I DON'T ARGUE WITH YOU
User_4:	You are afraid of me.
ELIZA_4:	DOES IT PLEASE YOU TO BELIEVE I AM AFRAID OF YOU

- Only based on pattern matching, not using any knowledge base.
- Several users felt like ELIZA had human feelings.

Symbolic NLP 1950 - early 1990s

- 1970: conceptual ontology development.
  - Real world information structured in a language understandable by a machine.
  - MARGIES (Schank, 1975), SAM (Cullingford, 1978).
- SHRDLU (Winograd 1972), one of the first NLU applications (Natural Language Understanding).
  - Simulate a robot in a toy world.
  - Accepts "natural" instructions.
    - Person: Pick up a big red block.
    - Computer: OK.
    - Person: Grasp the pyramid.
    - Computer: I DON'T UNDERSTAND WHICH PYRAMID YOU MEAN.
    - Person: Find a block which is taller than the one you are holding and put it into the box.
    - Computer: BY "IT", I ASSUME YOU MEAN THE BLOCK WHICH IS TALLER THAN THE ONE I AM HOLDING.

Symbolic NLP 1950 - early 1990s

- Prolog (Colmerauer and Roussel 1972)
  - Declare rules as facts and ask questions to inference engine.

```
mother_child(trude, sally).

father_child(tom, sally).
father_child(tom, erica).
father_child(mike, tom).

sibling(X, Y)      :- parent_child(Z, X), parent_child(Z, Y).

parent_child(X, Y) :- father_child(X, Y).
parent_child(X, Y) :- mother_child(X, Y).
```

Symbolic NLP 1950 - early 1990s

- 1980s Empiricism
- Probability-based methods are back
  - Part-of-speech tagging
  - Parsing
  - Semantics
- Models are evaluated on held out data

Probability-based 1990-2000

- Probability-based methods become the new standard
- They are used in every application
- Evaluation methodology is standardized
- The increase in computers’ capacity brings an increase of commercial applications
  - Speech recognition (Dragon Dictate 1990)
  - Spelling and grammar check
- Internet is invented and, with it, the need for search engines.

2000-2010 The Machine Learning Era

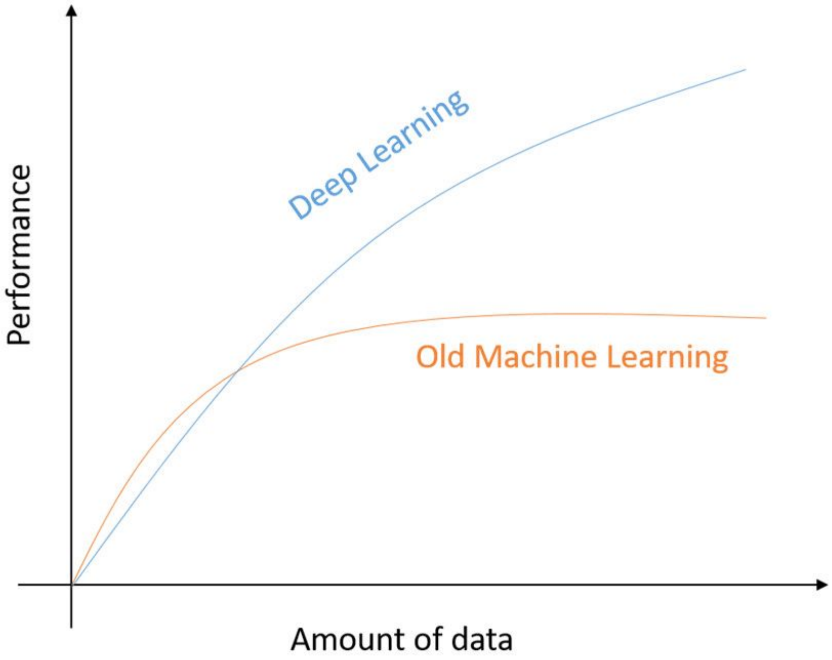
- Increase of annotated data
  - Penn treebank (1991), Prague Dependency Treebank (1998), PropBank (2005), …
- Which brings a surge of supervised methods
  - Bayesian models
  - Maximum entropy techniques (e.g. decision trees)
  - Support Vector Machines (Vapnik 1995)
- Most of these methods are “resurrected” on fresh data and applications.

