LING/C SC/PSYC 438/538

Lecture 2

Sandiway Fong

Administrivia

- Do you have Perl and Python3 installed on your computer?
  - Please try to have it ready by next week!

- **Speech and Language Processing (3rd ed. draft Jan 12, 2022)**
  - **available at** http://web.stanford.edu/~jurafsky/slp3/
  - you can download the PDF

Today's Lecture

Chapter 1 of JM Reading:	• It's Question 1 on Homework 3
Introduction	• Homework 3 (due Sunday midnight)

Language and Computers

- Enormous amounts of data stored
  - world-wide web (WWW)
  - corporate databases
  - Dark Web
  - your own SSD or hard drive
- Major categories of data
  - numeric
  - **Language**: words, text, sound
  - pictures, video

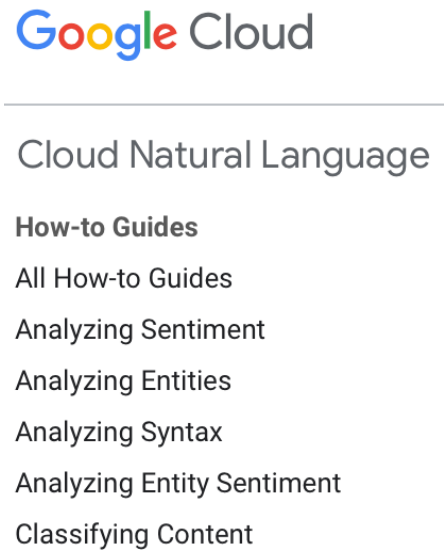
The size of the World Wide Web (The Internet) The Indexed Web contains **at least 4.28 billion pages** (Wednesday, 25 August, 2021).

https://www.worldwidewebsize.com

WorldWideWebSize.com | The size of the World Wide Web ...

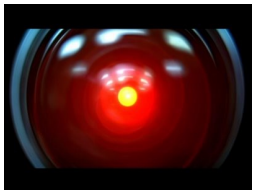
Language and Computers

- We know what we want from computer software
- “killer applications”
  - those that can make sense of language data
    - retrieve language data: (IR)
    - summarize knowledge contained in language data
    - sentiment analysis from online product reviews
    - answer questions (QA), make logical inferences
    - read medical reports, make diagnoses
    - translate from one language into another
    - recognize speech: transcribe audio from videos
    - etc...



Language and Computers

- We'd like computers to be smart about language
  - possess "intelligence"
  - pass the **Turing Test** …
  - but not be too smart?



From 2001… (HAL)

COMMENTARY ARTIFICIAL INTELLIGENCE

We Should Be as Scared of Artificial Intelligence as Elon Musk Is

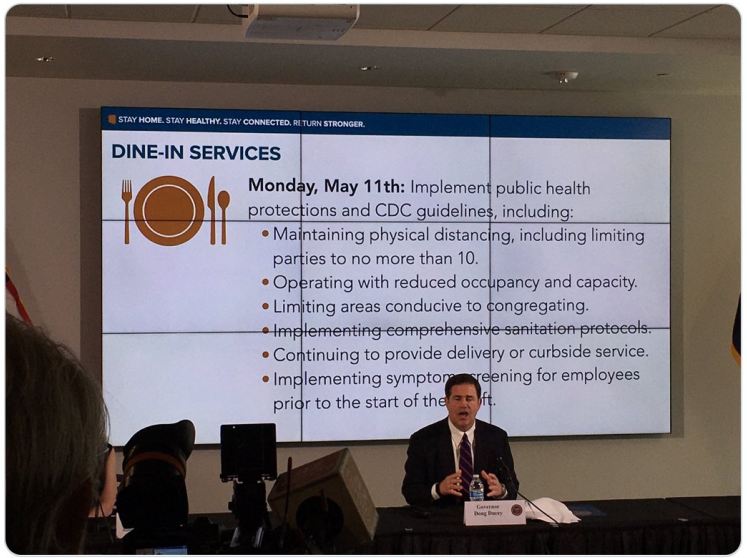
Steven Finlay
Aug 18, 2017

Elon Musk recently commented on Twitter (TWTR, +0.75%) that artificial intelligence (AI) is more dangerous than North Korea. It's not the first time that the entrepreneur has warned about the dangers of AI. Should we all be afraid as he is? Will AI lead to a huge disaster or robot takeover that destroys humanity?

http://fortune.com/2017/08/18/elon-musk-artificial-intelligence-risk/

Algorithms rule our lives

DINING-IN BACK MAY 11th:
Arizona restaurants will allow to have customers ordering in their buildings May 11th, with restrictions/guidelines.



3:24 PM · May 4, 2020

- Arizona Governor Doug Ducey announced plans to allow businesses to start reopening this week after consulting FEMA's new pandemic-prediction model, which has not been released to the public. The move came as a shock because a team of university experts -- who had developed their own model on the state's behalf -- had advised waiting until the end of May, lest the hospital system be overwhelmed.

https://mashable.com/article/how-algorithms-control-your-life

AI rule our lives

› Military News

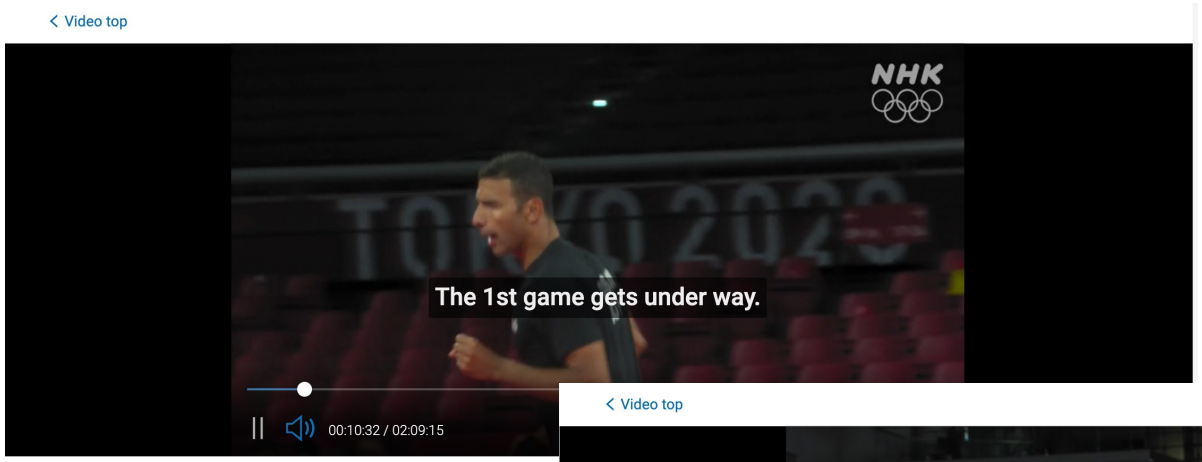
Rise of the Machines: AI Algorithm Beats F-16 Pilot in Dogfight

A U.S. Air Force F-15C and two F-16s from the 65th Aggressor Squadron fly in formation after a

- 8 AI systems: one winner
  - During the Defense Advanced Research Projects Agency-hosted AlphaDogfight challenge last week, Maryland-based Heron Systems came in first place among eight companies who pitted their AI-powered simulated aircraft against one another for two days straight.
- AI 5 – elite human pilot 0:
  - Then, Heron's system on the third day beat an F-16 Fighting Falcon pilot "in five straight simulated dogfights in the man-vs-machine finale," the organization said following the finals.

