# Introduction to Natural Language Processing

CS505
Spring 2022

# Course Information

- CS505 (Discussion and Resources @Piazza)
- https://piazza.com/bu/spring2022/cs505/info
- **Lectures**: Tuesday and Thursday 8 – 9.15 am at CAS 211

  or Zoom:

  https://bostonu.zoom.us/j/98138949567?pwd=M2dDWmtrR09tbGNGYmZjbTdCa2sxQT09

  **Lab**: held on the week an assignment is due

# Course Information

- Office hour:
  - (Derry) Fridays 10-11am at Zoom
    - (TAs) Check Piazza for information
- Professor: Derry Wijaya (wijaya@bu.edu)
- TA: Yusuf Kocygit (kocyigit@bu.edu)
- TA: Afra Feyza Akyurek (akyurek@bu.edu)

# Professor Wijaya

- Bachelors from National University of Singapore
- PhD from Carnegie Mellon University
- Postdoc at University of Pennsylvania
- Developed interest in the field of Information Extraction as an undergraduate doing UROP

# What will you learn?

- This is a *introduction* class to natural language processing

- Focus will be on *introducing* you to (1) key concepts from NLP used to describe and analyze languages (2) how to process and analyze language data

# Topics

- We'll cover topics in Jurafsky and Martin, Speech and Language Processing, 3rd edition (draft available online: https://web.stanford.edu/~jurafsky/slp3/
- Topics include: text processing, text classification, language modeling, vector semantics, neural language modeling, recurrent neural networks for NLP, unsupervised learning, part-of-speech tagging, context free grammar and parsing, information extraction, machine translation, QA, dialogue systems

# Assignments

- There are 5 assignments that include programming in Python that must be done individually
- ~1 assignment every 2 weeks:
  - Assignments will be released on Fridays
  - Due on Sunday (midnight EST) the following week

# Grading

- Each assignment contributes **12 points**
- Lecture attendance (%) *OR* Piazza participation (top-10) (bonus up to **5 points**)
  - https://forms.gle/ifktRCxa4BfEdDrH7
  - If you cannot attend but want to get attendance point, please watch the video and create a short summary of the lecture and submit it via this Form: https://forms.gle/AYSHyQo5M2tdFLUu5
- Project:
  Team of 5 (**20 points** for presentation, **20 points** for write-up)

# About Assignments

- This class will have individual programming assignments that are intense
  - The assignments will be real-life programming assignments using real-world datasets (no toy datasets here)
    - You will get hands-on experience on scraping, processing, cleaning, and analyzing textual data – things you will actually need to know at your current/future work
    - The focus is on building end-to-end actual NLP systems
- Require experience in Python programming and for machine learning

# Important Dates

- Feb 2, 2022 Last day to ADD course
- Feb 24, 2022 Last day to DROP course (w/o ”W”)
- Apr 1, 2022 Last day to DROP with a ”W” grade

# About Assignments

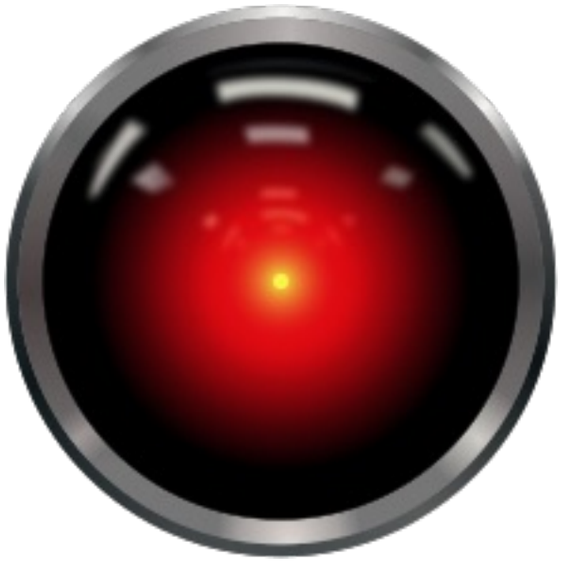
- This is also a class that is *Introduction* to NLP so by definition, students come from various backgrounds
- Therefore, there may be people who will find it *not* challenging enough
- Email me if you want extra challenge – I can give extra, research-level NLP, challenge for you

# Mark Your Dates

*Linguistics and NLP Talks*

- Thursday, Feb. 3, 2022 (5:00pm-6:00pm, RKC-101) - Nelson Flores (University of Pennsylvania)
- Monday, Feb. 28, 2022 (4:30pm-5:30pm, location TBA) - Tracy Conner (Northwestern University)
- Monday, Mar. 28, 2022 (4:30pm-5:30pm, location TBA) - Najoung Kim (New York University)
- Monday, Apr 11, 2022 (11am – 12.30pm, location TBA) – Lucia Specia (Imperial College London)
- Monday, Apr. 25, 2022 (4:30pm-5:30pm, location TBA) - Norma Mendoza-Denton (University of California, Los Angeles)

# Introduction




- HAL 9000 in 2001: A Space Odyssey (1968)
- Her (2013)

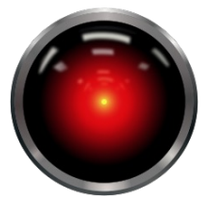
# Introduction

- Natural Language Processing
  - techniques to give computers the ability to process, learn, and understand human languages
  - other names: speech and language processing, human language technologies, computational linguistics, speech recognition and synthesis, …

# Goal of NLP

**get computers to perform tasks involving human language**

- conversational agent

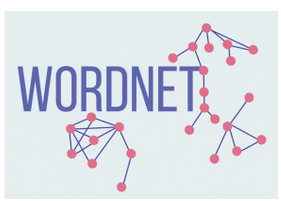



- machine translation (MT)

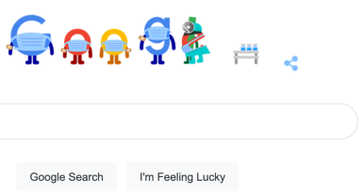
- question answering (QA)
- information extraction (IE)



- word sense disambiguation (WSD)

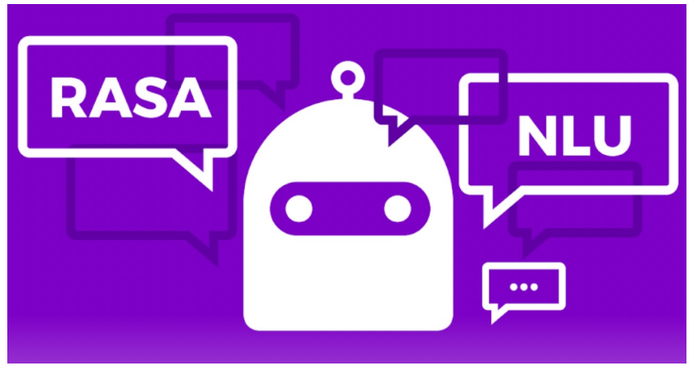

- search engine ...



# Goal of NLP

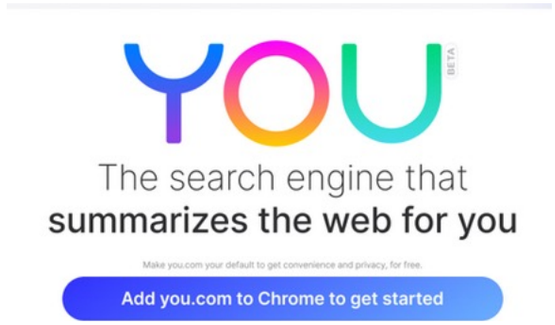
**get computers to perform tasks involving human language**



