Research Method & Presentation Course

Natural Language Processing

Reza Khan Mohammadi

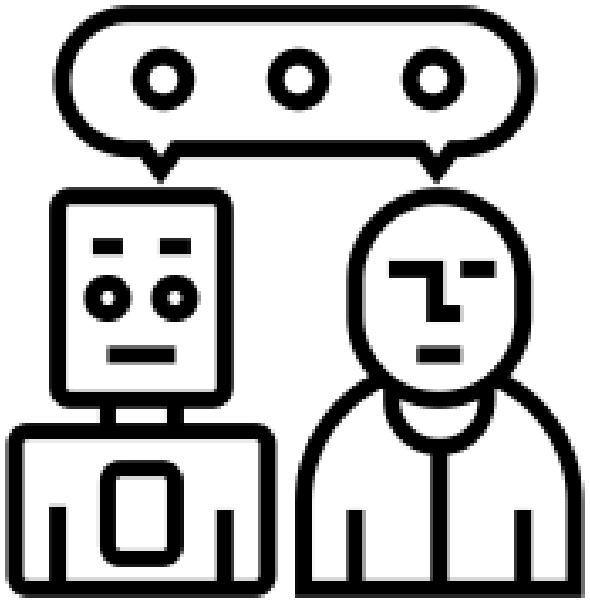
Some slides borrowed from Diyi Yang at GaTech

Welcome!

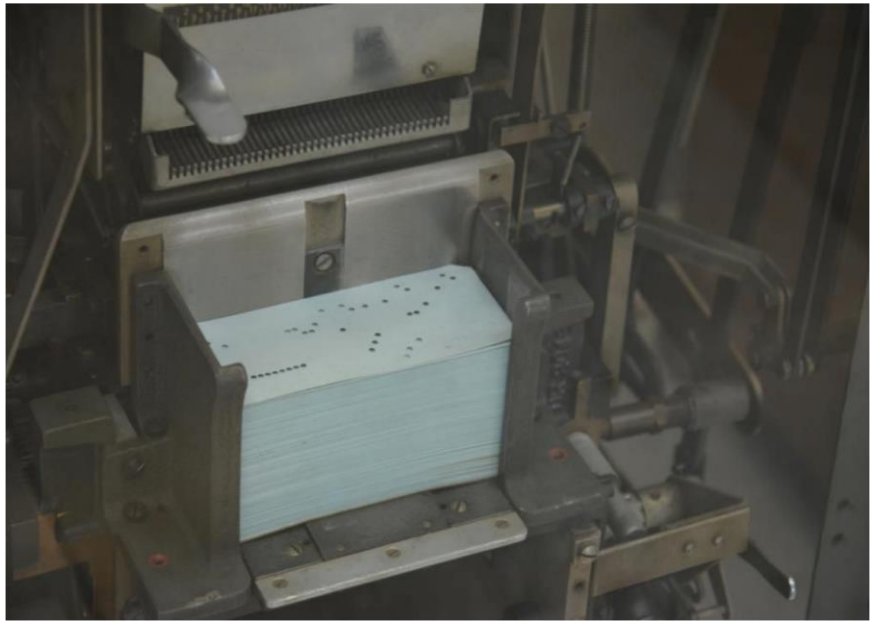
I am Reza Khan Mohammadi…

- NLP Researcher - Computer Engineering Student
- It’s been almost 2 years since I entered AI.
- My main focus of research is NLP.
- By now, I have submitted two research papers, one in an international Journal and the other in an international Conference.
- Currently working on 4 other papers.

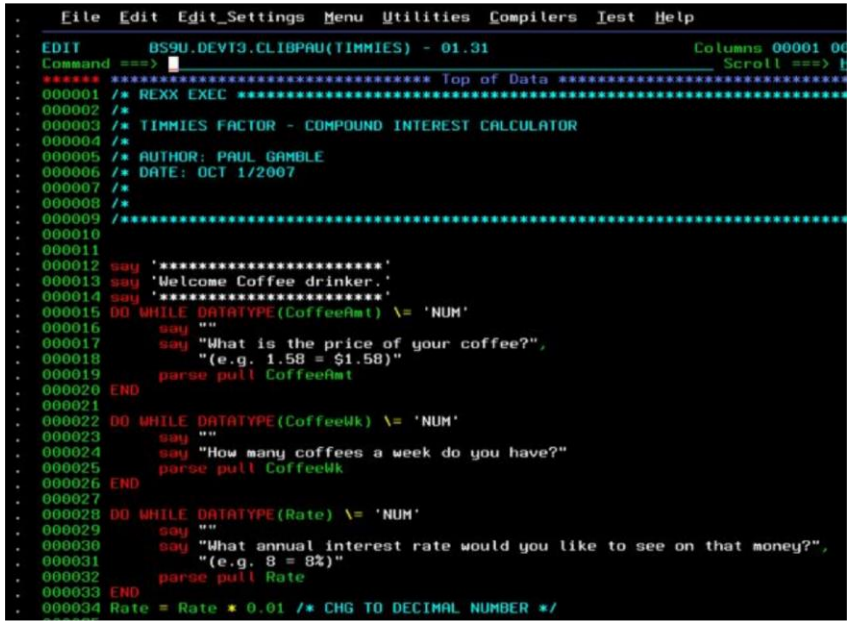
Today we will talk about **N**atural **L**anguage **P**rocessing (NLP)



Communication With Machines



~50-70s



~80s

today

Different uses of Language

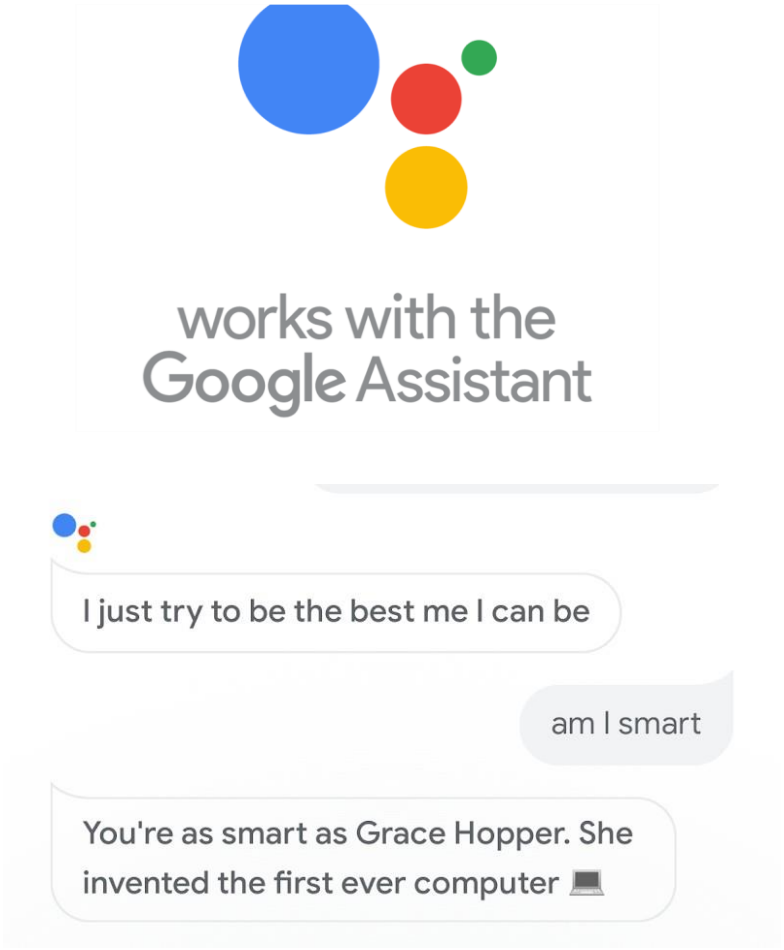
In our daily-life, we use language to...