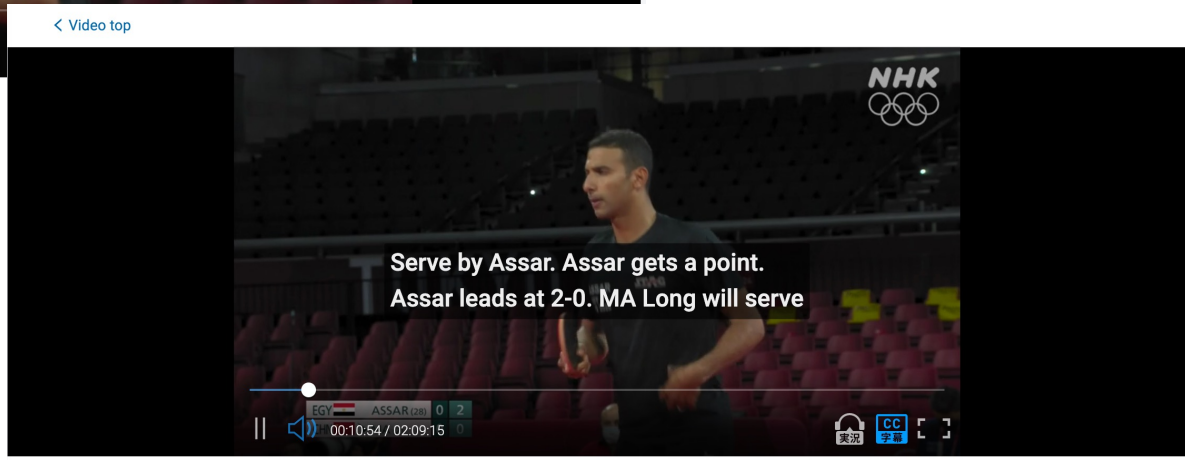
https://www.military.com/daily-news/2020/08/24/f-16-pilot-just-lost-algorithm-dogfight.html

Accessibility: robot subtitles (Tokyo Olympics)



The subtitles in this stream are brought by "robot play-by-play broadcast". It is automatically generated by computers using data provided by the organizer of the event. The contents may differ from the actual live voice coverage.

⇠ subtitle



Much harder task: true CC

Much harder task: true CC

Mathieu van der Poel

- vanderbogo
- round the ball
- vanderbolt
- vanderbald

Language and Computers

- (Un)fortunately, we're not quite there yet…
  - still gap between what computers can do and what we want them to be able to do

Often quoted (**but never verified**):

"The spirit is strong, but the flesh is weak" was translated into Russian and then back to English, the result was "The vodka is good, but the meat is rotten."

but with Google translate ~~or babelfish~~, it's not difficult to find (funny) examples…

Language and Computers

- and how can we tell if the translation is right anyway?

http://fun.drno.de/pics/english/only-in-china/TranslateServerError.jpg

Applications

- *technology is still in development*
- Deep Learning is accelerating the deployment of NLP applications

- even if we are willing to pay...
  - machine translation has been worked on since after World War II (1950s)
  - still not perfected today
  - **why?**
  - what are the properties of human languages that make it hard?

Language and Computers

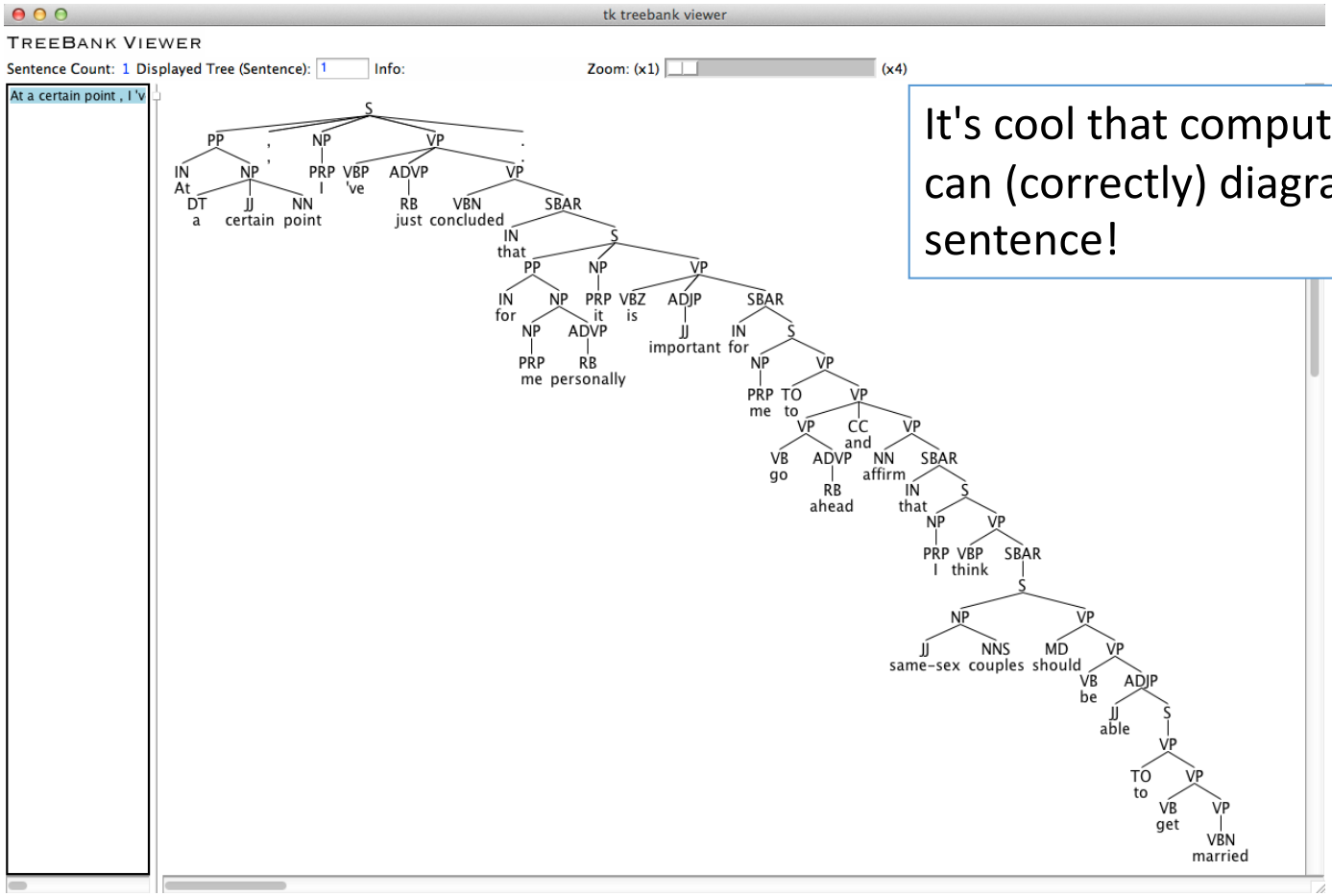
We can exploit the recursive nature of language …

Language and Computers

- Obama: "*At a certain point, I've just concluded that for me personally it is important for me to go ahead and affirm that I think same-sex couples should be able to get married*."

Is this sentence complicated? Why?

Language and Computers



It's cool that computer programs can (correctly) diagram this long sentence!