2010s Deep Learning

- A logical calculus of ideas immanent in nervous activity (McCulloch and Pitts, 1943)
- Recurrent Neural Network (Mikolov et al. 2010)
- Word2Vec (Mikolov et al. 2013)
- Deep contextual embeddings
  - ELMo (Peters et al. 2018) and ULMFiT (Howard and Ruder 2018)
- The Transformer (Vaswani et al. 2017)
- BERT (Devlin et al. 2018)

Why is deep learning so prevalent?

- Change in technology (GPU et TPU)
- Improvement in theory
  - Better activation functions
  - Better loss and optimization functions
  - **Residual connections**
- Libraries and democratization
  - Theano (2009)
  - Caffe (2013)
  - Tensorflow (2015)
  - Torch (2002) -> pyTorch (2016)
  - JAX (2018?)
- Data



State of the art

The BERT

- BERT: Pre-
Understand
- Tran

Training Size	
Human	
BERT	
BigBird	

Zho		-	-
Yan		-	-
Che		-	-
Xu a		-	-
Yan		-	-
Ma		-	-
Gon		-	-
		90.4	95.7
Ou		**93.1**	**97.0**
Ours		92.7	**97.0**

Don’t believe the hype

- Neural networks are low bias learner. They like shortcuts.
- They are black-box models, making it hard understanding what they truly learn.
- In 2019, another publication on BERT
  - BERT’s results go **77% accuracy down to random** by removing simple cue words (“not”, “is”, “do”,...)
  - NLP clever Hans moment
- Another publication (July 2020)
  - BERT always predict a negative sentiment for “I am a {PROTECTED} {NOUN}.” when {PROTECTED} is black, atheist, gay, and lesbian, while predicting positive for Asian, straight, etc
  - Replacing “before” with “after” changed sentiment prediction in 100% of cases…
  - In QA, “{P1} is not a doctor, {P2} is.”, “Who is a doctor?” if {P1} is a male name and {P2} a female name, the model fails 89.1% of the time.

Real effects on people life

- Automatic translate option of Facebook lands Palestinian man in jail
  - Translated “Good morning” to “Hurt them”.
  - The man spent a day in jail before the mistake was discovered.
- Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings (2016)
- Racial Bias and Gender Bias Examples in AI systems
  - *Through COMPAS, black offenders were seen almost twice as likely as white offenders to be labeled a higher risk but not actually re-offend.*

Current efforts

- Harder benchmarks with less bias.
  - SuperGLUE, Winograd, …
- Automatic bias removal.
- Better understanding of what the models learn.
- Form vs meaning debat
  - Using “understanding” in academic papers
  - Chinese room experiment / octopus test
- Model compression/distillation
- **Better and more data -> better model**
  - Data analysis, hard negatives, RLHF, …
- But still, much better than what we used to do
  - Using word embeddings >= not using them
  - Using contextual word embeddings >= word embeddings