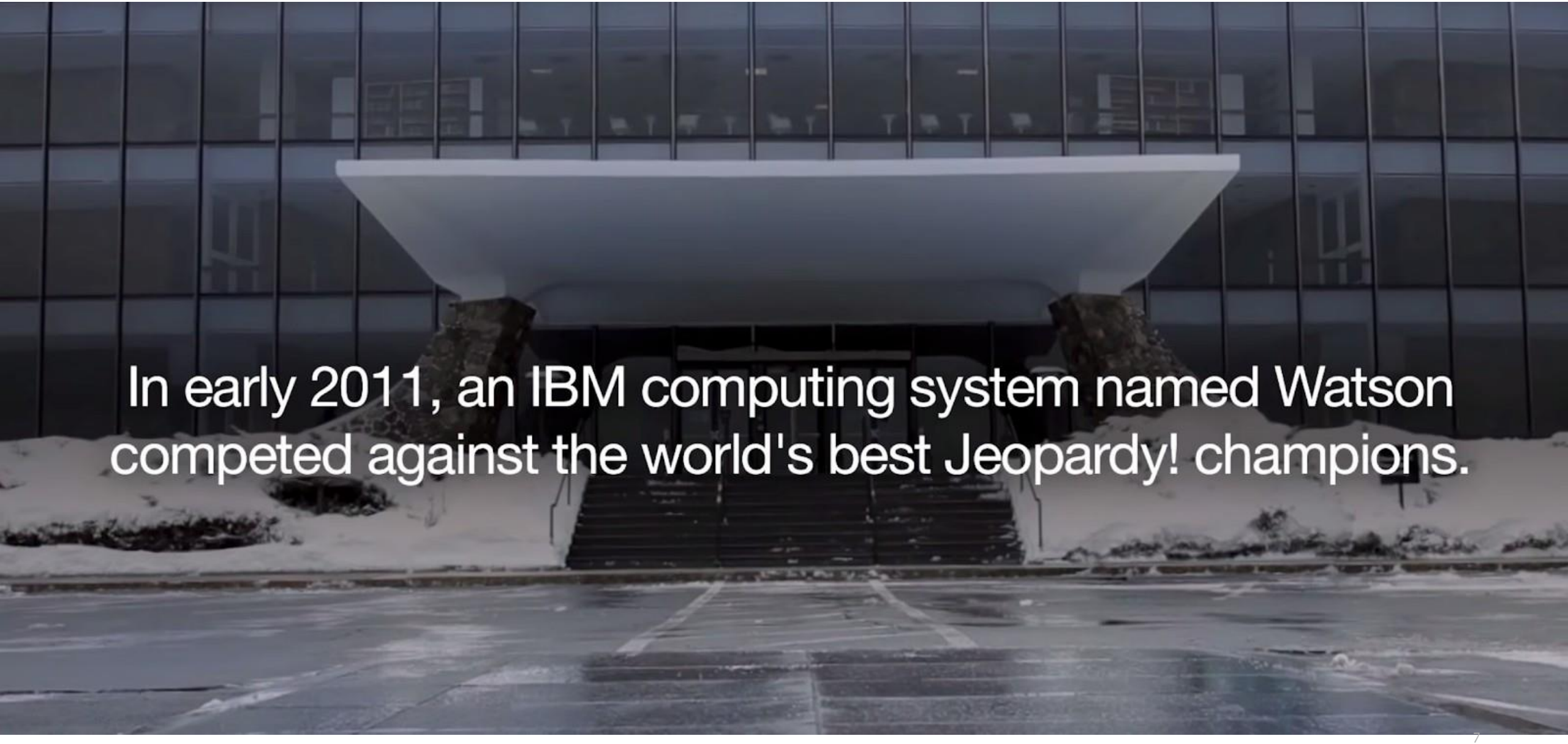
- write something.
- Chat.
- ask questions and answer.
- Describe something.
- ...

Conversational Agents

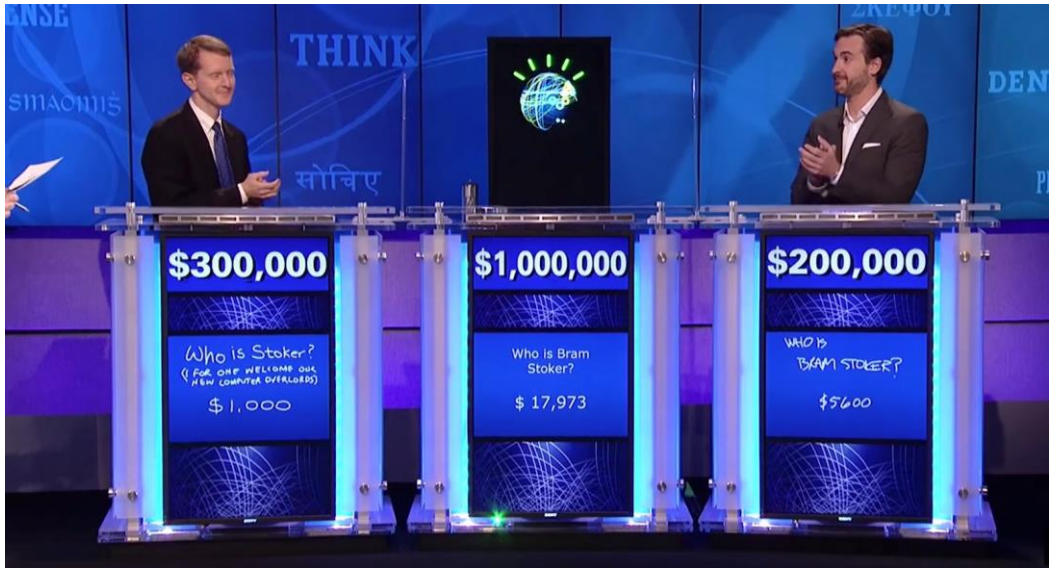
Conversational agents contain:

- Speech recognition
- Language analysis
- Dialogue processing
- Information retrieval
- Text to speech




In early 2011, an IBM computing system named Watson
competed against the world's best Jeopardy! champions.

Question Answering



- What does “divergent” mean?
- What year was Abraham Lincoln born?
- How many states were in the United States that year?
- How much Chinese silk was exported to England in the end of the 18th century?
- What do scientists think about the ethics of human cloning?

Textual Question Answering (Reading Comprehension)

The first recorded travels by Europeans to China and back date from this time. The most famous traveler of the period was the Venetian Marco Polo, whose account of his trip to "Cambaluc," the capital of the Great Khan, and of life there astounded the people of Europe. The account of his travels, Il milione (or, The Million, known in English as the Travels of Marco Polo), appeared about the year 1299. Some argue over the accuracy of Marco Polo's accounts due to the lack of mentioning the Great Wall of China, tea houses, which would have been a prominent sight since Europeans had yet to adopt a tea culture, as well the practice of foot binding by the women in capital of the Great Khan. Some suggest that Marco Polo acquired much of his knowledge through contact with Persian traders since many of the places he named were in Persian.

How did some suspect that Polo learned about China instead of by actually visiting it?

Answer: through contact with Persian traders

(Rajpurkar et al, 2016): SQuAD: 100,000+ Questions for Machine Comprehension of Text

Textual Question Answering

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

(Richardson et al, 2013): MCTest: A Challenge Dataset for the Open-Domain Machine Comprehension of Text

Conversational Question Answering

The Virginia governor's race, billed as the marquee battle of an otherwise anticlimactic 2013 election cycle, is shaping up to be a foregone conclusion. Democrat Terry McAuliffe, the longtime political fixer and moneyman, hasn't trailed in a poll since May. Barring a political miracle, Republican Ken Cuccinelli will be delivering a concession speech on Tuesday evening in Richmond. In recent ...

Q: What are the candidates **running** for?

A: Governor

Q: **Where?**

A: Virginia

Q: Who is the democratic candidate?

A: Terry McAuliffe

Q: Who is **his** opponent?