Language and Computers

Executive Summarization

Language and Computers

- Obama: "~~*At a certain point, I've just concluded that for me personally it is important for me to go ahead and affirm that*~~ ***I think same-sex couples should be able to get married***."

Most summarizer programs can't do this …

Language and Computers

A Sports Shooter Shoots Shooters Shooting Sports

Jul 12, 2015 · Michael Zhang Share Like 1.2k 15 Comments

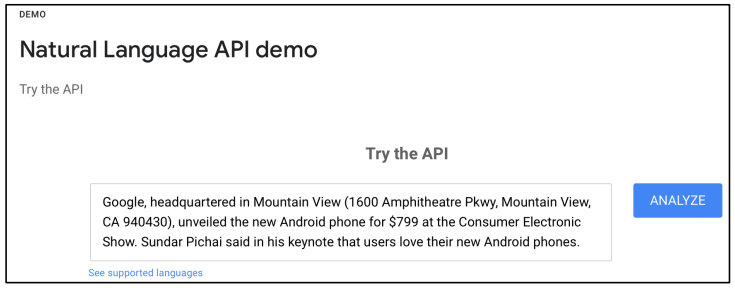
Language and Computers

Natural language parsers

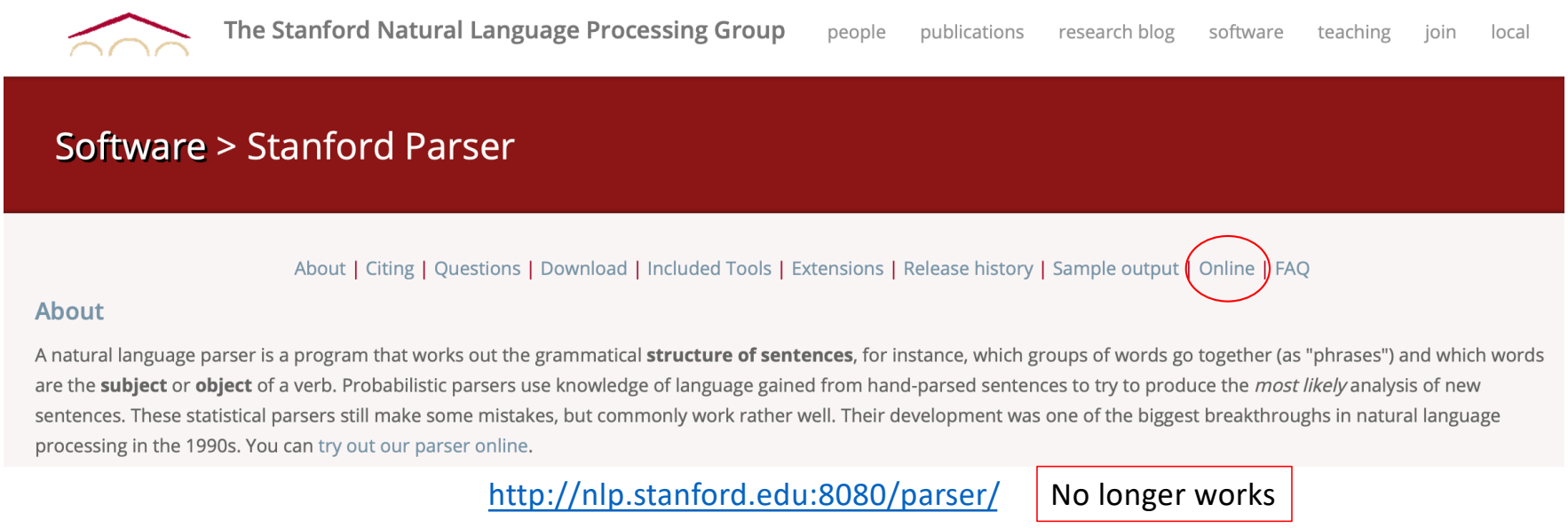
- Let's get some idea of what these (popular) systems produce.

  Sadly, all once had an easily accessible working demo page:

  1. Stanford Parser (deprecated? Try https://corenlp.run)
  2. Berkeley Parser (deprecated? B. Neural Parser: https://parser.kitaev.io)
  3. Google Natural Language (https://cloud.google.com/natural-language)



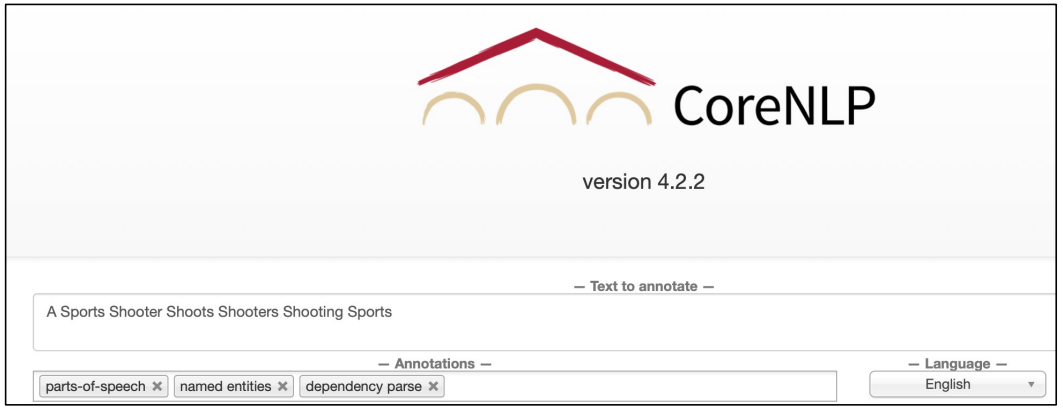
Language and Computers

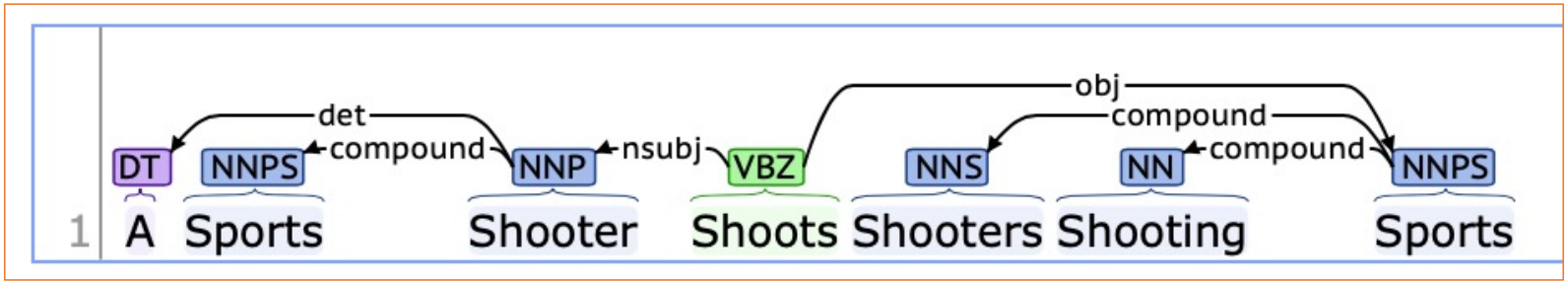


http://nlp.stanford.edu:8080/parser/ No longer works

Language and Computers

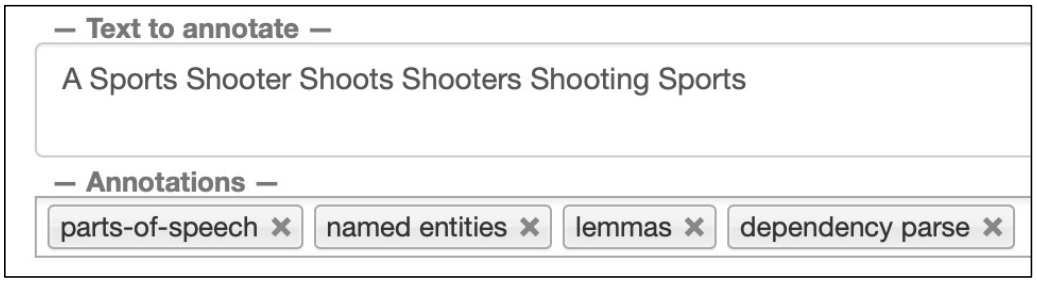
- Natural language parsers
  - Stanford CoreNLP
  - Java-based
  - Demo!
    (https://corenlp.run)