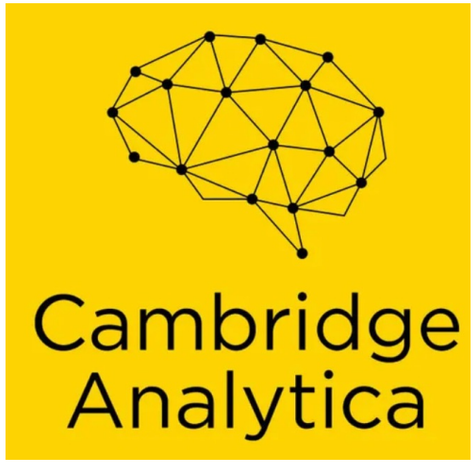


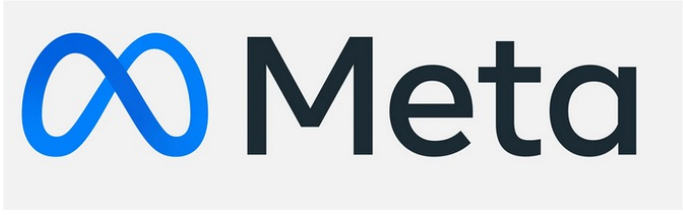


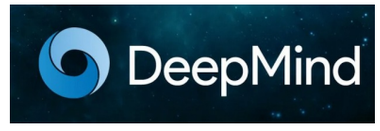


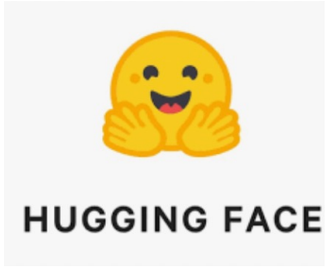








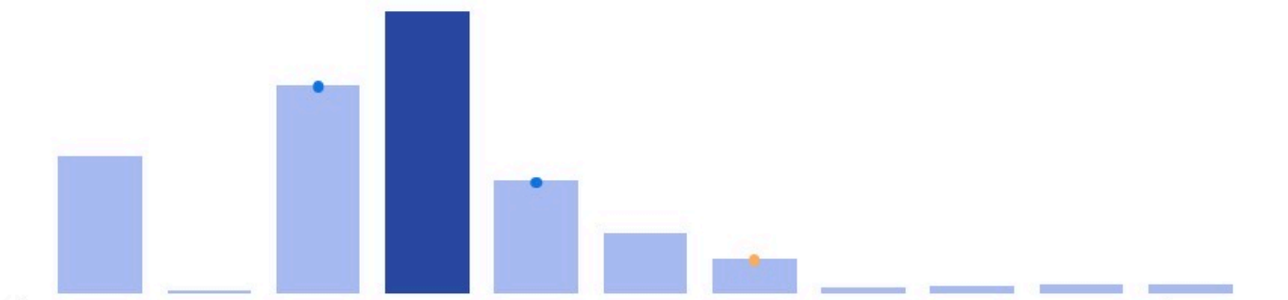



# Goal of NLP-practitioner

**get computers to perform tasks involving human language**

**Artificial Intelligence Engineer Salary**

Yearly | Monthly | Weekly | Hourly | Table View




$164,769 /year

$79 /hour

**Artificial Intelligence Engineer Salary Comparison by Location**

# To A Computer

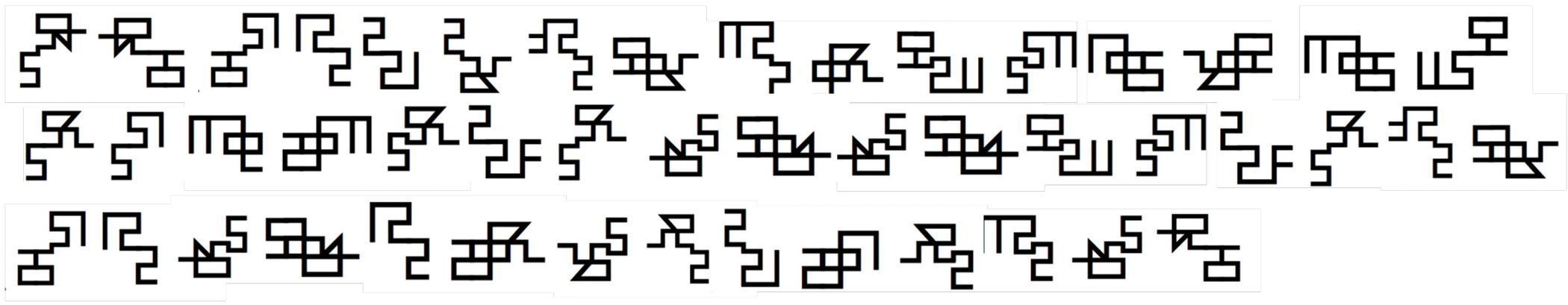
This text

bakaji bavwa bamona muana. baluma bavwa bamona bakaji. Tubambwa tuvwa tumona baluma. kanzolu kavwa kamona tubambwa. cimuma civwa cimona ntambwe. ntambwe uvwa mumona tubimuma.

To a non-speaker of the language

# To A Computer

This text



To a non-speaker of the language

# Goal of NLP

Computer
solving task
involving human
language

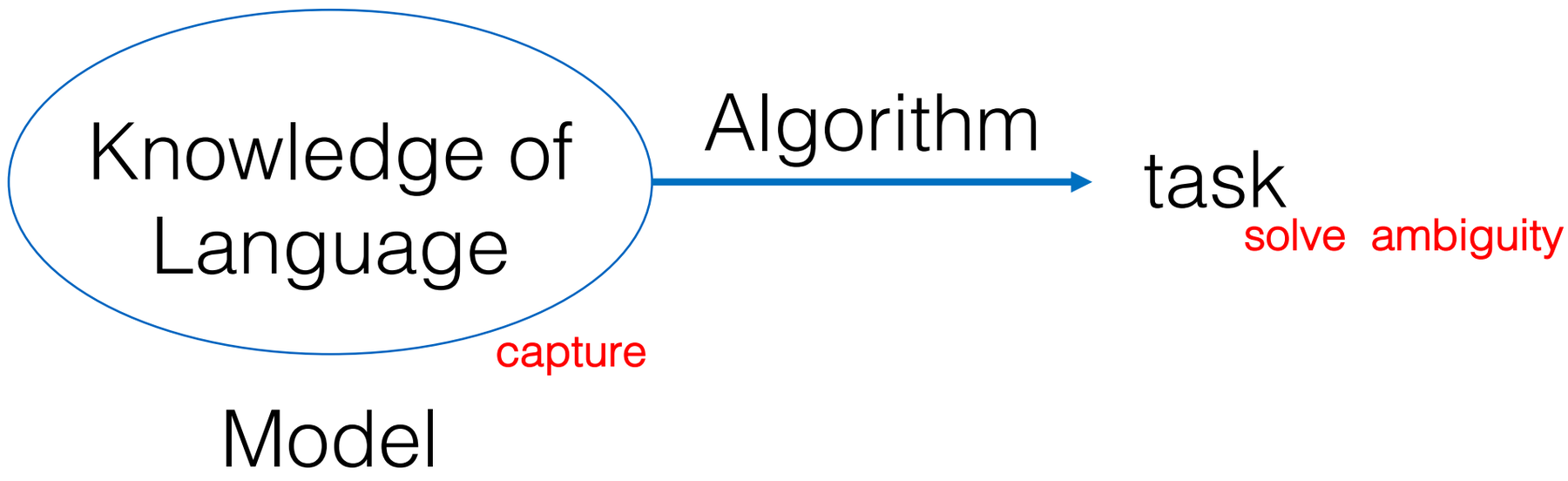
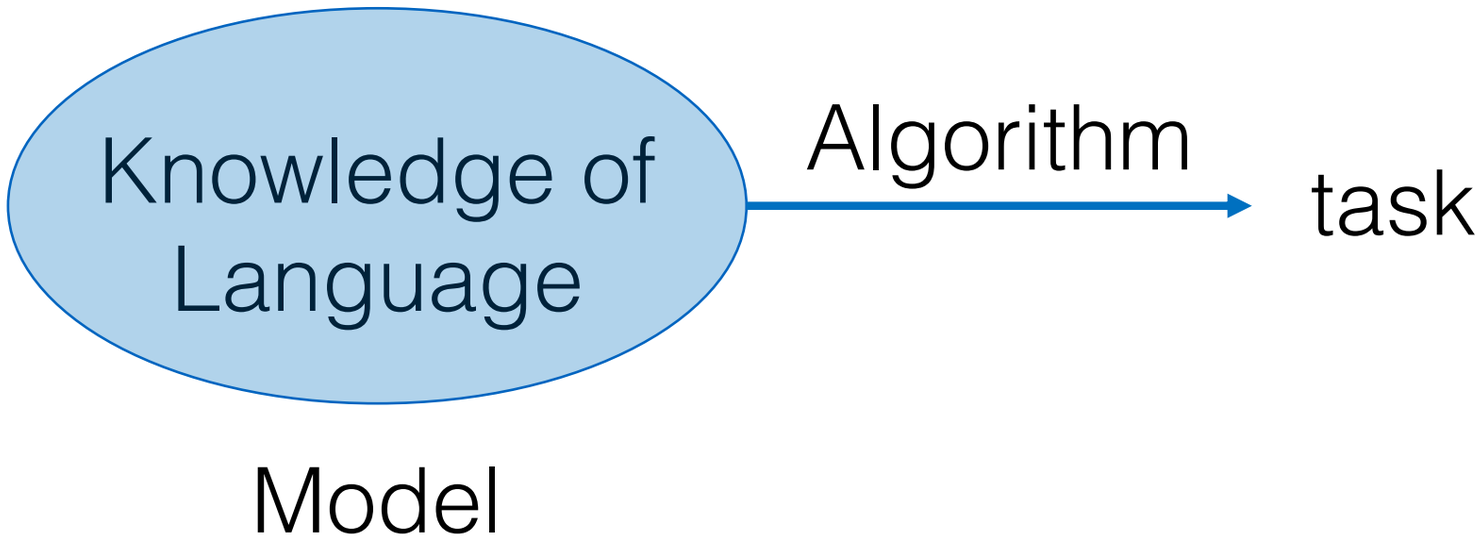



Image source: medium.com

# Goal of NLP

Computer solving task involving human language




Model

Image source: medium.com

# Knowledge of Language

**phonetics and phonology**

- phonetics: <u>production</u> of sounds: *physical description* of sounds (position of mouth, tongue, etc.) i.e., how sounds are used in human speech

- phonology: <u>patterns</u> of sounds: description of *sound interrelation* and *function* i.e., how speech sounds are used in a particular language e.g., **c**art vs. **c**oot

# Knowledge of Language

**morphology**

- different *variations of words* e.g., door vs doors (singular vs. plural), kejar vs. dikejar (chase vs. to be chased, in Indonesian language)

**syntax**

- *structural knowledge* to order and group words together
  - I'm sorry Dave, I'm afraid I can't do that
  - (*) I'm I do, sorry that afraid Dave I'm I can't

# Knowledge of Language

**lexical semantics**

- *meaning of words* e.g., to answer the question:

> How much Chinese **silk** was **exported** to Western Europe by the end of the 18th century?