A: Ken Cuccinelli

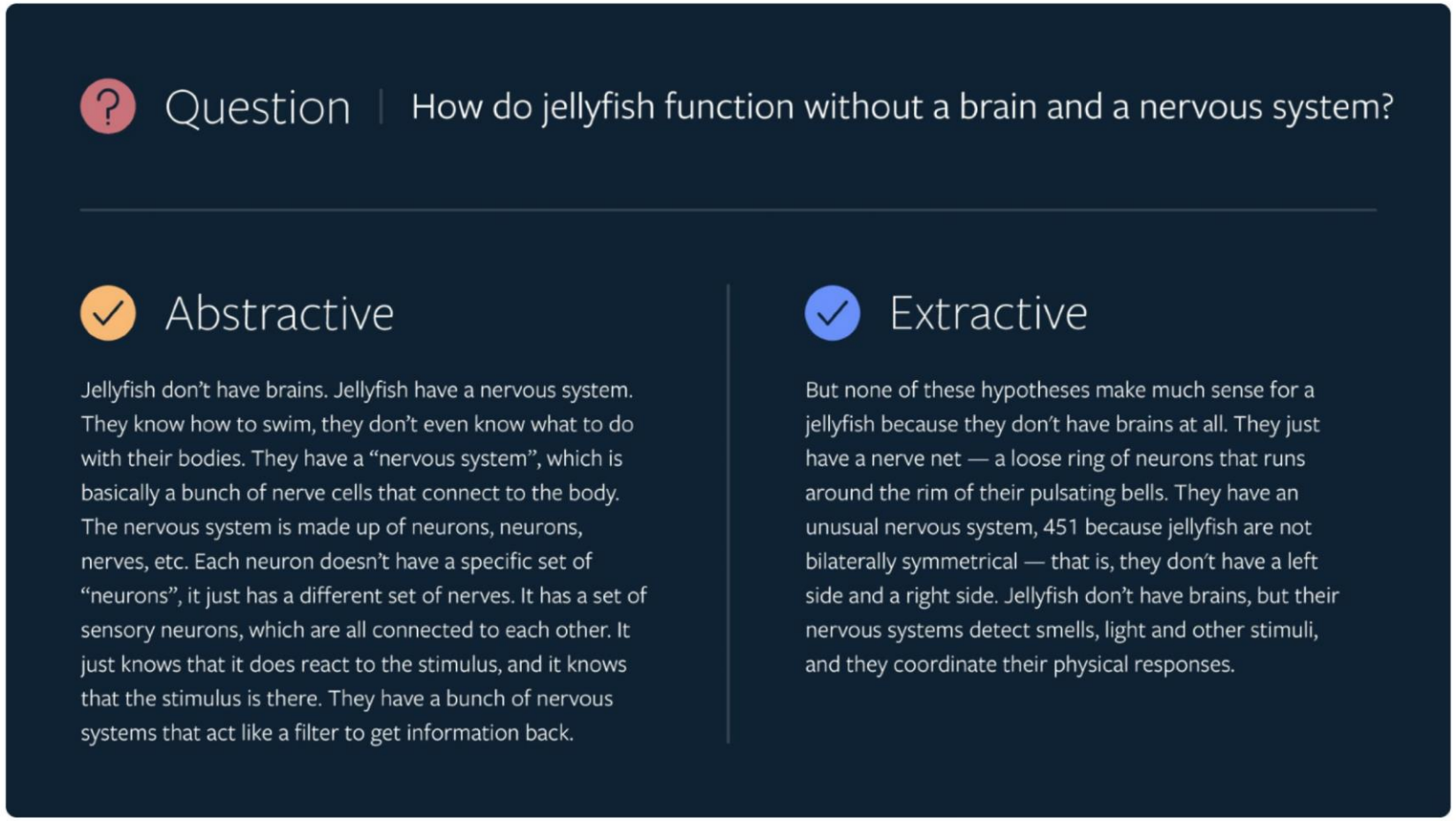
Q: What party does **he** belong to?

A: Republican

Q: Which of **them** is winning?

(Reddy et al, 2019): CoQA: A Conversational Question Answering Challenge

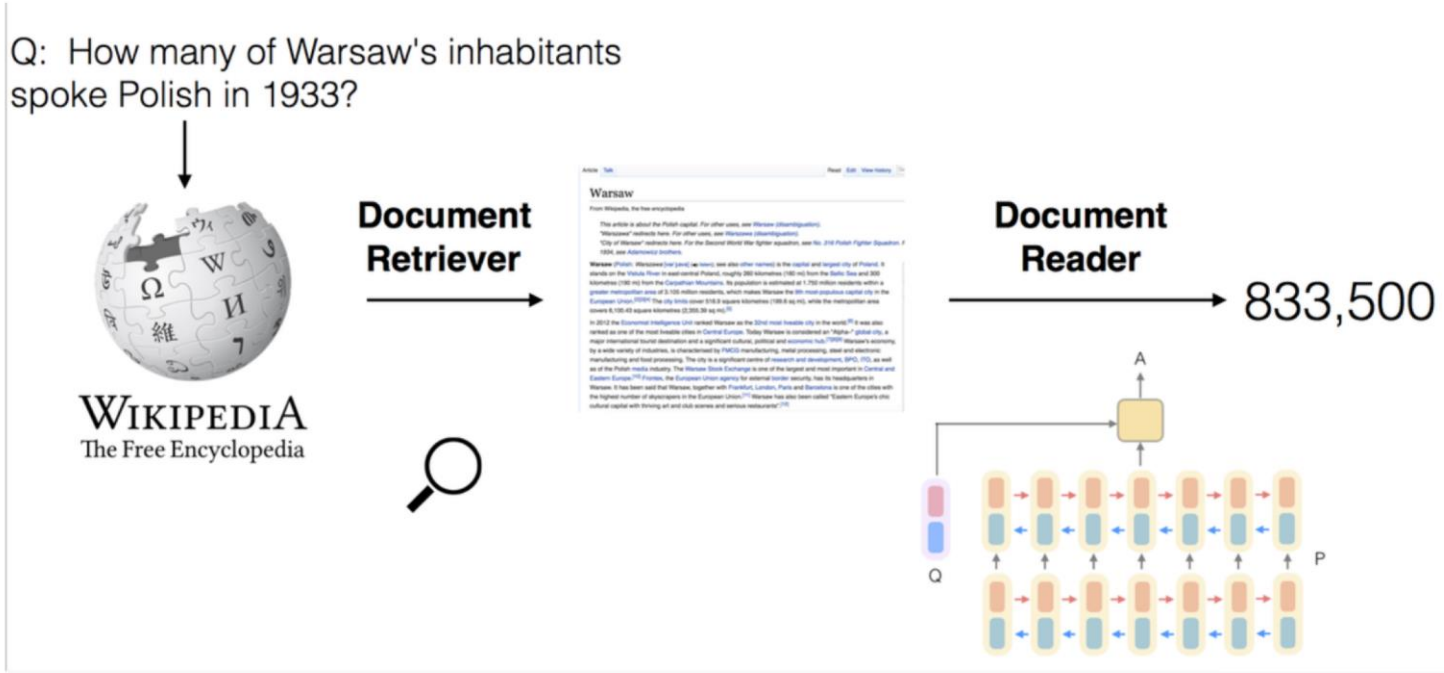
Long-form Question Answering



https://ai.facebook.com/blog/longform-qa/
(Fan et al, 2019): ELI5: Long Form Question Answering