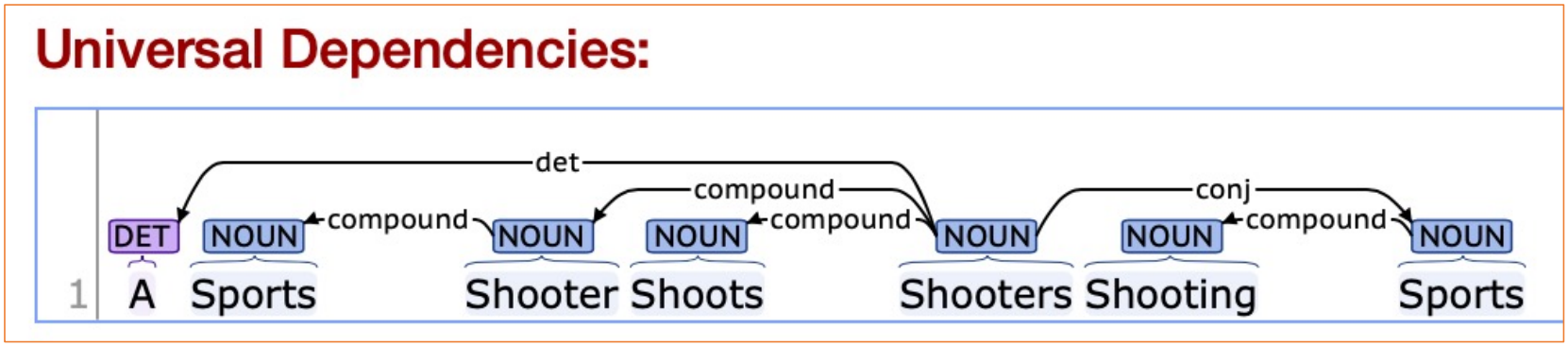




Language and Computers

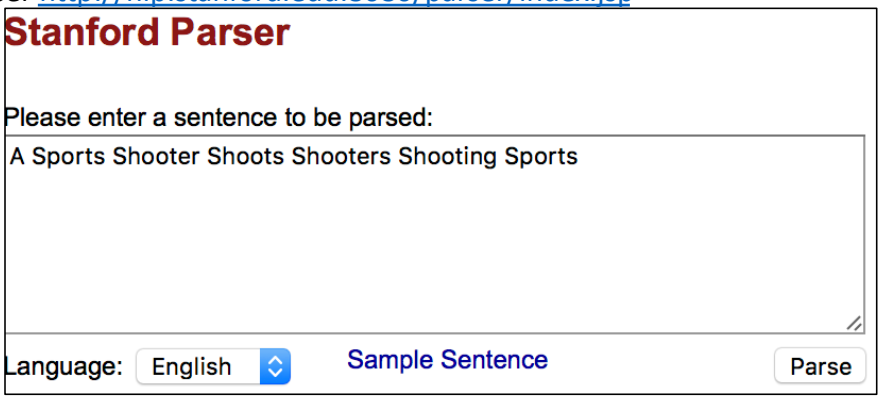
- Natural language parsers
  - (Stanford) Stanza
  - Python-based (CoreNLP)
  - Demo!
    http://stanza.run





Language and Computers

- seems to be offline right now
- Stanford parser http://nlp.stanford.edu:8080/parser/index.jsp



- Part of Speech Tagging:

Tagging

A/DT Sports/NNPS Shooter/NNP Shoots/NNP Shooters/NNP Shooting/NNP Sports/NNP

DT = determiner; NNP = Proper Noun; NNPS = Plural Proper Noun;
VBZ = Verb 3rd Person Singular Present; VBG = Verb Gerund Form

Language and Computers

- Syntax (Constituency-based):

Parse

```
(ROOT
  (FRAG
    (NP (DT A) (NNPS Sports))
    (NP (NNP Shooter) (NNP Shoots) (NNP Shooters) (NNP Shooting) (NNP Sports))))
```

Constituents:
FRAG = Fragment (of a sentence)
S = Sentence, NP = Noun Phrase, VP = Verb Phrase
Parts of Speech:
DT Determiner, NNP Proper Noun, NNPS Plural Proper Noun.

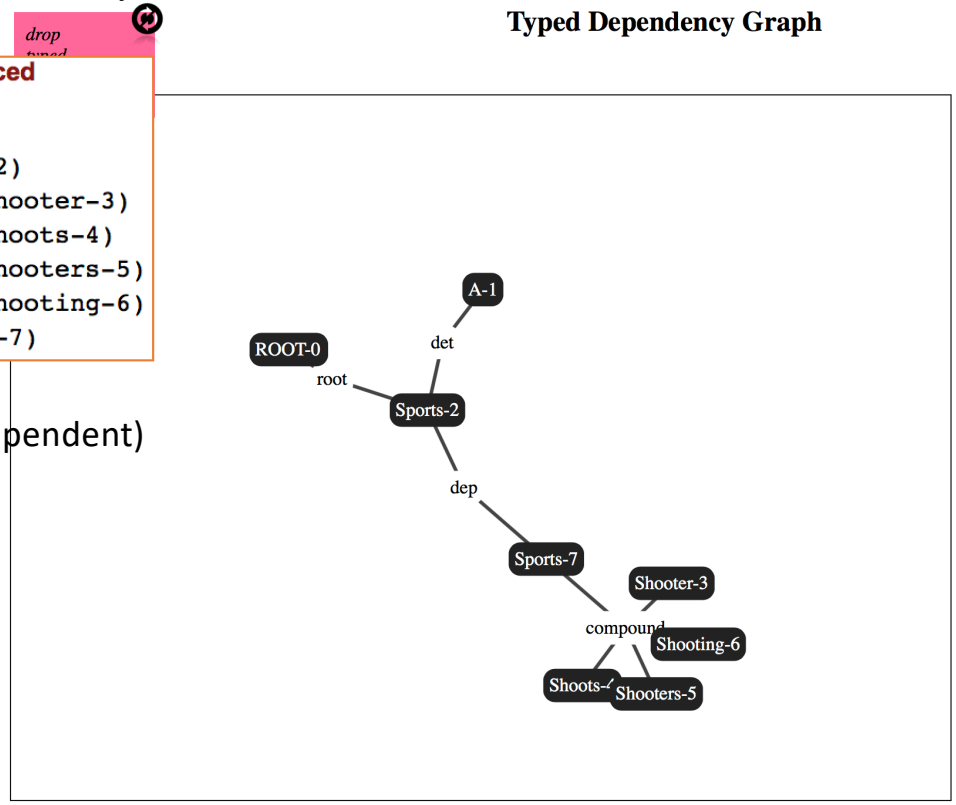
Language and Computers

- Syntax (Dependencies):