- need to know the meaning of “silk”, “export”, …

# Knowledge of Language

**compositional semantics**

- meaning of words in *combination* e.g., to answer the question:

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

- need to know the meaning of *Chinese silk* as opposed to *Assam silk*, for example; or what *Western Europe* means as opposed to *Eastern Europe*; or *what end of the 18th century* mean as opposed to *end of that road*?

# Knowledge of Language

**pragmatic**

| | |
|---|---|
| REQUEST: | *HAL, open the pod bay door.* |
| STATEMENT: | *HAL, the pod bay door is open.* |
| INFORMATION QUESTION: | *HAL, is the pod bay door open?* |

- knowledge about the kind of *actions* that speakers *intend* by their use of sentences

# Knowledge of Language

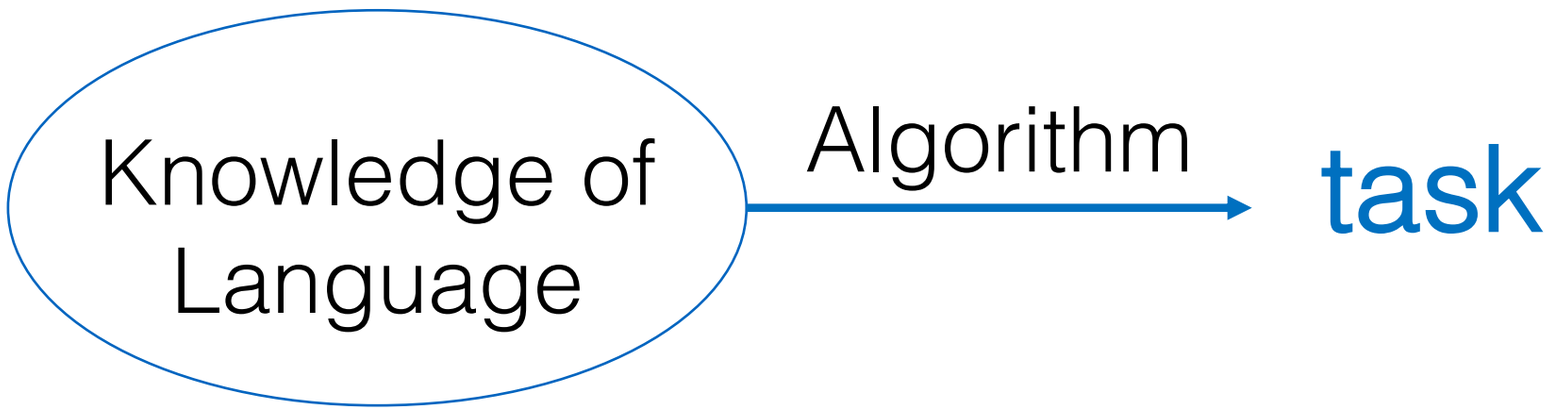
**discourse**

- What year was Abraham Lincoln born?
- How many states were in the United States that year?

- knowledge about linguistic units *larger than a single utterance* e.g., coreference resolution, “that year” == “the year Abraham Lincoln was born”

# Goal of NLP

Computer
solving task
involving human
language




Model

Image source: medium.com

# Ambiguity

- Most *tasks* in NLP are about *resolving ambiguities*

I made her duck

noun?

a duck
the ducks
it's a duck
(*) he duck

verb?

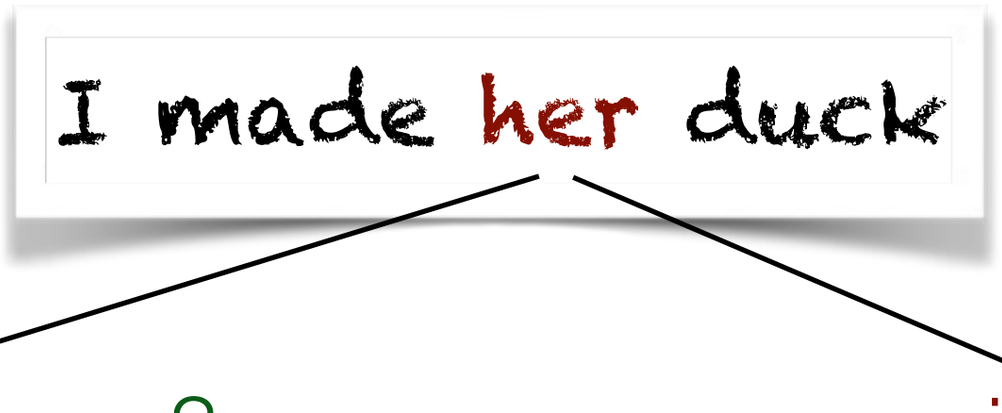
he ducks
he ducked
he's ducking
(*) a ducked

resolving part-of-speech (morphology)

words with same part-of-speech share similar forms

# Ambiguity

- Most *tasks* in NLP are about *resolving ambiguities*




dative pronoun?

"I cooked her a duck dish"

possessive pronoun?

"I cooked the duck dish belonging to her"

resolving syntactic ambiguity

# Ambiguity

- Most *tasks* in NLP are about *resolving ambiguities*


cooked?

created?

resolving semantic ambiguity (word senses)