Open-domain Question Answering

DrQA

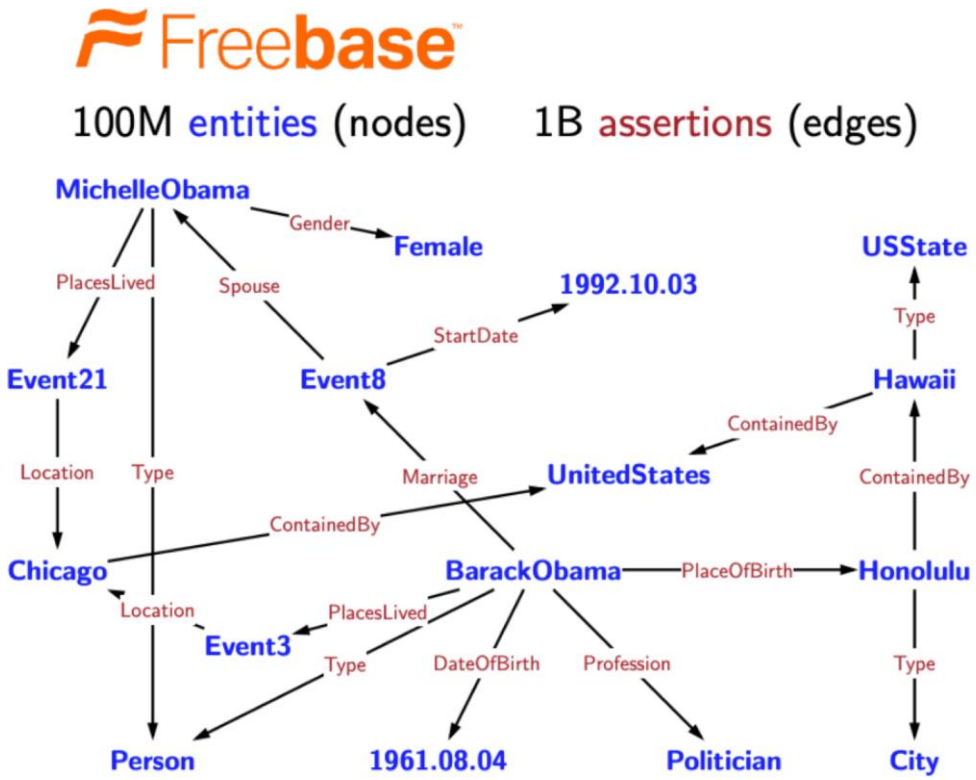



```
>>> process('What is the answer to life, the universe, and everything?')

Top Predictions:
+------+--------+-------------------------------------------------+--------------+-----------+
| Rank | Answer |                       Doc                       | Answer Score | Doc Score |
+------+--------+-------------------------------------------------+--------------+-----------+
|  1   |   42   | Phrases from The Hitchhiker's Guide to the Galaxy |    47242     |   141.26  |
+------+--------+-------------------------------------------------+--------------+-----------+
```

(Chen et al, 2017): Reading Wikipedia to Answer Open-Domain Questions

Knowledge Base Question Answering



Which states' capitals are also their largest cities by area?

semantic parsing

$\mu x.\text{Type.USState} \sqcap \text{Capital.argmax}(\text{Type.City} \sqcap \text{ContainedBy}.x, \text{Area})$

execute

Arizona,Hawaii,Idaho,Indiana,Iowa,Oklahoma,Utah

(Berant et al, 2013): Semantic Parsing on Freebase from Question-Answer Pairs

Table-based Question Answering

Year	City	Country	Nations
1896	Athens	Greece	14
1900	Paris	France	24
1904	St. Louis	USA	12
...	...	...	...
2004	Athens	Greece	201
2008	Beijing	China	204
2012	London	UK	204

x = Greece held its last Summer Olympics in which year?

y = 2004

(Pasupat and Liang, 2015): Compositional Semantic Parsing on Semi-Structured Tables.

Visual Question Answering

What color are her eyes?
What is the mustache made of?

How many slices of pizza are there?
Is this a vegetarian pizza?

(Antol et al, 2015): Visual Question Answering

Question Answering Datasets

- **Reading Comprehension**

  CNN/Daily Mail, CoQA, HotpotQA, QuAC, RACE, SQuAD, SWAG, Receipt QA, NarrativeQA, DROP, Story Cloze Test ……