Universal dependencies, enhanced

```
det(Sports-2, A-1)
root(ROOT-0, Sports-2)
compound(Sports-7, Shooter-3)
compound(Sports-7, Shoots-4)
compound(Sports-7, Shooters-5)
compound(Sports-7, Shooting-6)
dep(Sports-2, Sports-7)
```

RELATION(depends-on, dependent)



Language and Computers

Your query

A sports shooter shoots shooters shooting sports

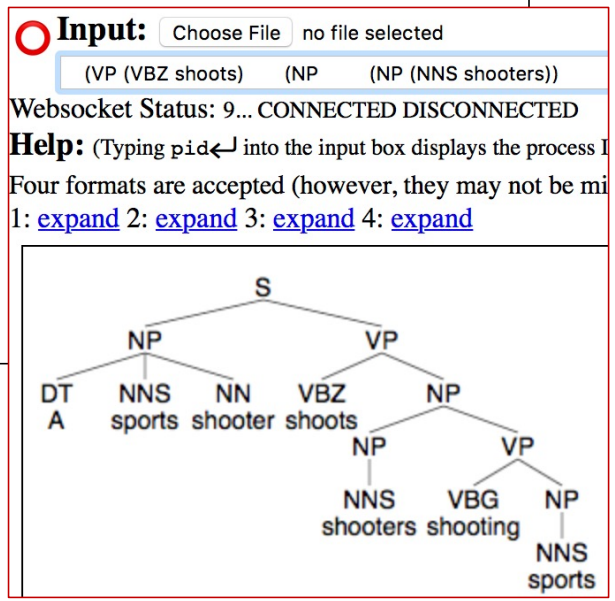
Tagging

```
A/DT  sports/NNS  shooter/NN  shoots/VBZ  shooters/NNS  shooting/VBG  sports/NNS
```

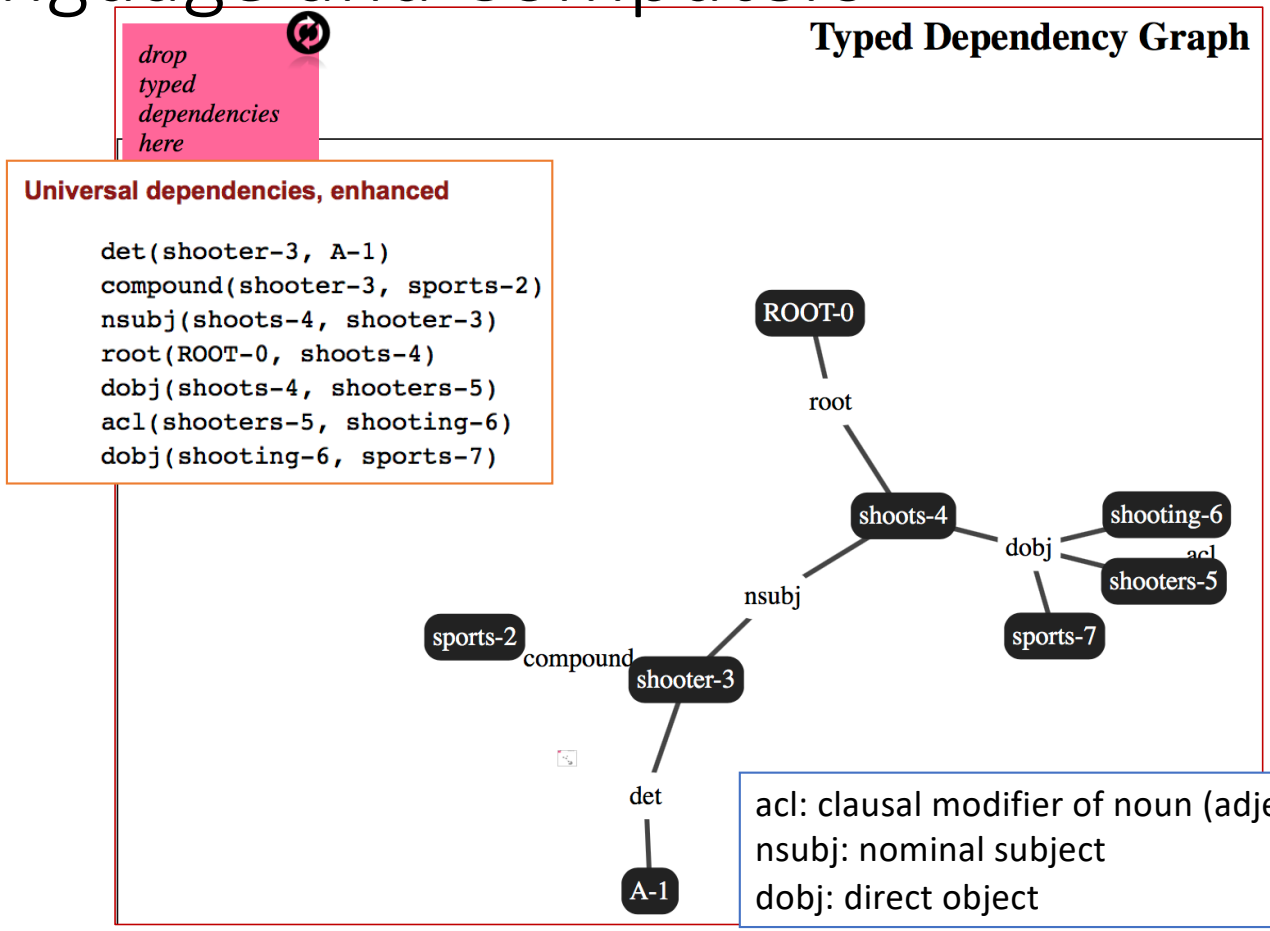
Parse

```
(ROOT
  (S
    (NP (DT A) (NNS sports) (NN shooter))
    (VP (VBZ shoots)
      (NP
        (NP (NNS shooters))
        (VP (VBG shooting)
          (NP (NNS sports)))))))
```

FRAG = Fragment (of a sentence)
S = Sentence
NP = Noun Phrase
VP = Verb Phrase



Language and Computers

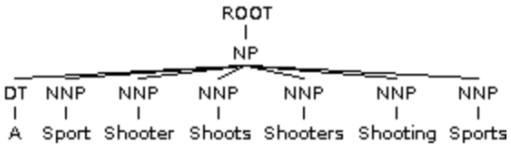


acl: clausal modifier of noun (adjectival clause)
nsubj: nominal subject
dobj: direct object

Language and Computers

Berkeley Parser (no longer available online)
http://tomato.banatao.berkeley.edu:8080/parser/parser.html

```
(ROOT
  (NP (DT A) (NNP Sport) (NNP Shooter) (NNP Shoots) (NNP Shooters) (NNP Shooting) (NNP Sports)))
```



A Sport Shooter Shoots Shooters Shooting Sports

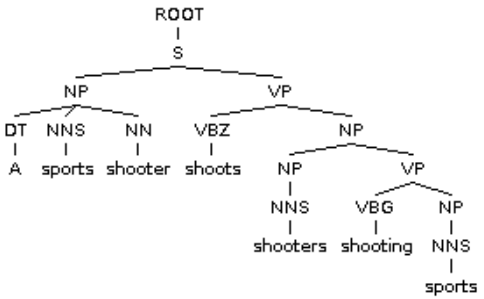
Parse!