# Goal of NLP

Computer
solving task
involving human
language

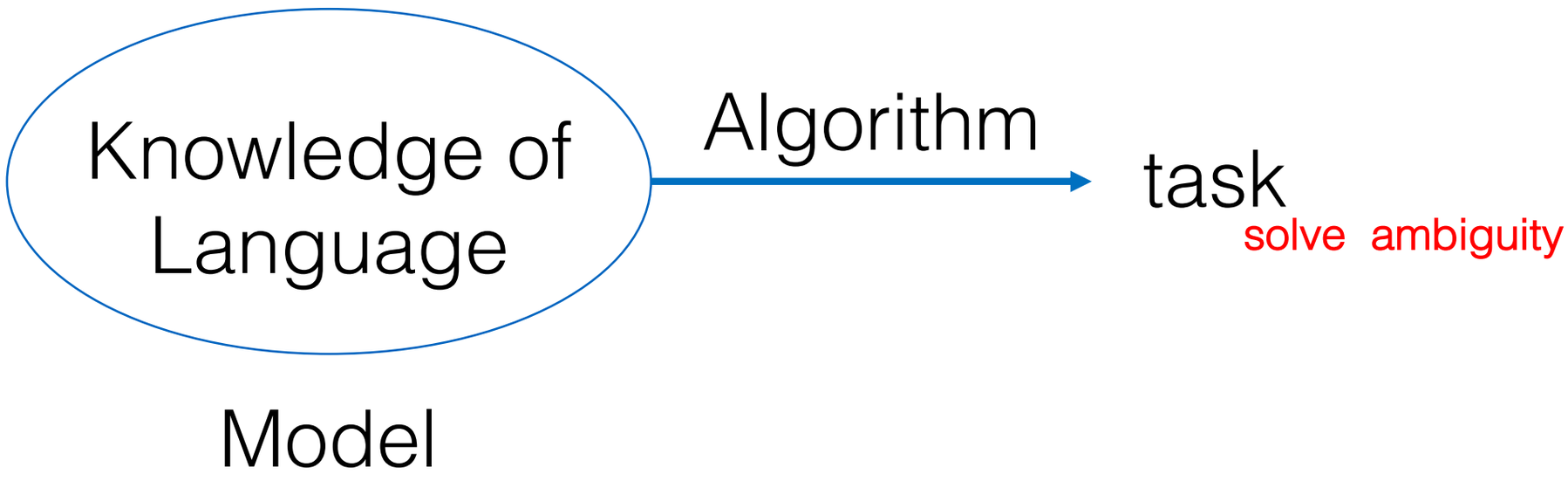



Image source: medium.com

# Goal of NLP

Computer
solving task
involving human
language

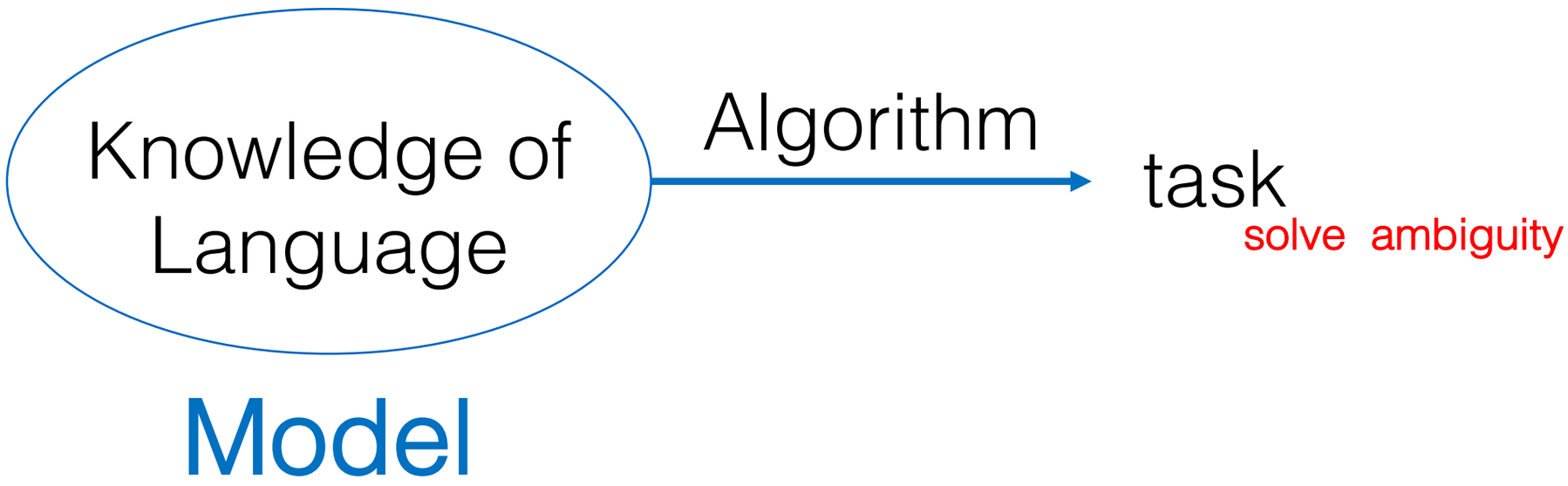



Image source: medium.com

# Models

- NLP systems rely on *models to capture knowledge* of language e.g., formal rule systems to capture phonology, morphology, and syntax

context free grammar

a set of **production rules** that describe all possible strings

| | |
|---|---|
| S → | NP① VP② |
| VP → | VP② PP① |
| VP → | AUX② V① |
| PP → | P② NP① |
| NP → | *Hamid Ansari* |
| NP → | *Vice President* |
| V → | *nominated* |
| P → | *for* |
| AUX → | *was* |

"Hamid Ansari was nominated for Vice President"

"Vice President was nominated for Hamid Ansari"

"Hamid Ansari was nominated"

"Vice President was nominated"

rules can be applied **regardless** of context

# Models

- e.g., probabilistic models: ambiguity problems can be recast as "***given N choices for some ambiguous input, choose the most probable one***"

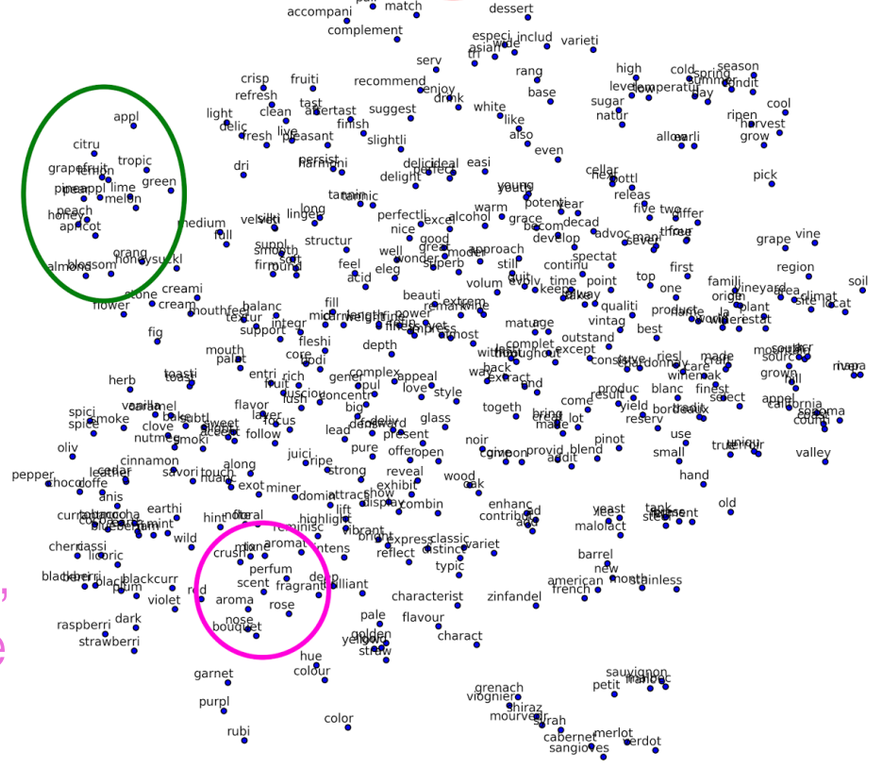
# Models

- e.g., vector space models to capture word meanings "**you shall know a word by the company it keeps (Firth, J. R. 1957:11)**"