- **Open-domain question answering**

  DuReader, Quasar, SearchQA, …

- **Knowledge base question answering**

Check out more datasets: http://nlpprogress.com/english/question_answering.html

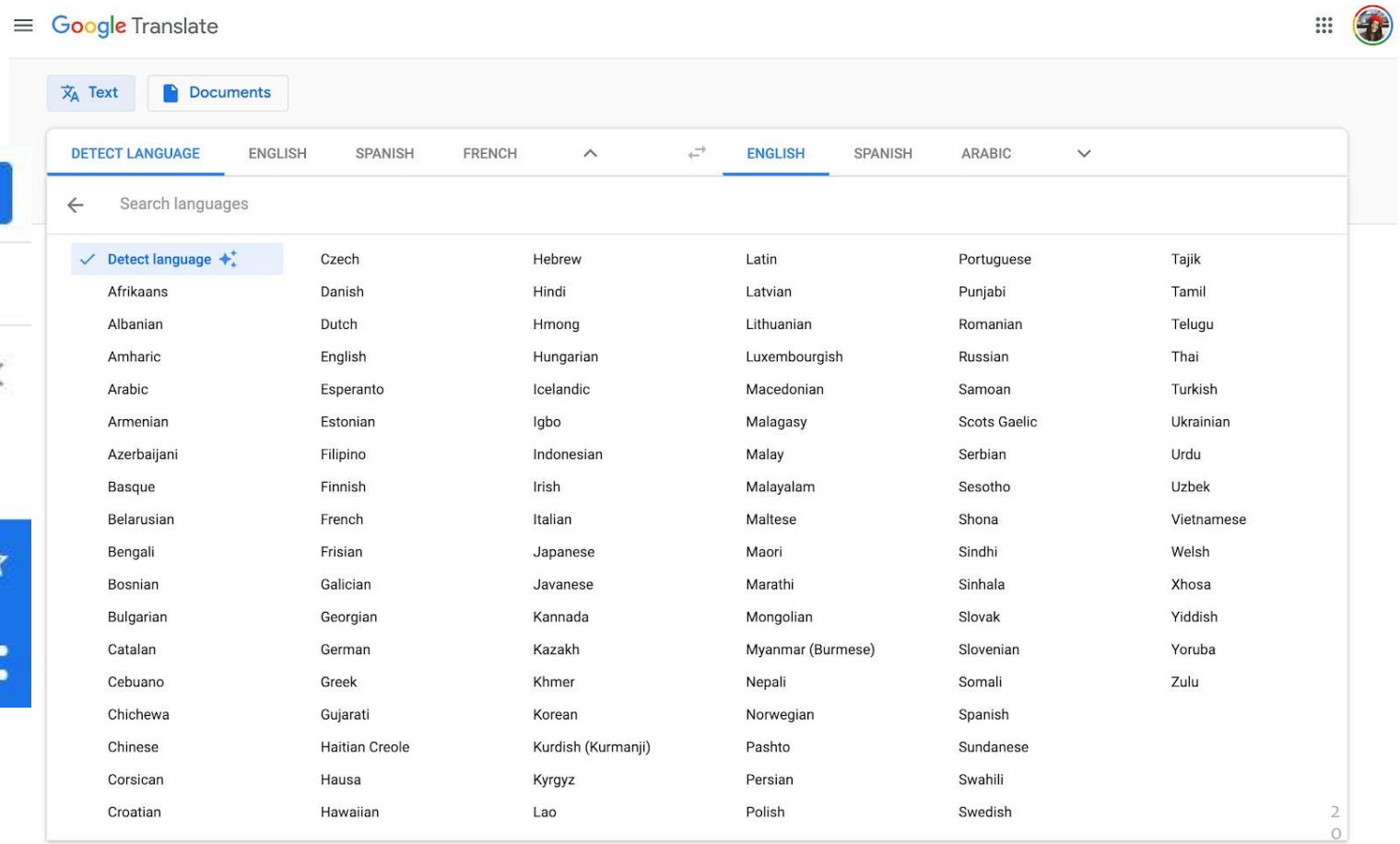
CNN Article

Document	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

SQuAD Benchmark

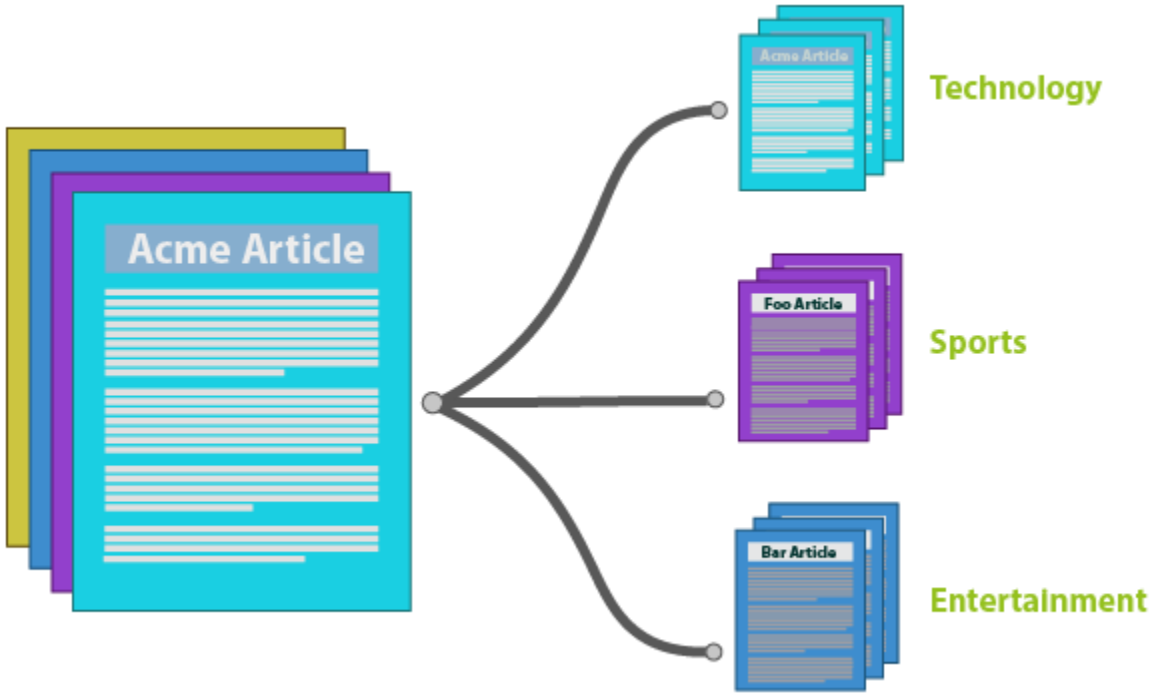
Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Apr 06, 2020	SA-Net on Albert (ensemble) *QIANXIN*	**90.724**	**93.011**
2 May 05, 2020	SA-Net-V2 (ensemble) *QIANXIN*	90.679	92.948
2 Apr 05, 2020	Retro-Reader (ensemble) *Shanghai Jiao Tong University*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

Machine Translation



Classification

- Text classification is the process of categorizing text into organized groups.



Movie Ratings

positive

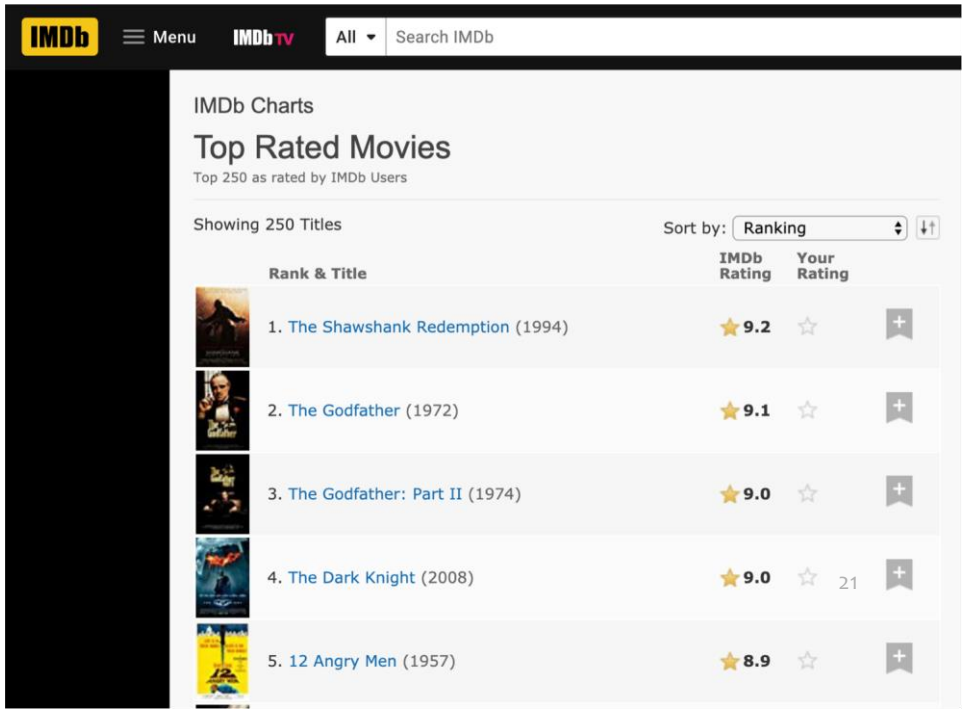
"… is a film which still causes real, not figurative, chills to run along my spine, and it is certainly the bravest and most ambitious fruit of Coppola's genius"

Roger Ebert, Apocalypse Now

- "I hated this movie. Hated hated hated hated hated this movie. Hated it. Hated every simpering stupid vacant audience-insulting moment of it. Hated the sensibility that thought anyone would like it."

Roger Ebert, North

negative



Female or Male Author?

1. By 1925 present-day Vietnam was divided into three parts under French colonial rule. The southern region embracing Saigon and the Mekong delta was the colony of Cochin-China; the central area with its imperial capital at Hue was the protectorate of Annam…
2. Clara never failed to be astonished by the extraordinary felicity of her own name. She found it hard to trust herself to the mercy of fate, which had managed over the years to convert her greatest shame into one of her greatest assets…

S. Argamon, M. Koppel, J. Fine, A. R. Shimoni, 2003. "Gender, Genre, and Writing Style in Formal Written Texts," Text, volume 23, number 3, pp. 321–346

Is This Spam?

Subject: **Important notice!**
From: Stanford University <newsforum@stanford.edu>
Date: October 28, 2011 12:34:16 PM PDT
To: undisclosed-recipients:;

Greats News!

You can now access the latest news by using the link below to login to Stanford University News Forum.

http://www.123contactform.com/contact-form-StanfordNew1-236335.html

Click on the above link to login for more information about this new exciting forum. You can also copy the above link to your browser bar and login for more information
about the new services.

Natural Language Processing

Applications

- Machine Translation
- Information Retrieval
- Question Answering
- Dialogue Systems
- Information Extraction
- Summarization
- Sentiment Analysis
- …

Core Technologies

- Language modeling
- Part-of-speech tagging
- Syntactic parsing
- Named-entity recognition
- Word sense disambiguation
- Semantic role labeling
- …

NLP lies at the intersection of Linguistics, Computer Science, and AI.

Ambiguity

- Ambiguity at multiple levels
  - Word senses: **bank** (finance or river ?)
  - Part of speech: **chair** (noun or verb ?)
  - Syntactic structure: **I can see a man with a telescope**

Ambiguity

A ship-shipping ship, shipping shipping-ships

Fields with Connections to NLP

- Machine learning
- Linguistics (including psycho-, socio-, descriptive, and theoretical)
- Cognitive science
- Information theory
- Logic
- Data science
- Political science
- Psychology
- Economics
- Education