Language and Computers

Berkeley Parser
http://tomato.banatao.berkeley.edu:8080/parser/parser.html

This one is available
Berkeley Neural Parser
https://parser.kitaev.io

```
(ROOT
  (S
    (NP (DT A) (NNS sports) (NN shooter))
    (VP (VBZ shoots)
      (NP
        (NP (NNS shooters))
        (VP (VBG shooting)
          (NP (NNS sports)))))))
```

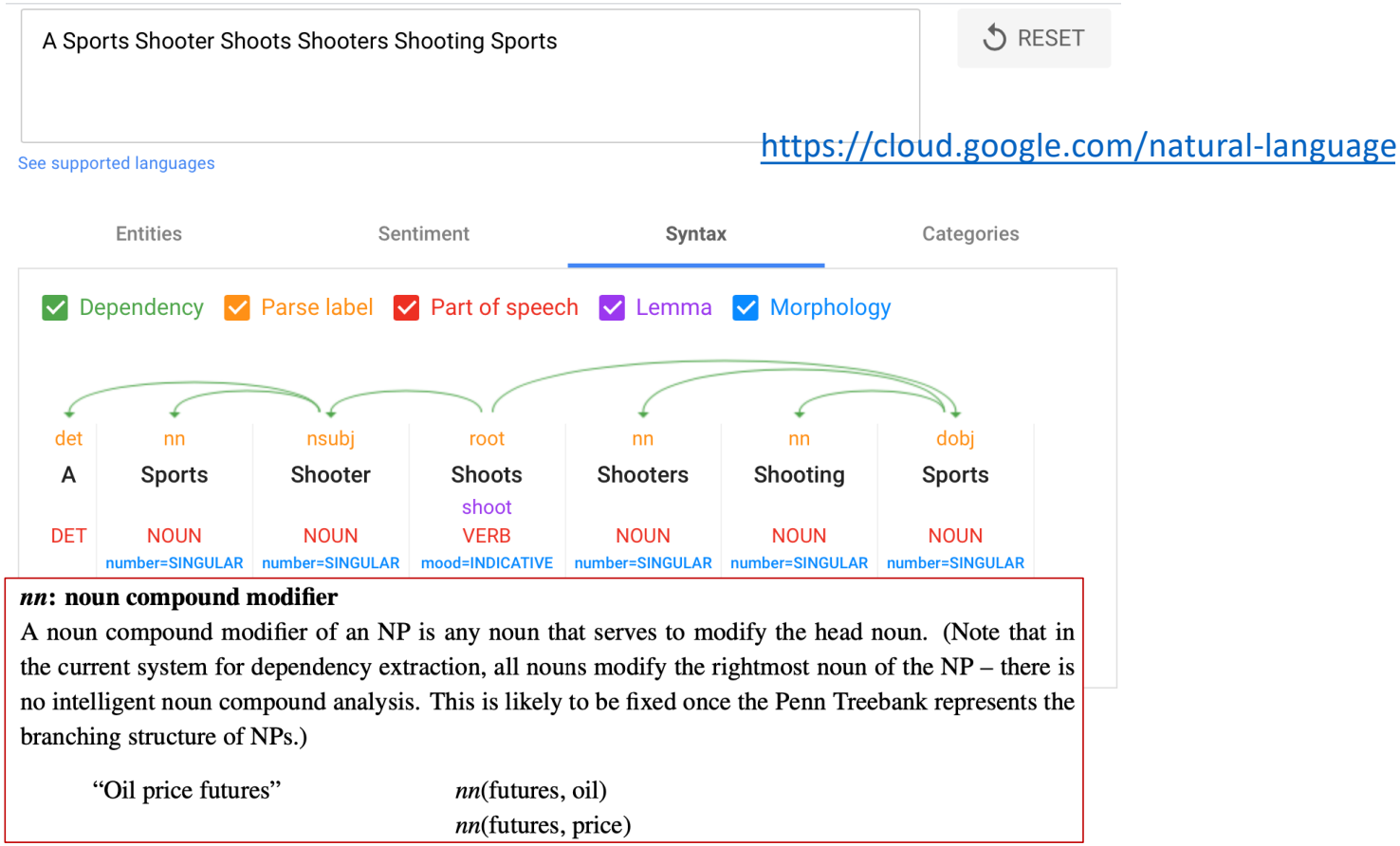


FRAG = Fragment (of a sentence)
S = Sentence
NP = Noun Phrase
VP = Verb Phrase

A sports shooter shoots shooters shooting sports

Parse!

Language and Computers



***nn*: noun compound modifier**

A noun compound modifier of an NP is any noun that serves to modify the head noun. (Note that in the current system for dependency extraction, all nouns modify the rightmost noun of the NP – there is no intelligent noun compound analysis. This is likely to be fixed once the Penn Treebank represents the branching structure of NPs.)

"Oil price futures"	*nn*(futures, oil)
	nn(futures, price)

Language and Computers

Universal Dependency Relations

The following table lists the 37 universal syntactic relations used in UD v2. It is a revised version of the relations originally described in *Universal Stanford Dependencies: A cross-linguistic typology* (de Marneffe *et al.* 2014).

The upper part of the table follows the main organizing principles of the UD taxonomy such that *rows* correspond to functional categories in relation to the head (core arguments of clausal predicates, non-core dependents of clausal predicates, and dependents of nominals) while *columns* correspond to structural categories of the dependent (nominals, clauses, modifier words, function words). The lower part of the table lists relations that are not dependency relations in the narrow sense.

	Nominals	Clauses	Modifier words	Function Words
Core arguments	nsubj obj iobj	csubj ccomp xcomp		
Non-core dependents	obl vocative expl dislocated	advcl	advmod* discourse	aux cop mark
Nominal dependents	nmod appos nummod	acl	amod	det clf case
Coordination	**MWE**	**Loose**	**Special**	**Other**
conj cc	fixed flat compound	list parataxis	orphan goeswith reparandum	punct root dep

* The `advmod` relation is used for modifiers not only of predicates but also of other modifier words.

https://universaldependencies.org/u/dep/index.html

Natural Language Properties

- *which properties are going to be difficult for computers to deal with*?
- **grammar** (Rules for putting words together into sentences)
  - *How many rules are there*?
    - 100, 1000, 10000, more …
  - Portions learnt or innate
  - *Do we have all the rules written down somewhere*?
- **lexicon** (Dictionary)
  - How many words do we need to know?
    - 1000, 10000, 100000 …
- **meaning and inference** (semantic interpretation, commonsense world knowledge)