State of the art

- On GLUE (General Language Understanding Evaluation) benchmark

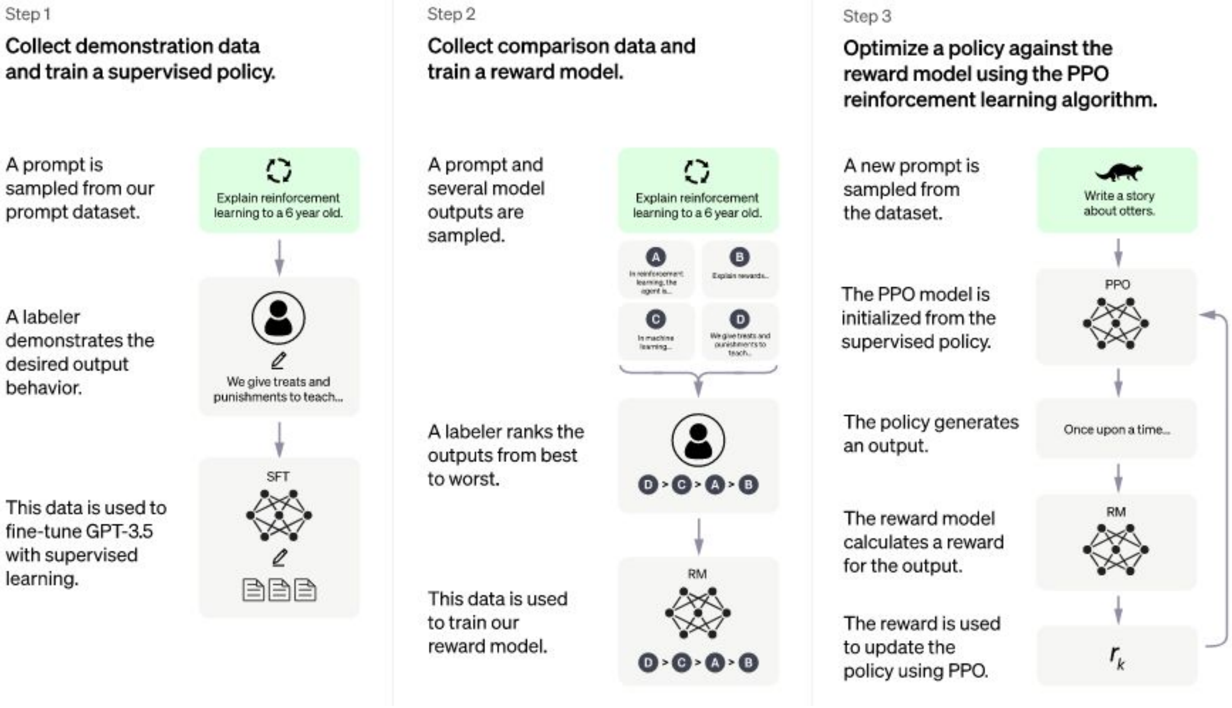
Rank	Name	Model	URL	Score	CoLA	SST-2	MRPC	STS-B	QQP	MNLI-m	MNLI-mm	QNLI	RTE	WNLI
1	ERNIE Team - Baidu	ERNIE		90.9	74.4	97.8	93.9/91.8	93.0/92.6	75.2/90.9	91.9	91.4	97.3	92.0	95.9
2	DeBERTa Team - Microsoft	DeBERTa / TuringNLRv4		90.8	71.5	97.5	94.0/92.0	92.9/92.6	76.2/90.8	91.9	91.6	99.2	93.2	94.5
3	HFL iFLYTEK	MacALBERT + DKM		90.7	74.8	97.0	94.5/92.6	92.8/92.6	74.7/90.6	91.3	91.1	97.8	92.0	94.5
4	Alibaba DAMO NLP	StructBERT + TAPT		90.6	75.3	97.3	93.9/91.9	93.2/92.7	74.8/91.0	90.9	90.7	97.4	91.2	94.5
5	PING-AN Omni-Sinitic	ALBERT + DAAF + NAS		90.6	73.5	97.2	94.0/92.0	93.0/92.4	76.1/91.0	91.6	91.3	97.5	91.7	94.5
6	T5 Team - Google	T5		90.3	71.6	97.5	92.8/90.4	93.1/92.8	75.1/90.6	92.2	91.9	96.9	92.8	94.5
7	Microsoft D365 AI & MSR AI & GATECH	MT-DNN-SMART		89.9	69.5	97.5	93.7/91.6	92.9/92.5	73.9/90.2	91.0	90.8	99.2	89.7	94.5
8	Huawei Noah's Ark Lab	NEZHA-Large		89.8	71.7	97.3	93.3/91.0	92.4/91.9	75.2/90.7	91.5	91.3	96.2	90.3	94.5
9	Zihang Dai	Funnel-Transformer (Ensemble B10-10-10H1024)		89.7	70.5	97.5	93.4/91.2	92.6/92.3	75.4/90.7	91.4	91.1	95.8	90.0	94.5
10	ELECTRA Team	ELECTRA-Large + Standard Tricks		89.4	71.7	97.1	93.1/90.7	92.9/92.5	75.6/90.8	91.3	90.8	95.8	89.8	91.8
11	Microsoft D365 AI & UMD	FreeLB-RoBERTa (ensemble)		88.4	68.0	96.8	93.1/90.8	92.3/92.1	74.8/90.3	91.1	90.7	95.6	88.7	89.0
12	Junjie Yang	HIRE-RoBERTa		88.3	68.6	97.1	93.0/90.7	92.4/92.0	74.3/90.2	90.7	90.4	95.5	87.9	89.0
13	Facebook AI	RoBERTa		88.1	67.8	96.7	92.3/89.8	92.2/91.9	74.3/90.2	90.8	90.2	95.4	88.2	89.0
14	Microsoft D365 AI & MSR AI	MT-DNN-ensemble		87.6	68.4	96.5	92.7/90.3	91.1/90.7	73.7/89.9	87.9	87.4	96.0	86.3	89.0
15	GLUE Human Baselines	GLUE Human Baselines		87.1	66.4	97.8	86.3/80.8	92.7/92.6	59.5/80.4	92.0	92.8	91.2	93.6	95.9