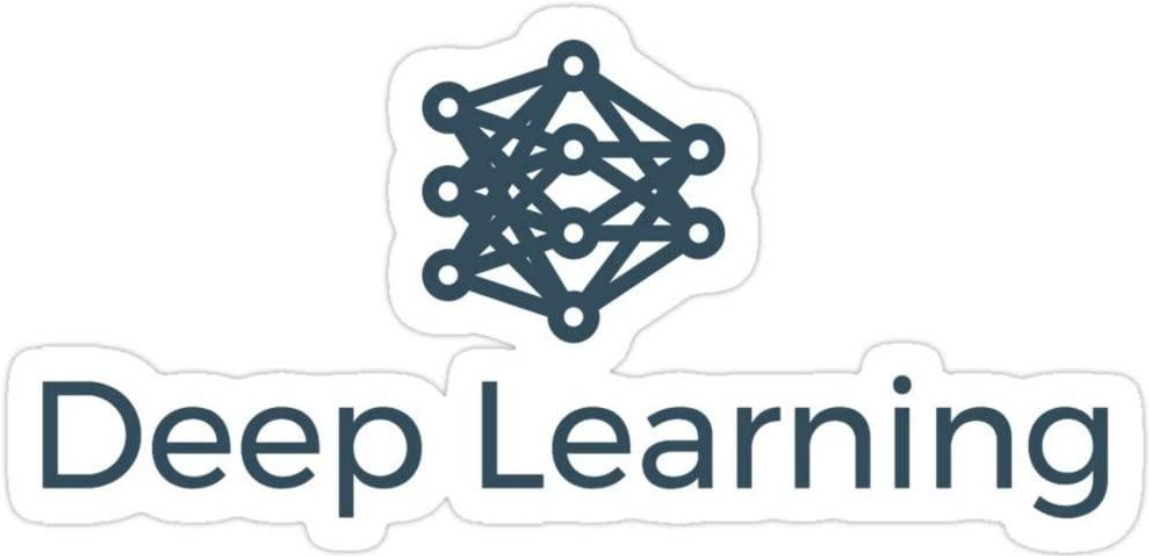
Today's Applications

- Conversational agents
- Information extraction and question answering
- Machine translation
- Opinion and sentiment analysis
- Social media analysis
- Visual understanding
- Essay evaluation
- Mining legal, medical, or scholarly literature

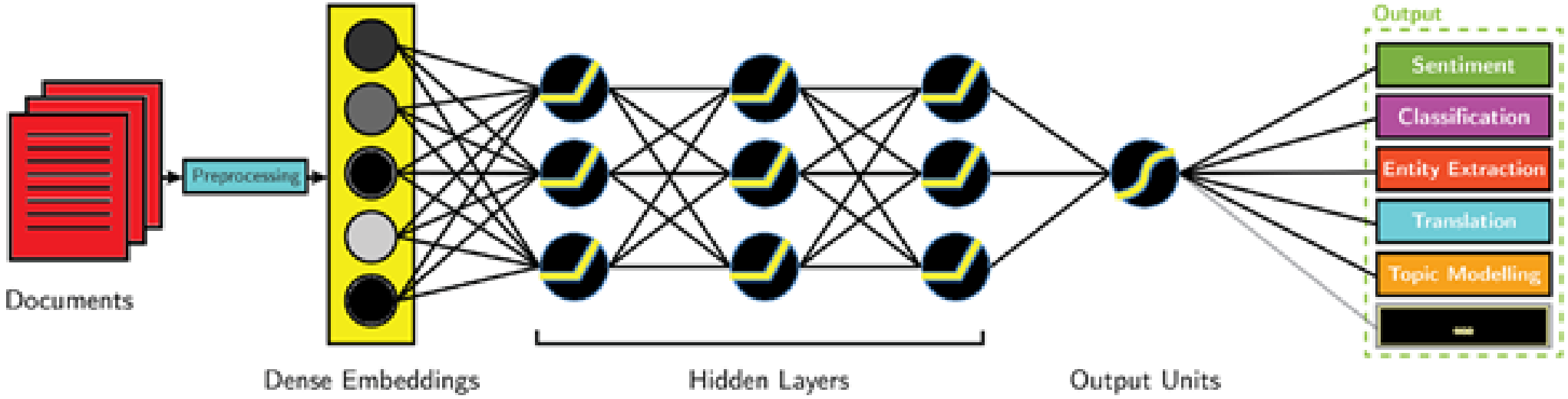
Factors Changing NLP Landscape

1. Increases in computing power
2. The rise of the web, then the social web
3. Advances in machine learning
4. Advances in understanding of language in social context

However, none of these were possible without…



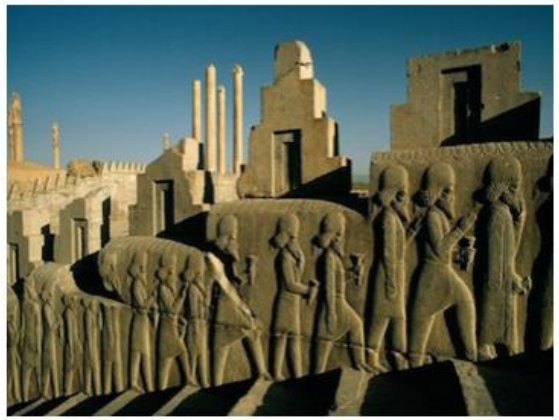
Deep Learning + NLP = Deep NLP



Personal Research Experience

Neural Style Transfer

content image



Ancient city of Persepolis

+

style image

The Starry Night (Van Gogh)

=

generated image

Persepolis
in Van Gogh style

Neural Style Transfer

Text Style Transfer

Task	Source	Target
UK to US	As the BMA's own study of alternative therapy showed, life is not as simple as that.	As the *F.D.A.'s* own study of alternative therapy showed, life is not as simple as that.
US to UK	The Greenburgh Drug and Alcohol Force and investigators in the Westchester District Attorney's Narcotics Initiative Program Participated in the arrest.	The *Royal Commission on Drug and Attache Force* and investigators in the Westchester District Attorney's Initiative Program Participated in the arrest.
NYT to Reddit	The votes weren't there.	*There weren't any upvotes.*
Reddit to NYT	i guess you need to refer to bnet website then.	*I* guess you need to refer to *the* bnet website then.
Pop to Hip Hop	My money's low	My money's *on the low*
Hip Hop to Pop	Yo, where the hell you been?	Yo, where the hell *are you?*

(Lee et al, 2019)

Gender Style Transfer

Example #5	GENDER (Male to Female)	GENDER (Female to Male)
SRC	this is a spot that ' s making very solid food , with good quality product .	this a great place for a special date or to take someone from out of town .
B-GST	**this is a cute spot that ' s making me very happy , with good quality product .**	**this a great place for a bachelor or to meet someone from out of town .**

(Sudhakar et al, 2019)

PGST: a Polyglot Gender Style Transfer method

https://arxiv.org/abs/2009.01040

Computer Science > Computation and Language

[Submitted on 2 Sep 2020]

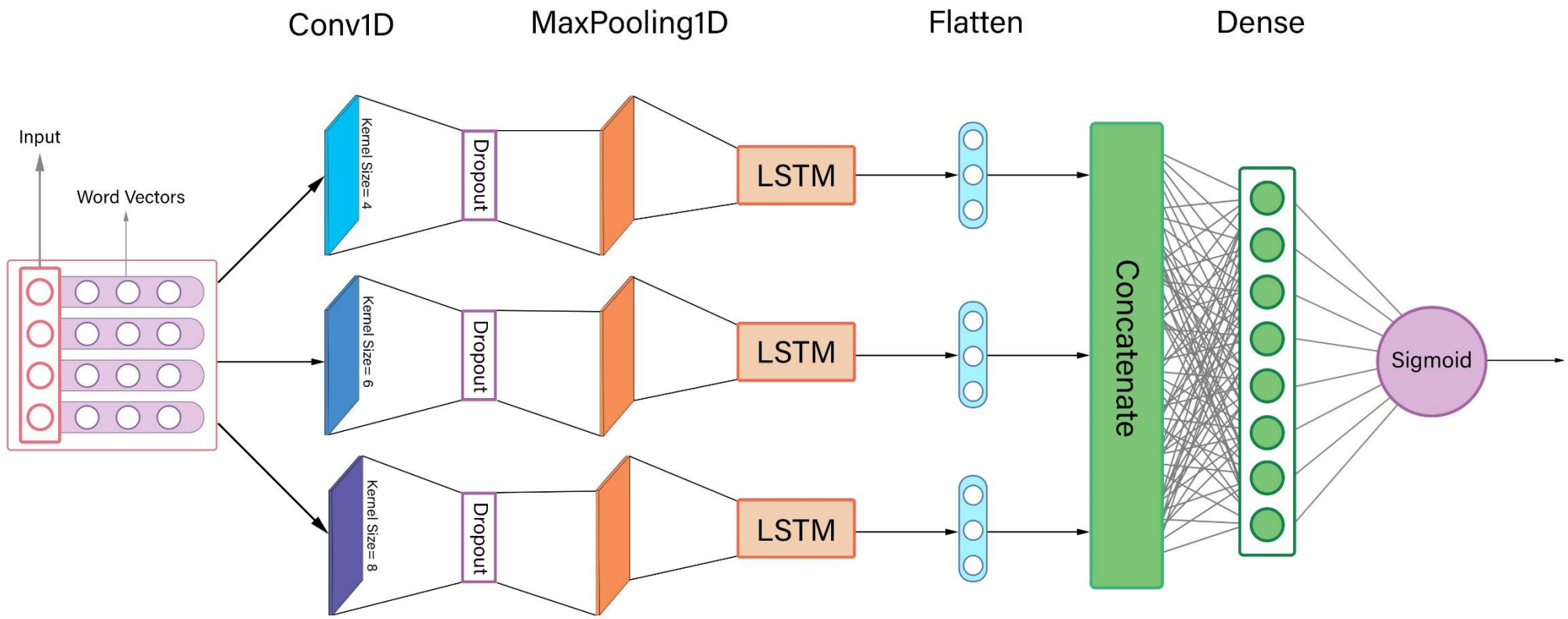
Defeating Author Gender Identification with Text Style Transfer