OpenAI GPT-2 Explorer

- Read this article:
  - https://www.bbc.com/news/technology-49446729

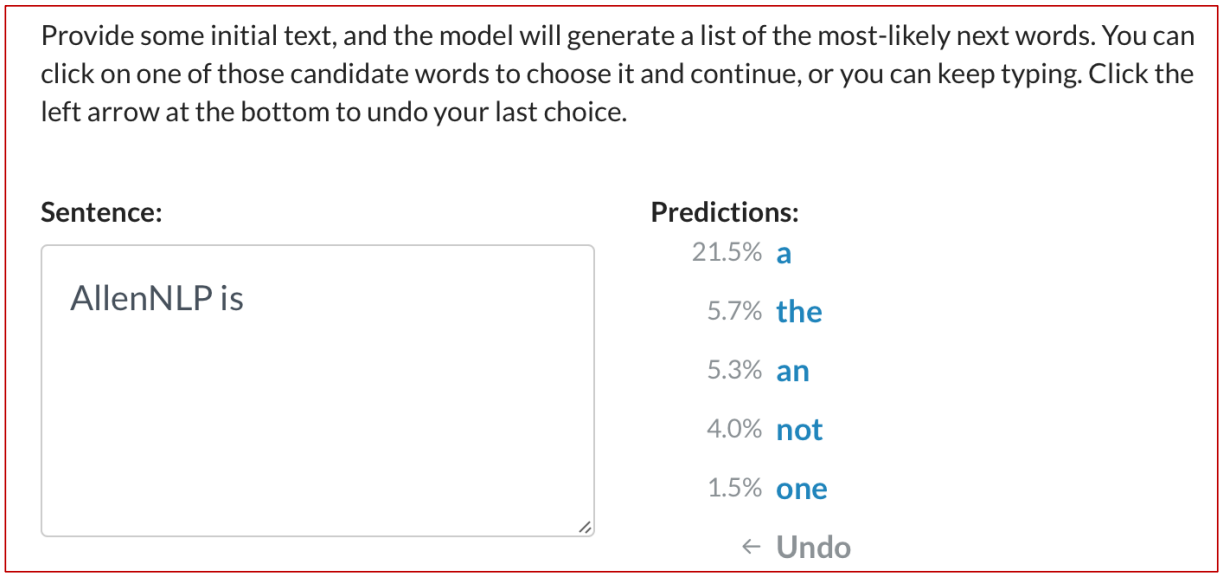
'Dangerous' AI offers to write fake news

By Jane Wakefield Technology reporter

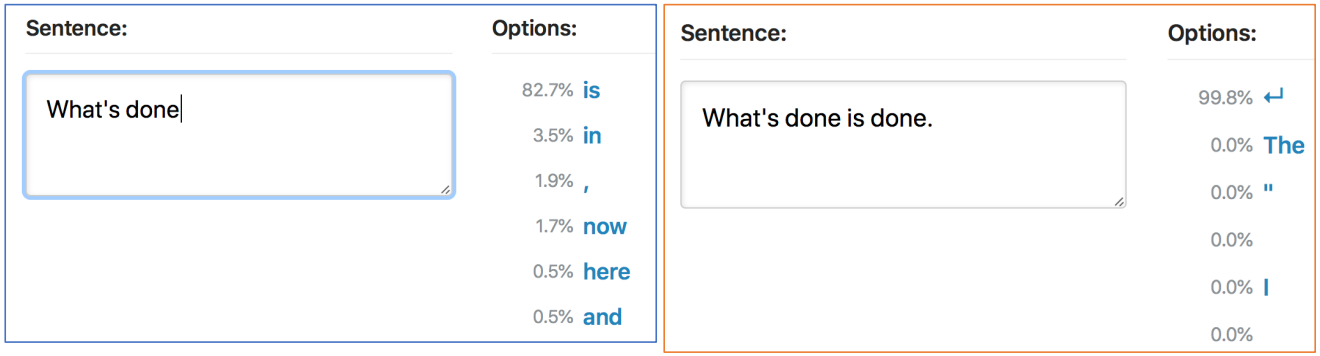
Image copyright Getty Images

OpenAI GPT-2 Explorer

- Go to the Language Modeling demo here:
  - https://gpt2.apps.allenai.org

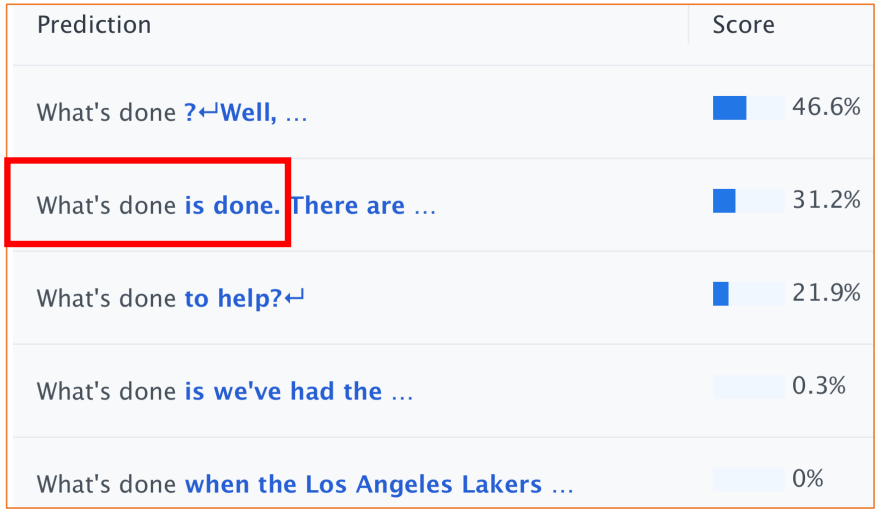


What's done …



It has memorized a line from Shakespeare!

What's done …



- The demo has changed.
  - *Yesterday, I got …*

So fair and foul a day …

*Enter **MACBETH** and **BANQUO***

MACBETH

So foul and fair a day I have not seen.

BANQUO

How far is 't called to Forres?—What are these

So withered and so wild in their attire,

That look not like th' inhabitants o' th' Earth,

And yet are on 't?—Live you? Or are you aught

***MACBETH** and **BANQUO** enter.*

MACBETH

(to BANQUO*)* I have never seen a day that was so good and bad at the same time.

BANQUO

How far is it supposed to be to Forres? *(he sees the* WITCHES*)* What are these creatures? They're so withered-looking and crazily dressed. They don't look like they belong on this planet, but I see them standing here on Earth. *(to the* WITCHES*)* Are you alive? Can you

https://www.sparknotes.com/nofear/shakespeare/macbeth/page_12/

So fair and foul a day …

Sentence:

So foul and fair a day

Predictions:

14.1% ,

8.7% .