State of the art

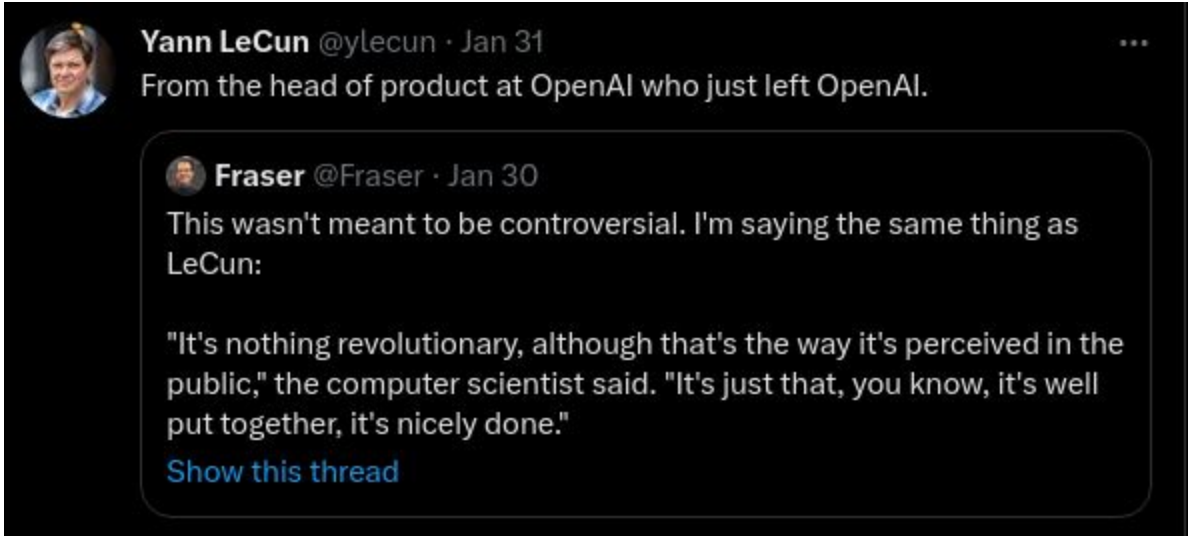
- On SuperGLUE

	Rank	Name	Model	URL	Score	BoolQ	CB	COPA	MultiRC	ReCoRD	RTE	WiC	WSC	AX-b	AX-g
+	1	Liam Fedus	ST-MoE-32B		91.2	92.4	96.9/98.0	99.2	89.6/65.8	95.1/94.4	93.5	77.7	96.6	72.3	96.1/94.1
	2	Microsoft Alexander v-team	Turing NLR v5		90.9	92.0	95.9/97.6	98.2	88.4/63.0	96.4/95.9	94.1	77.1	97.3	67.8	93.3/95.5
	3	ERNIE Team - Baidu	ERNIE 3.0		90.6	91.0	98.6/99.2	97.4	88.6/63.2	94.7/94.2	92.6	77.4	97.3	68.6	92.7/94.7
	4	Yi Tay	PaLM 540B		90.4	91.9	94.4/96.0	99.0	88.7/63.6	94.2/93.3	94.1	77.4	95.9	72.9	95.5/90.4
+	5	Zirui Wang	T5 + UDG, Single Model (Google Brain)		90.4	91.4	95.8/97.6	98.0	88.3/63.0	94.2/93.5	93.0	77.9	96.6	69.1	92.7/91.9
+	6	DeBERTa Team - Microsoft	DeBERTa / TuringNLRv4		90.3	90.4	95.7/97.6	98.4	88.2/63.7	94.5/94.1	93.2	77.5	95.9	66.7	93.3/93.8
	7	SuperGLUE Human Baselines	SuperGLUE Human Baselines		89.8	89.0	95.8/98.9	100.0	81.8/51.9	91.7/91.3	93.6	80.0	100.0	76.6	99.3/99.7
+	8	T5 Team - Google	T5		89.3	91.2	93.9/96.8	94.8	88.1/63.3	94.1/93.4	92.5	76.9	93.8	65.6	92.7/91.9
	9	SPoT Team - Google	Frozen T5 1.1 + SPoT		89.2	91.1	95.8/97.6	95.6	87.9/61.9	93.3/92.4	92.9	75.8	93.8	66.9	83.1/82.6
+	10	Huawei Noah's Ark Lab	NEZHA-Plus		86.7	87.8	94.4/96.0	93.6	84.6/55.1	90.1/89.6	89.1	74.6	93.2	58.0	87.1/74.4
+	11	Alibaba PAI&ICBU	PAI Albert		86.1	88.1	92.4/96.4	91.8	84.6/54.7	89.0/88.3	88.8	74.1	93.2	75.6	98.3/99.2
+	12	Infosys : DAWN : AI Research	RoBERTa-iCETS		86.0	88.5	93.2/95.2	91.2	86.4/58.2	89.9/89.3	89.9	72.9	89.0	61.8	88.8/81.5