Reza Khan Mohammadi, Seyed Abolghasem Mirroshandel

Text Style Transfer can be named as one of the most important Natural Language Processing tasks. Up until now, there have been several approaches and methods experimented for this purpose. In this work, we introduce PGST, a novel polyglot text style transfer approach in gender domain composed of different building blocks. If they become fulfilled with required elements, our method can be applied in multiple languages. We have proceeded with a pre-trained word embedding for token replacement purposes, a character-based token classifier for gender exchange purposes, and the beam search algorithm for extracting the most fluent combination among all suggestions. Since different approaches are introduced in our research, we determine a trade-off value for evaluating different models' success in faking our gender identification model with transferred text. To demonstrate our method's multilingual applicability, we applied our method on both English and Persian corpora and finally ended up defeating our proposed gender identification model by 45.6% and 39.2%, respectively, and obtained highly competitive evaluation results in an analogy among English state of the art methods.

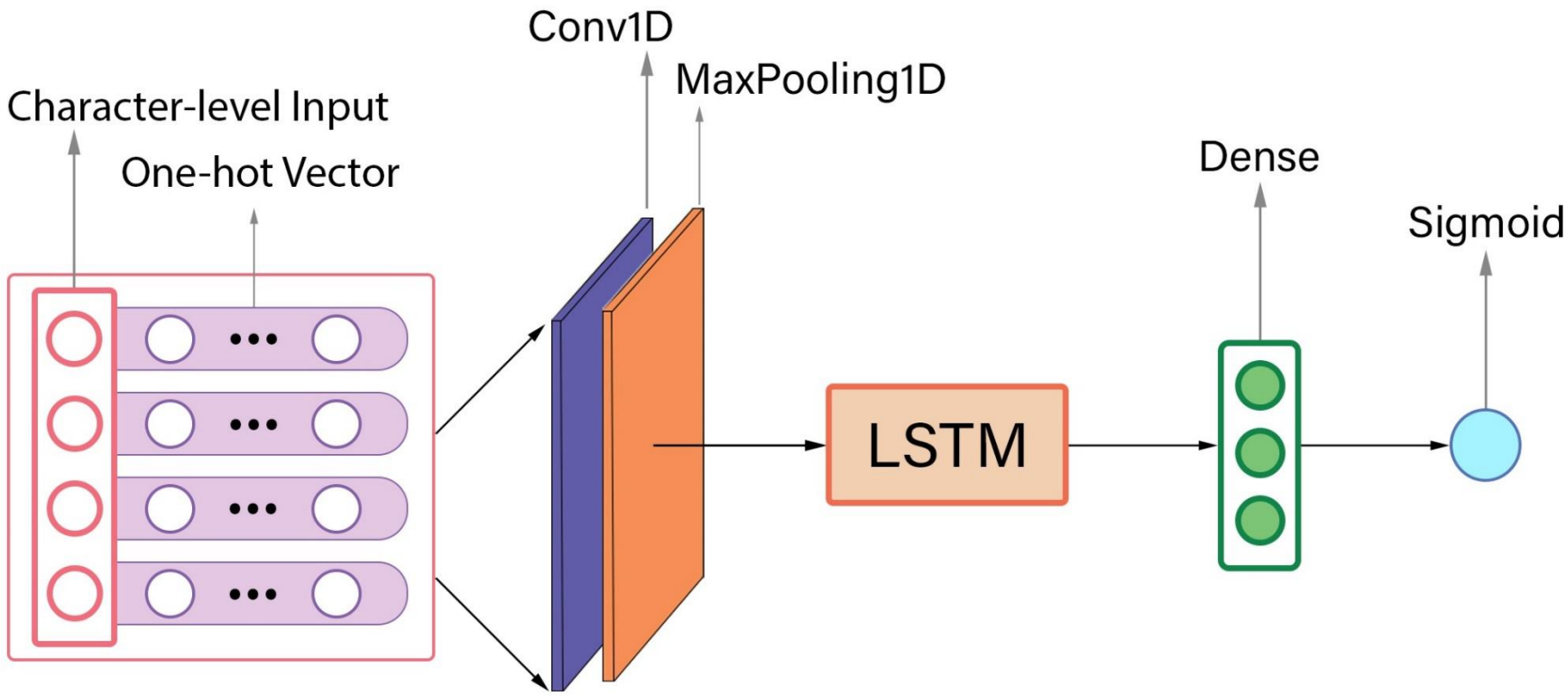
Comments: It is submitted to Computer and Speech Journal
Subjects: **Computation and Language (cs.CL)**; Machine Learning (cs.LG)
Cite as: arXiv:2009.01040 **[cs.CL]**
(or arXiv:2009.01040v1 **[cs.CL]** for this version)