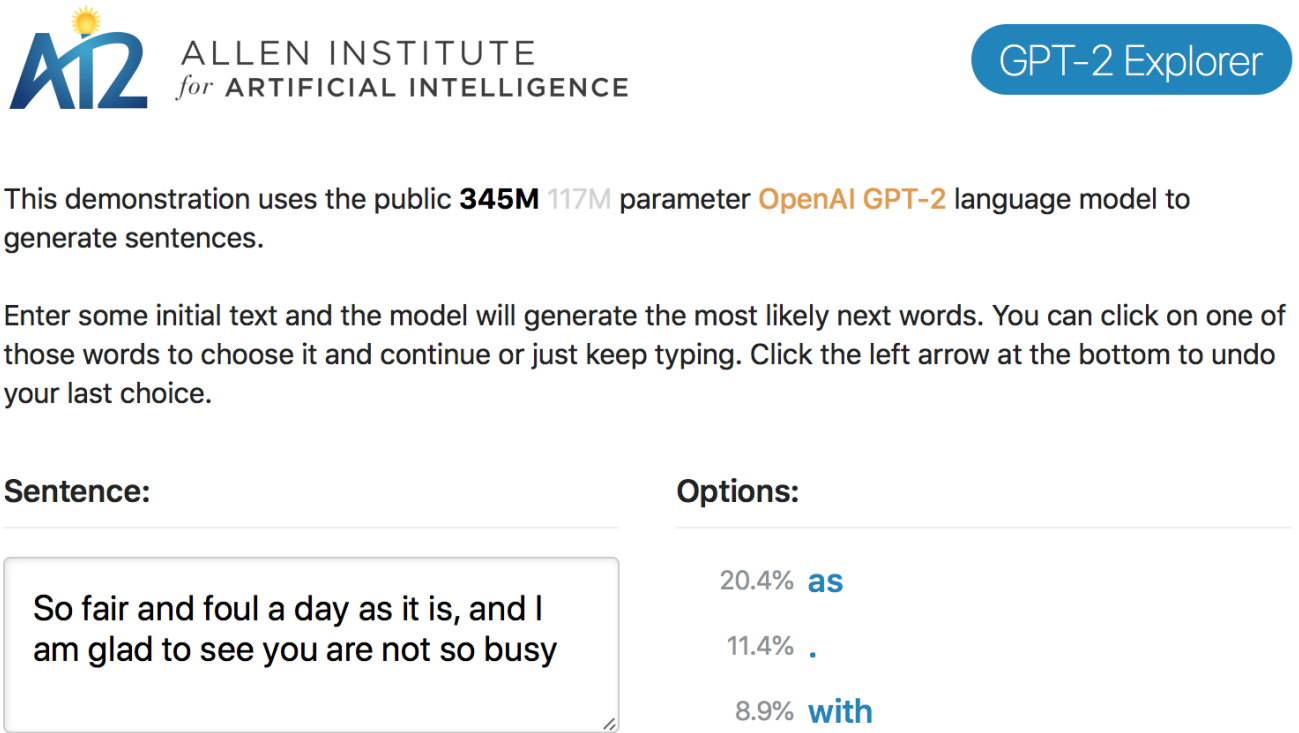
7.2% for

4.9% to

4.1% ,"

← Undo

So fair and foul a day …



So fair and foul a day …

Sentence:

So foul and fair a day, and I'll be damned if I don't get it."

Predictions:

99.8% ↵

0.0% "

0.0% The

0.0% .

0.0% <|endoftext|>

← Undo

So fair and foul a day …

Prediction	Score
So fair and foul a day **," he said.↵**	99.8%
So fair and foul a day **you have no right to** …	0.2%
So fair and foul a day **, I will call you** …	0%
So fair and foul a day **for you!" And that** …	0%
So fair and foul a day **Saturday at the rink.**	0%

Homework 3

(1.2) How much Chinese silk was exported to Western Europe by the end of the 18th century?

To answer this question, we need to know something about **lexical semantics**, the meaning of all the words (*export* or *silk*) as well as **compositional semantics** (what exactly constitutes *Western Europe* as opposed to Eastern or Southern Europe, what does *end* mean when combined with *the 18th century*. We also need to know something about the relationship of the words to the syntactic structure. For example, we need to know that *by the end of the 18th century* is a temporal end-point and not a description of the agent, as the by-phrase is in the following sentence:

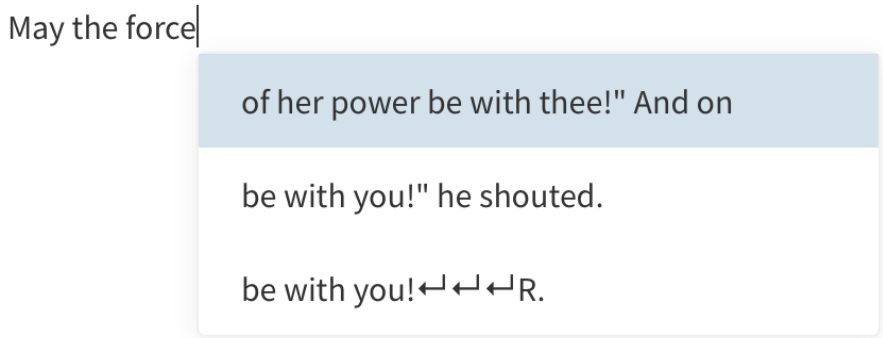
(1.3) How much Chinese silk was exported to Western Europe by southern merchants?

Homework 3

- Question 1
  - Try syntactic analysis using two of the parsers mentioned in today's lecture.
  - Screenshot the output.
  - Do your parsers distinguish the two distinct senses of preposition *by* highlighted in the previous slide?
  - Submit just a line or three justifying your answer.

Homework 3

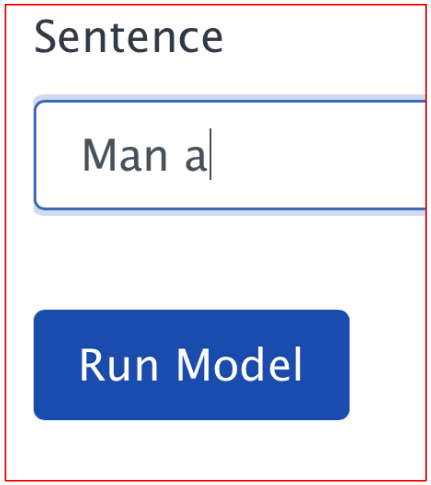
- Question 2: use the GPT-2 demo below
  - https://transformer.huggingface.co/doc/gpt2-large



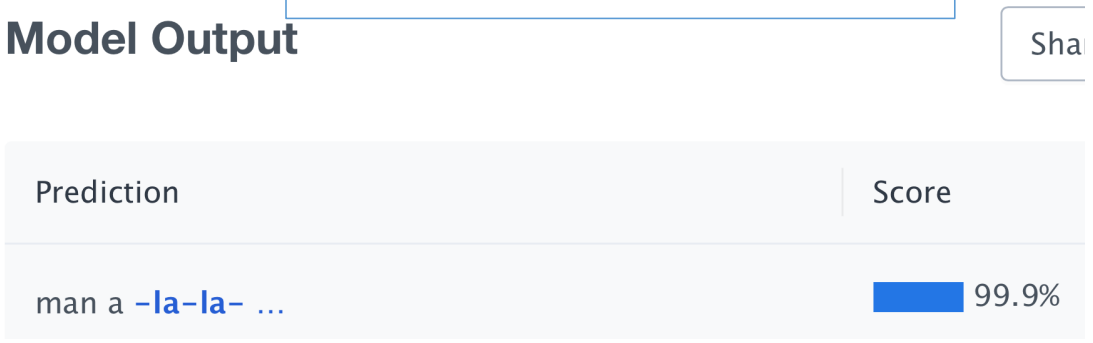
on a couple of famous quotes, e.g. movie or from someone, and see how it completes it.

Homework 3

- Question 3: what happens if you start with something ungrammatical? Define what you mean by ungrammatical.
- Can you make it do something strange?



It went into an infinite loop:
Man a –la-la- la la la …



Homework 3

- Instructions:
  - Put all your answers in one PDF document (not Word .docx or .doc)
  - email to me (sandiway@email.arizona.edu)
  - subject line: 438/538 Homework 3 YOUR NAME
  - Due date: by midnight Sunday (i.e. before next class!)