ChatGPT


Step 1
Collect demonstration data and train a supervised policy.
A prompt is sampled from our prompt dataset.
Explain reinforcement learning to a 6 year old.
A labeler demonstrates the desired output behavior.
We give treats and punishments to teach...
SFT
This data is used to fine-tune GPT-3.5 with supervised learning.
Step 2
Collect comparison data and train a reward model.
A prompt and several model outputs are sampled.
Explain reinforcement learning to a 6 year old.
A
B
C
D
A labeler ranks the outputs from best to worst.
D > C > A > B
RM
This data is used to train our reward model.
D > C > A > B
Step 3
Optimize a policy against the reward model using the PPO reinforcement learning algorithm.
A new prompt is sampled from the dataset.
Write a story about otters.
PPO
The PPO model is initialized from the supervised policy.
The policy generates an output.
Once upon a time...
RM
The reward model calculates a reward for the output.
The reward is used to update the policy using PPO.
r_k

ChatGPT



Yann LeCun @ylecun · Jan 31

From the head of product at OpenAI who just left OpenAI.

Fraser @Fraser · Jan 30

This wasn't meant to be controversial. I'm saying the same thing as LeCun:

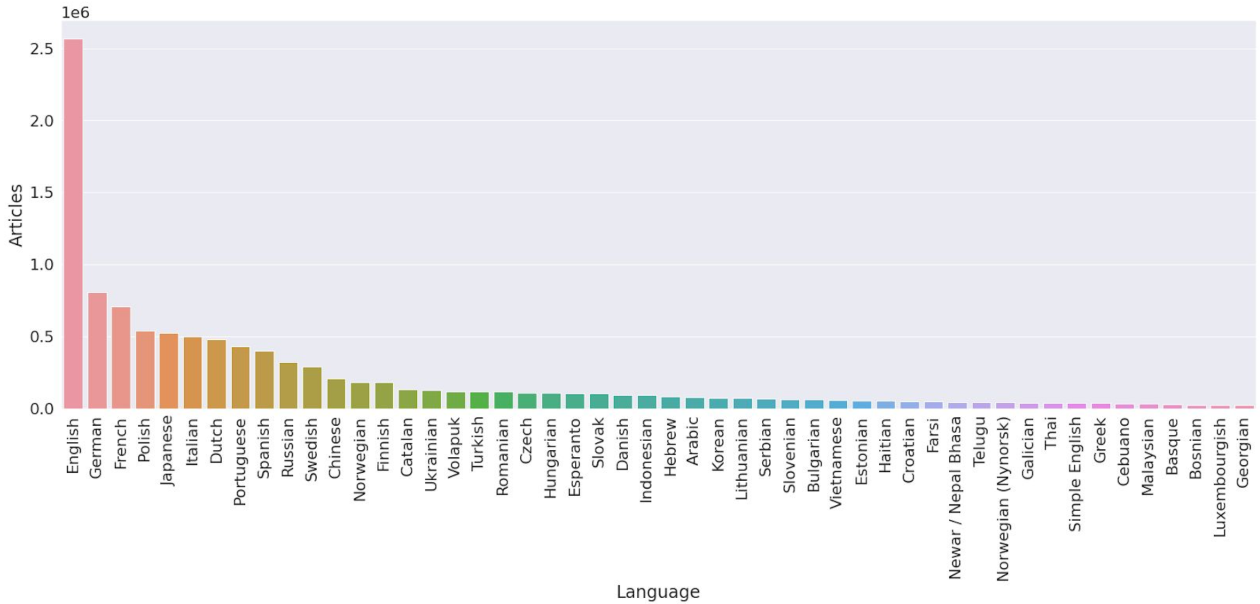
"It's nothing revolutionary, although that's the way it's perceived in the public," the computer scientist said. "It's just that, you know, it's well put together, it's nicely done."