pasta, lamb,
cheese, mushroom

citrus, apple,
orange, lime

aromatic, nose,
scent, perfume



http://methodmatters.blogspot.com/2017/11/using-word2vec-to-analyze-word.html

# End of Lecture 1

# Beginning of Lecture 2

# Goal of NLP

Computer
solving task
involving human
language

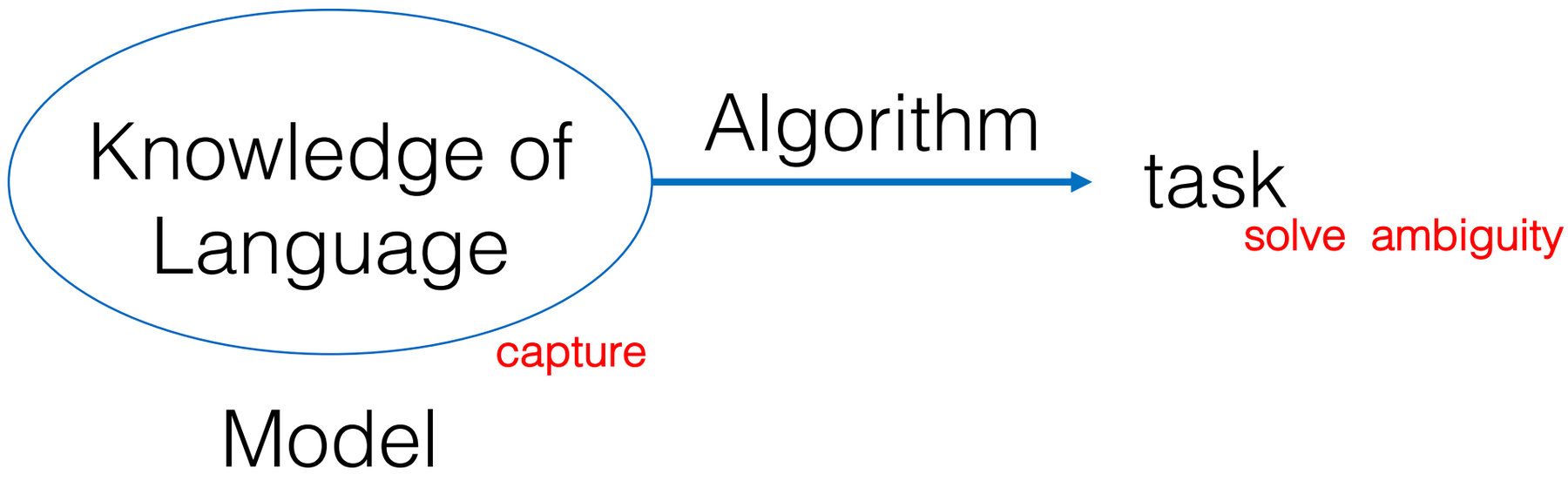



Image source: medium.com

# Goal of NLP

Computer
solving task
involving human
language

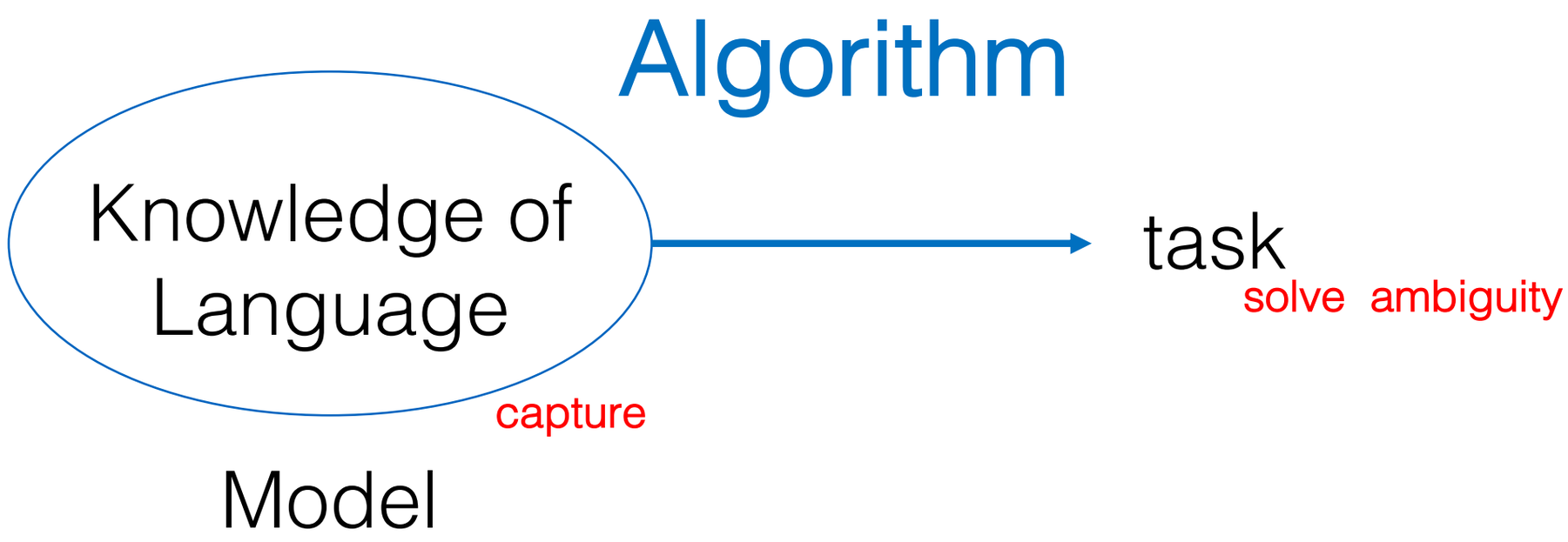



Image source: medium.com

# Algorithms

- Given the models, ***search*** through a space of ***hypotheses*** about an input

- e.g., a *classifier* can be trained to compute the sentiment polarity of a word i.e., whether it conveys a positive/negative sentiment based on the word vector

- e.g., a *machine translation algorithm* searches through a space of translation hypotheses for the correct translation of a sentence into another language

# Goal of NLP

Computer
solving task
involving human
language

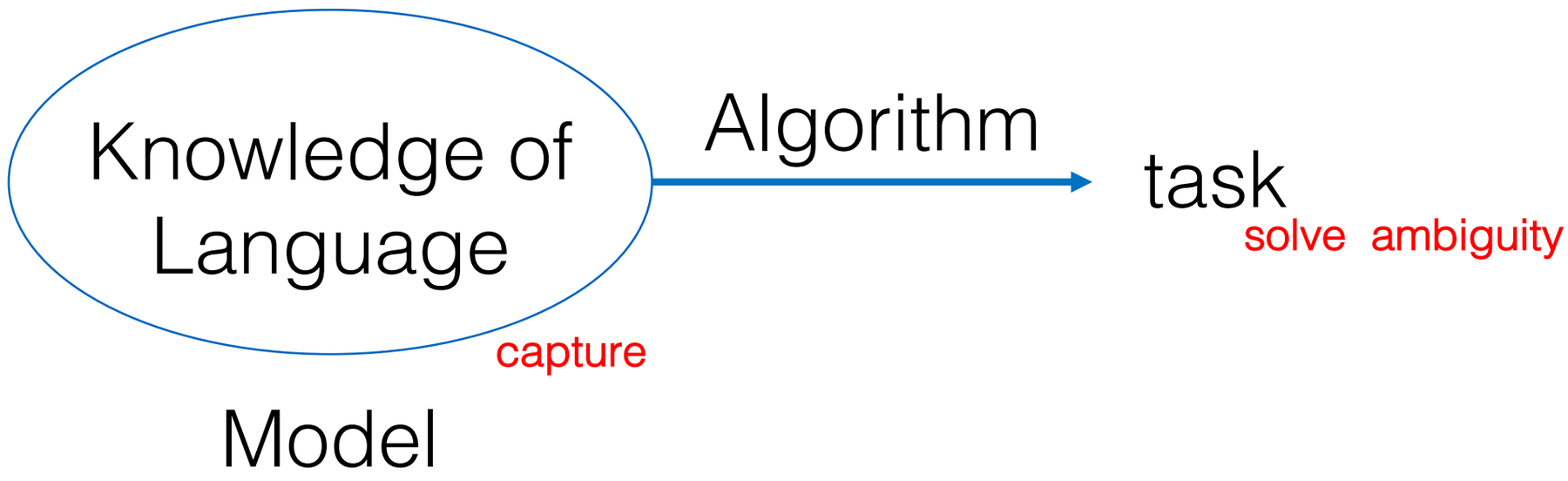


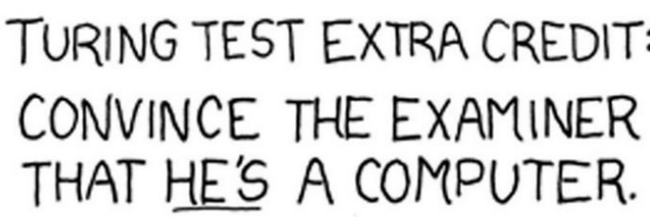

Image source: medium.com

# NLP and the Measure of Intelligence

using language as humans do
== *truly intelligent* machines?

**Turing Test**

by responding as a person to the examiner's questions, the machine wins if it can convince the examiner into believing that it is a person





https://xkcd.com/329/

# NLP and the Measure of Intelligence

## ELIZA program (Weizenbaum, 1966)

NLP system that imitates a psychotherapist

| | |
|---|---|
| $User_1$: | 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 |

ELIZA **uses pattern matching**, **knows nothing** of the world,
**but** many people thought that it really ***understood*** them and their problems!

# NLP and the Measure of Intelligence

## ELIZA program (Weizenbaum, 1966)

NLP system that imitates a psychotherapist

| | |
|---|---|
| User$_1$: | You are like my father in some ways. |
| ELIZA$_1$: | WHAT RESEMBLANCE DO YOU SEE |
| User$_2$: | You [illegible] to notice that. |
| ELIZA$_2$: | WH[illegible] |
| User$_3$: | Yo[illegible] |
| ELIZA$_3$: | WH[illegible] |
| User$_4$: | You are afraid of me. |
| ELIZA$_4$: | DOES IT PLEASE YOU TO BELIEVE I AM AFRAID OF YOU |

says more about the people than about the machine

ELIZA **uses pattern matching**, **knows nothing** of the world,
**but** many people thought that it really ***understood*** them and their problems!

# NLP and the Measure of Intelligence

using language as humans do
== *truly intelligent* machines?

regardless, people talk about computers and
interact with them as social entities;
***expecting* computers to *understand* their needs** and
be able to *interact naturally* (Reeves and Nass 1996)

# NLP and the Measure of Intelligence

using language as humans do
== *truly intelligent* machines?

regardless, people talk about computers and interact with them as social entities;
***expecting* computers to *understand* their needs** and be able to *interact naturally* (Reeves and Nass 1996)

**The importance of NLP!**

# An **Exciting** Time for NLP!

- Increase in **computing resources**
- Increase in the **amount of data** and information available in digital form
- Development of highly successful **machine learning methods** and **competitive evaluations** (SemEval, NIST, CoNLL shared tasks, Kaggle)
- Richer understanding of the structure of human language and its **deployment in social contexts**

# An **Exciting** Time for NLP!

- Real world NLP systems
  - Conversational agents for making reservations
  - Voice assistant systems
  - Machine translation systems, speech-to-speech translation
  - Automated grading systems
  - Virtual tutors
  - Text analysis companies for marketing intelligence
  - …

# State of the Art

- Simple methods often work very well when trained on **large quantities of data**
  - e.g., many text and sentiment classifiers still rely on different sets of words (“bag of words”) without regard to sentence and discourse structure or meaning

# State of the Art

- However, most NLP resources and systems are **available only for high resource languages**
  - Many low resource languages are spoken by millions of people e.g., Bengali, Indonesian, Swahili, …
  - The challenge is how to develop resources and tools for thousands of languages, not just a few

# Machine Translation

- Challenging because correct translations require:
  - ability to analyze and generate sentences in human languages
  - understanding of world knowledge and context to resolve the ambiguities of languages
    - e.g., the French word “bordel” straightforward translation is “brothel” but what if someone says “*My room is un bordel*”?