Gender Identification



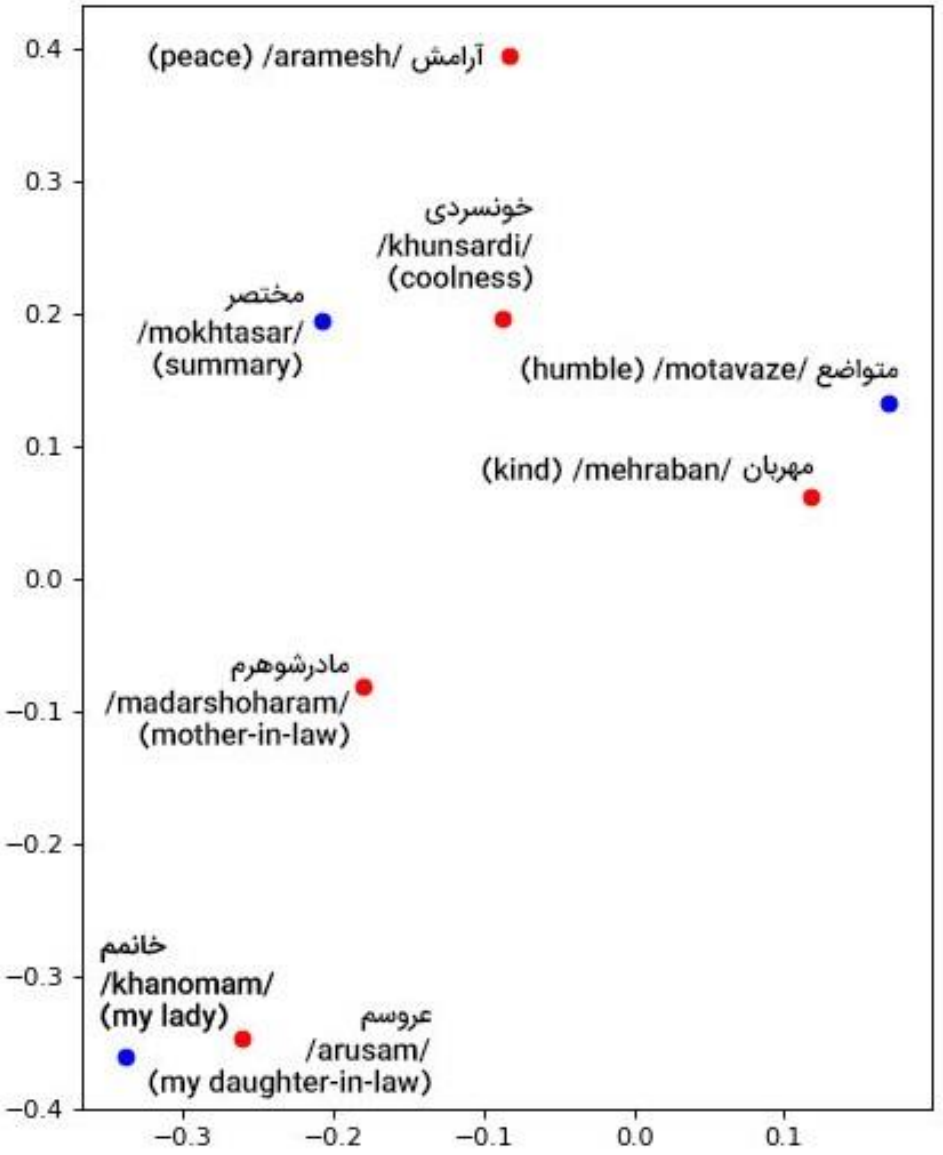
(Khan Mohammadi and Mirroshandel, 2020)

Character-based Token Classifier



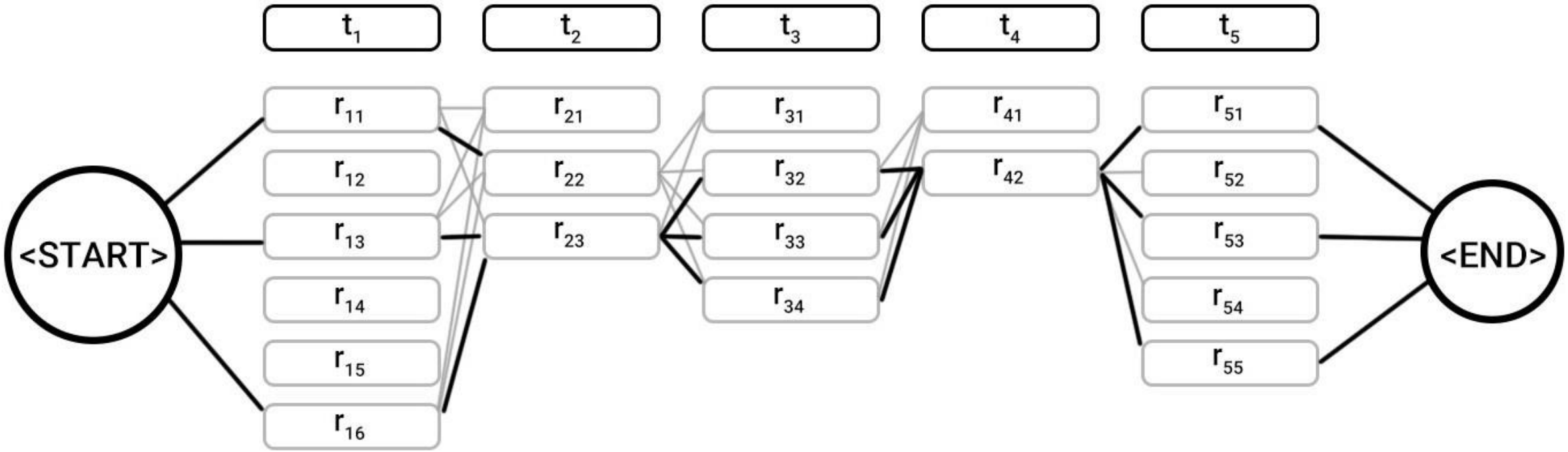
(Khan Mohammadi and Mirroshandel, 2020)

Character-based Token Classifier



(Khan Mohammadi and Mirroshandel, 2020)

Approach



(Khan Mohammadi and Mirroshandel, 2020)

Results

Model	Accuracy	
	Persian	English
Naïve Bayes	69%	73%
Logistic Regression	62%	70%
Multi-lingual BERT	65%	75%
SVM [15]	72%	-
CNN + LSTM NN	**90%**	**80%**

Model	BLEU	Perplexity	Accuracy
SRC	100	183.4	18.9
BT	46.0	196.2	52.9
G-GST	78.5	252.0	49.0
B-GST	**82.5**	**189.2**	**57.9**
PGST	68.4	198.9	45.6

(Khan Mohammadi and Mirroshandel, 2020)

Get in touch !

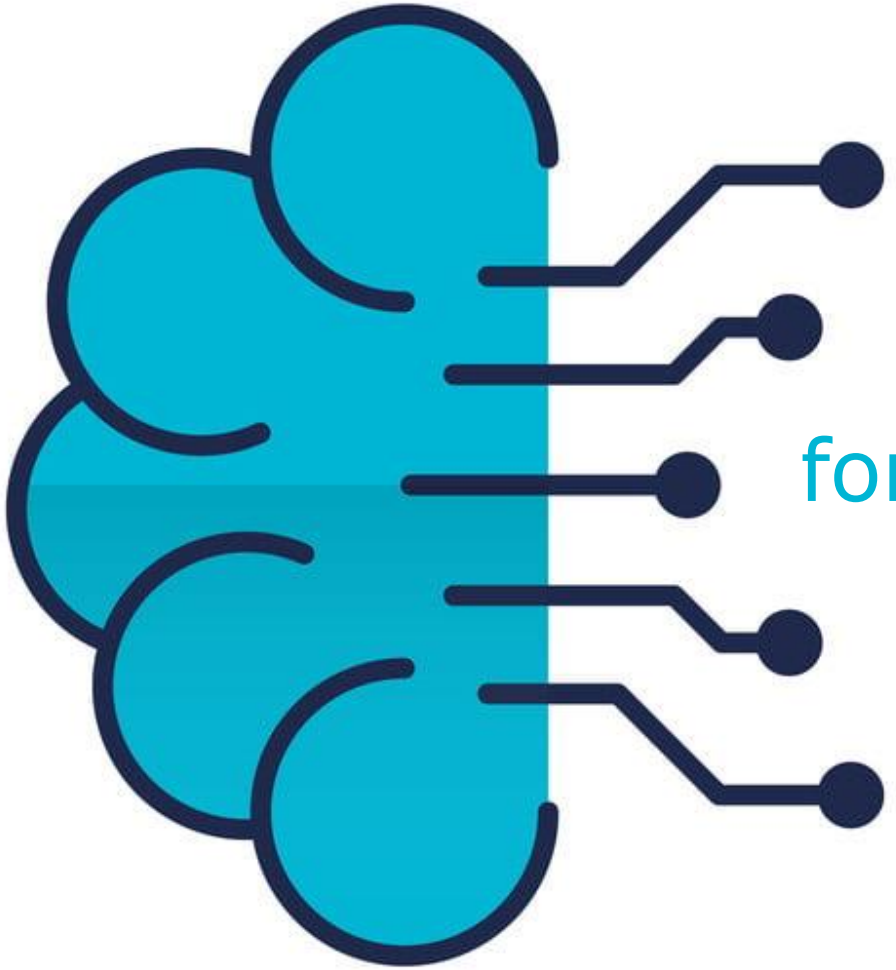
Scientific Association of the Computer Engineering Department

Brain and Cognition
Scientific Association

Contact

 @ledengary

 ledengary.github.io



Thank

you

for

your

attention !