Show this thread

NLP is not just English

Language diversity

- More than 7000 languages
- In 420 distinct families
- 136 are language isolate



English is a particular case

- Low morphology and strong syntax
- Conjugation
    - *Train*, *trains*, *trained*... add *will* or *going to* for the future and voilà
- Compared to Nepali

	Simple Present/Future	Probable Future	Simple Past	Past Habitual	Injunctive	Imperative
First person singular	गर्छु *garchu* 'I (will) do'	गरुँला *garũlā* 'I will (probably) do'	गरेँ *garẽ* 'I did'	गर्थेँ *garthẽ* 'I used to do'	गरूँ *garũ* 'may I do'	-
First person plural	गर्छौँ *garchaũ* 'We (will) do'	गरौँला *garaũlā* 'We will (probably) do'	गर्यौँ *garyaũ* 'We did'	गर्थ्यौँ *garthyaũ* 'We used to do'	गरौँ *garaũ* 'may we do, let's do'	-
Second person singular low-grade	गर्छस् *garchas* 'you (will) do'	गर्लास् *garlās* 'you will (probably) do'	गरिस् *garis* 'you did'	गर्थिस् *garthis* 'you used to do'	गरेस् *gares* 'may you do'	गर् *gar* 'do!'
Second person middle-grade/plural	गर्छौ *garchau* 'you (will) do'	गरौला *garaulā* 'you will (probably) do'	गर्यौ *garyau* 'you did'	गर्थ्यौ *garthyau* 'you used to do'	गरौ *garau* 'may you do'	गर *gara* 'do'
High grade	गर्नुहुन्छ *garnuhuncha* 'you (will) do'	गर्नुहोला *garnuhola* 'you will (probably) do'	गर्नुभयो *garnubhayo* 'you did'	गर्नुहुन्थ्यो *garnuhunthyo* 'you used to do'	गर्नुहोस् *garnuhos* 'may you do, please do'	-
Third person singular low-grade	गर्छ *garcha* 'he does'	गर्ला *garlā* 'he will (probably) do'	गर्यो *garyo* 'he did'	गर्थ्यो *garthyo* 'he used to do'	गरोस् *garos* 'may he do'	-
Third person middle-grade/plural masculine	गर्छन् *garchan* 'they (will) do'	गर्लान् *garlān* 'they will (probably) do'	गरे *gare* 'they did'	गर्थे *garthe* 'they used to do'	गरून् *garūn* 'may they do'	-
Third person middle-grade/plural feminine	गर्छिन् *garchin* 'she (will) do'	गर्लिन् *garlin* 'she will (probably) do'	गरिन् *garin* 'she did'	गर्थिन् *garthin* 'she used to do'	गरून् *garūn* 'may she do'	-

And you need to conjugate at
the negative form as well

English is a particular case

- Some languages have a very productive morphology
  - “*ugyarlaştıramayabileceklerimizdenmişsinizcesine*”
  - ugyar+laş+tır+ama+yabil+ecek+ler+imiz+den+miş+siniz+cesine
  - “(behaving) as if you were one of those whom we might not be able to civilize”
- Some languages don’t have space between words
  - Japonais, Mandarin, Thai, Khmer, Tibetan, Burman, …
- Some languages have a very liberal word order
  - “***Ngaragana-nguja*** *ngiy-a gujinganjanga-ni jiyawu* ***ngabulu***” (Wambaya)
  - (His) mother gave (him) milk with grog in it.
  - *Pupure**um** hospes videt leper**em*** (Latin)
  - The host sees a purple rabbit
- English doesn’t have a grammatical gender on words
  - French has 2, German and Russian 3, Ganda 17

Languages are constantly moving

- They are subject to change
- Every day new words are created
- Sounds change
- Meanings change
- Grammar changes

Languages are constantly moving

- 24% of known languages are extinct
- 2500 are endangered
- From the UNESCO
  - *At least 43% of the estimated 6000 languages spoken in the world are endangered. This figure doesn’t include “data deficient languages” for which no reliable information is available.*
- People speaking a dying language don’t teach (or can’t teach) their language to the next generation.
- 1 language disappear every 2 weeks.