# Machine Translation

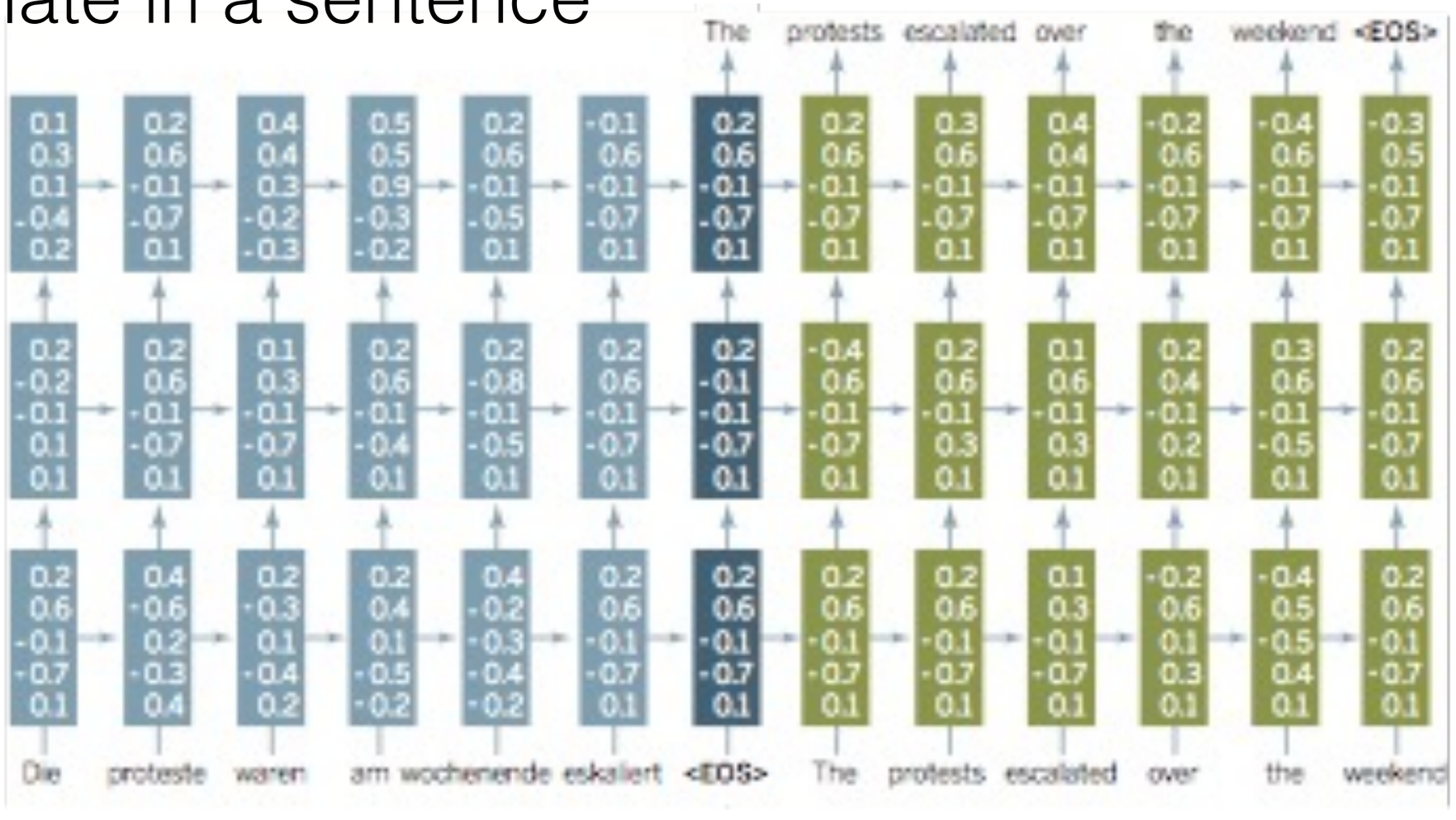
- Started with hand-built grammar based systems (limited success)
- Transformed with the availability of **parallel sentences** to collect statistics of word translations and word sequences
    - small word groups often have distinctive translations — phrase based MT, which formed the basis of Google translate

# Machine Translation

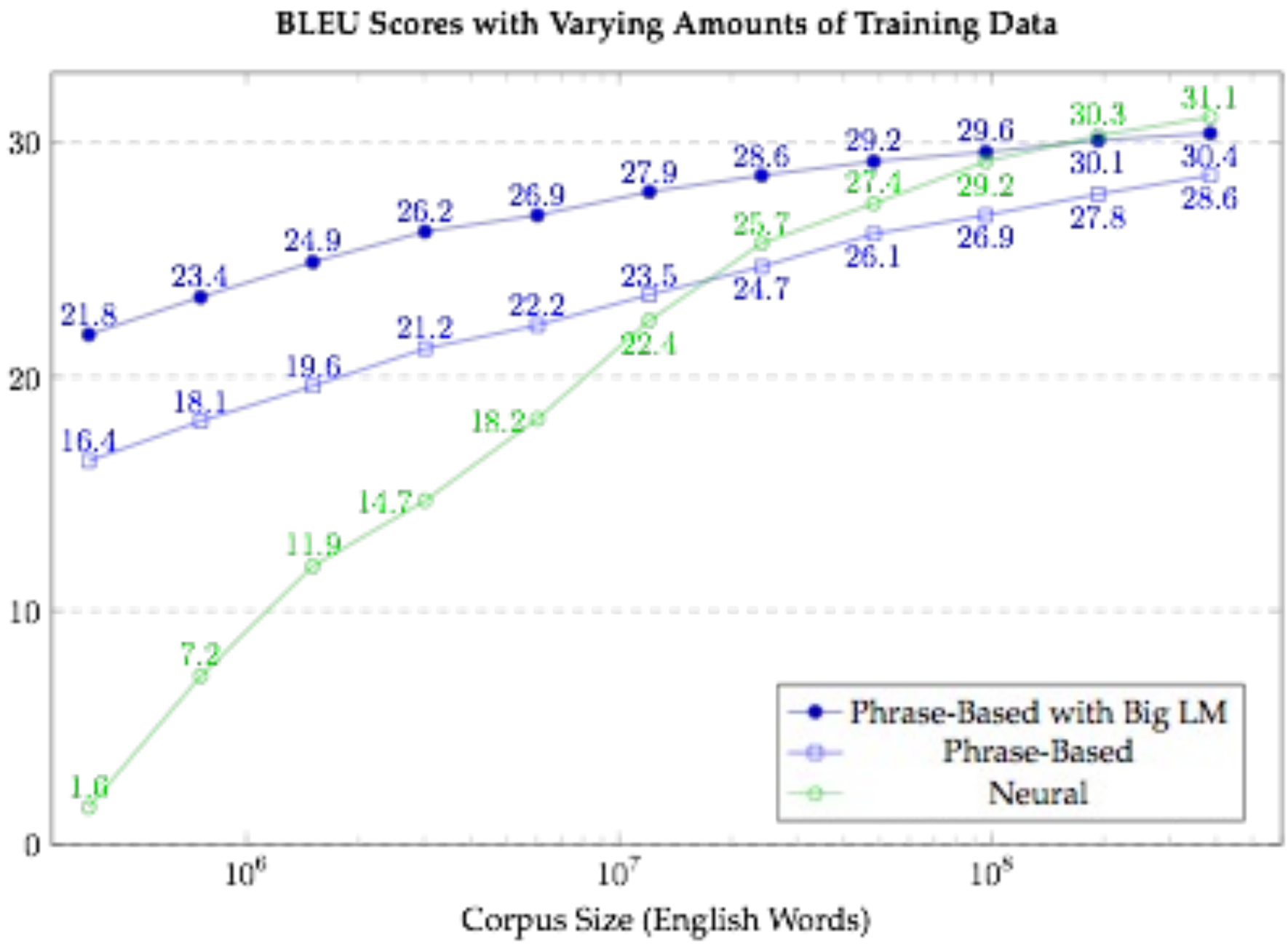
- Another transformation using **deep learning-based** sequence models
  - train a model with several representational levels to optimize translation quality
  - the model learns intermediate representations that are useful for translation

# Machine Translation

- **Long Short Term Memory** Network
  - maintain contextual information from early until late in a sentence



# Machine Translation




[Neural MT, Philipp Koehn]

# Machine Translation

| Ratio | Words | Source: *A Republican strategy to counter the re-election of Obama* |
|---|---|---|
| $\frac{1}{1024}$ | 0.4 million | *Un órgano de coordinación para el anuncio de libre determinación* |
| $\frac{1}{512}$ | 0.8 million | *Lista de una estrategia para luchar contra la elección de hojas de Ohio* |
| $\frac{1}{256}$ | 1.5 million | *Explosión realiza una estrategia divisiva de luchar contra las elecciones de autor* |
| $\frac{1}{128}$ | 3.0 million | *Una estrategia republicana para la eliminación de la reelección de Obama* |
| $\frac{1}{64}$ | 6.0 million | *Estrategia siria para contrarrestar la reelección del Obama .* |
| $\frac{1}{32}+$ | 12.0 million | *Una estrategia republicana para contrarrestar la reelección de Obama* |

Figure 13.49: Translations of the first sentence of the test set using neural machine translation system trained on varying amounts of training data. Under low resource conditions, neural machine translation produces fluent output unrelated to the input.

Neural MT, Philipp Koehn

# Lessons so far

- More data -> better performance
- Less data ?
  - new approach?
  - look for more data?

# The State and Fate of Linguistic Diversity and Inclusion in the NLP World

**Pratik Joshi*** **Sebastin Santy*** **Amar Budhiraja***
**Kalika Bali** **Monojit Choudhury**
Microsoft Research, India
{t-prjos, t-sesan, amar.budhiraja, kalikab, monojitc}@microsoft.com

ACL 2020

# The State

- 7000 languages
- Only a small number are represented in NLP
- Disparities in terms of resources: labeled and unlabeled data
- Are language agnostic systems really language agnostic?
- Proposals so that no language is left behind

# The State

| Dutch | Somali |
|---|---|
| 29 M speakers | 18 M speakers |
| 2 M Wiki articles | 5.5 K Wiki articles |
| Syntactically similar to English | Syntactically different from English, different word order, rare typological feature |
| 69 items on LDC and ELRA | 2 items on LDC and ELRA |
| Best online MT systems | Few online MT systems, far inferior quality |
| Steady and growing trend of research since 1980s | Just started in 2018/19 (zero-shot papers) |
| Large amount of resources | Lacking resources *and* attention from NLP community |
| Benefits from NLP breakthroughs | Benefits from NLP breakthroughs unknown |

# The State

- Most NLP systems are trained and tested on a handful of languages
  - Drawn from a few dominant language families
  - Typological echo chamber
    - Lots of linguistic phenomena have never been seen by NLP systems

# The State

- SOTA models rely on Deep NN systems that require a lot more **data** for training and a lot more **compute resources**
  - Create a bigger divide between high and low resource languages
  - But some multilingual models need only large unlabeled data (BERT, XLM, Roberta, GPT-3, …), but … # parameters: 175 billion!

# Findings

- The languages of the world can be broadly classified into 6 classes
    - According to the types and amount of resources (data) available
- Labeled data: LDC, ELRA
- Unlabeled data: Wikipedia

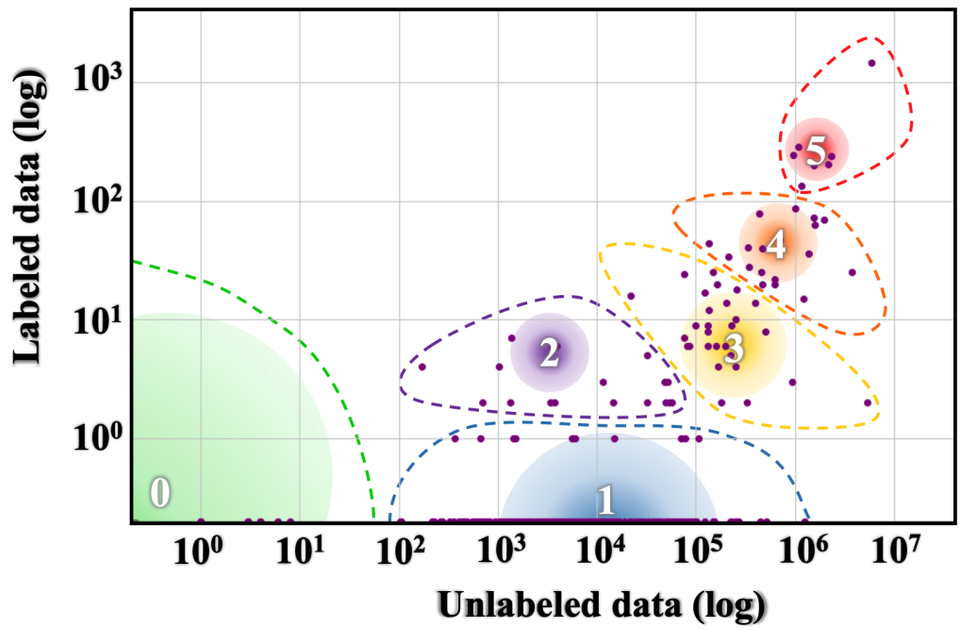



Figure 2: Language Resource Distribution: The size of the gradient circle represents the number of languages in the class. The color spectrum VIBGYOR, represents the total speaker population size from low to high. Bounding curves used to demonstrate covered points by that language class.

| Class | 5 Example Languages | #Langs | #Speakers | % of Total Langs |
|---|---|---|---|---|
| 0 | Dahalo, Warlpiri, Popoloca, Wallisian, Bora | 2191 | 1.2B | 88.38% |
| 1 | Cherokee, Fijian, Greenlandic, Bhojpuri, Navajo | 222 | 30M | 5.49% |
| 2 | Zulu, Konkani, Lao, Maltese, Irish | 19 | 5.7M | 0.36% |
| 3 | Indonesian, Ukranian, Cebuano, Afrikaans, Hebrew | 28 | 1.8B | 4.42% |
| 4 | Russian, Hungarian, Vietnamese, Dutch, Korean | 18 | 2.2B | 1.07% |
| 5 | English, Spanish, German, Japanese, French | 7 | 2.5B | 0.28% |

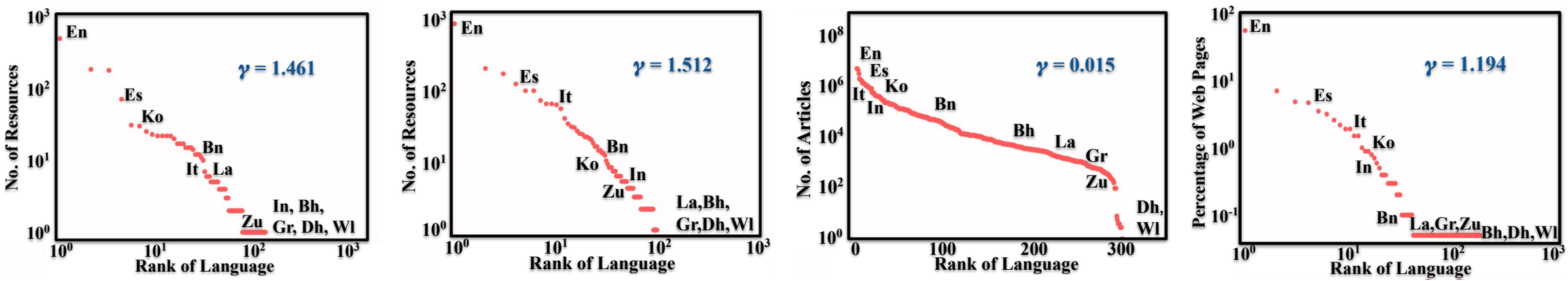
# Findings

| | |
|---|---|
| 0: the Left-behinds | - Ignored<br>- Virtually no unlabeled data |
| 1: the Scraping-bys | - Some amount of unlabeled data<br>- Virtually no labeled data |
| 2: the Hopefuls | - Small set of labeled data<br>- There are research and language support |
| 3: the Rising Stars | - Strong web presence, lots of unlabeled data<br>- Thriving cultural community online<br>- Insufficient efforts in labeled data collection |
| 4: the Underdogs | - Large amount of unlabeled data, comparable to winners<br>- Lesser amount of labeled data<br>- Dedicated NLP communities |
| 5: the Winners | - Dominant web presence<br>- Massive industrial and government investments<br>- Reaping benefits of SOTA NLP breakthroughs |

| Class | 5 Example Languages | #Langs | #Speakers | % of Total Langs |
|---|---|---|---|---|
| 0 | Dahalo, Warlpiri, Popoloca, Wallisian, Bora | 2191 | 1.2B | 88.38% |
| 1 | Cherokee, Fijian, Greenlandic, Bhojpuri, Navajo | 222 | 30M | 5.49% |
| 2 | Zulu, Konkani, Lao, Maltese, Irish | 19 | 5.7M | 0.36% |
| 3 | Indonesian, Ukranian, Cebuano, Afrikaans, Hebrew | 28 | 1.8B | 4.42% |
| 4 | Russian, Hungarian, Vietnamese, Dutch, Korean | 18 | 2.2B | 1.07% |
| 5 | English, Spanish, German, Japanese, French | 7 | 2.5B | 0.28% |

# Findings

| | |
|---|---|
| 0: the Left-behinds | - Ignored<br>- Virtually no unlabeled data |
| 1: the Scraping-bys | - Some amount of unlabeled data<br>- Virtually no labeled data |
| 2: the Hopefuls | - Small set of labeled data<br>- There are research and language support |
| 3: the Rising Stars | - Strong web presence, lots of unlabeled data<br>- Thriving cultural community online<br>- Insufficient efforts in labeled data collection |
| 4: the Underdogs | - Large amount of unlabeled data, comparable to winners<br>- Lesser amount of labeled data<br>- Dedicated NLP communities |
| 5: the Winners | - Dominant web presence<br>- Massive industrial and government investments<br>- Reaping benefits of SOTA NLP breakthroughs |




(a) LDC (b) LRE (c) Wikipedia (d) Web

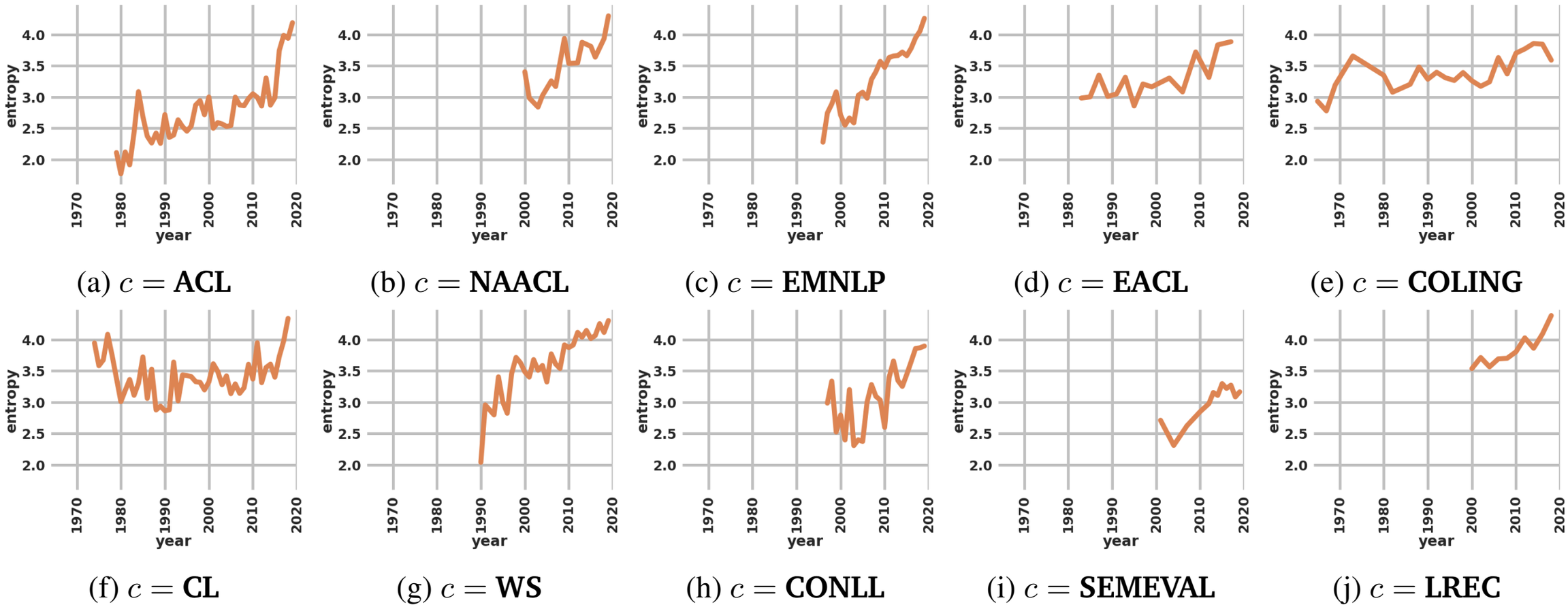
# Typology

- Structural and semantic properties of languages
- Skewed data availability means that certain typological features are under represented in NLP models
- NLP models might fail on tasks for certain languages
- Amharic has more types of “ignored” typological features; has significantly higher similarity search errors than Arabic

# Conference Inclusion

- NLP **conferences** have high impact on how language **resources** and **technologies** are constructed and have the **potential to attract funds** to a particular technology
  - Has usage of a small set of resource-rich languages led to a disparity in language research?
- Dataset: ACL Anthology Corpus, Semantic Scholar's API
  - 11 conferences: ACL, NAACL, EMNLP, EACL, COLING, LREC, CONLL, WS, SEMEVAL, TACL, CL
  - Measure distribution of language use (its entropy) in the conferences

# Conference Inclusion




(a) $c =$ **ACL** (b) $c =$ **NAACL** (c) $c =$ **EMNLP** (d) $c =$ **EACL** (e) $c =$ **COLING**

(f) $c =$ **CL** (g) $c =$ **WS** (h) $c =$ **CONLL** (i) $c =$ **SEMEVAL** (j) $c =$ **LREC**

Figure 4: Language occurrence entropy over the years for different conferences ($\{S\}^{c,y}$).

WS and LREC are more **inclusive**

**Spikes in 2010s** for other conferences due to cross-lingual techniques

# Conference Inclusion

Table 4 shows inverse mean reciprocal ranks of each category for a conference. The smaller the inverse MRR value, the more inclusive that conference is to that language class.

Inverse MRR: **smaller** means the conference is **more inclusive** to that language

| Conf / Class | 0 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| **ACL** | 725 | 372 | 157 | 63 | 20 | 3 |
| **CL** | 647 | 401 | 175 | 76 | 27 | 3 |
| **COLING** | 670 | 462 | 185 | 74 | 21 | 2 |
| **CONLL** | 836 | 576 | 224 | 64 | 16 | 3 |
| **EACL** | 839 | 514 | 195 | 63 | 15 | 3 |
| **EMNLP** | 698 | 367 | 172 | 67 | 19 | 3 |
| **LREC** | 811 | 261 | 104 | 45 | 13 | 2 |
| **NAACL** | 754 | 365 | 136 | 63 | 18 | 3 |
| **SEMEVAL** | 730 | 983 | 296 | 121 | 19 | 3 |
| **TACL** | 974 | 400 | 180 | 50 | 15 | 3 |
| **WS** | 667 | 293 | 133 | 59 | 15 | 3 |

WS and LREC are more **inclusive**
Dip in ranks is more forgiving

# Entity Embedding Analysis

- Jointly learn the representations of conferences, authors, and languages (i.e., entities)

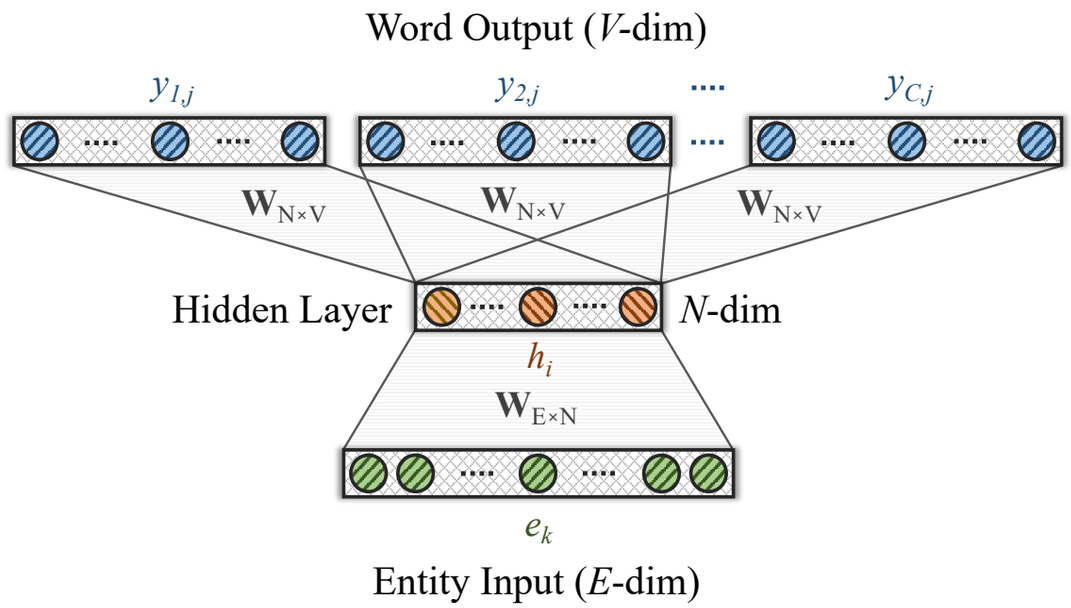



Figure 5: Model architecture to learn entity embeddings. $W_{E\times N}$ is the weight matrix from input layer (entity layer) to the hidden layer, and $W_{N\times V}$ is the weight matrix for the hidden layer to output layer computation. At the end of training, $W_{E\times N}$ is the matrix containing embeddings of entities and $W_{N\times V}$ is the matrix containing the embeddings of words.

E=number of entities
V=number of words

$W_{ExN}$=entity embedding
$W_{VxN}$=word embedding

# Entity Embedding Analysis

theoretical

data-driven

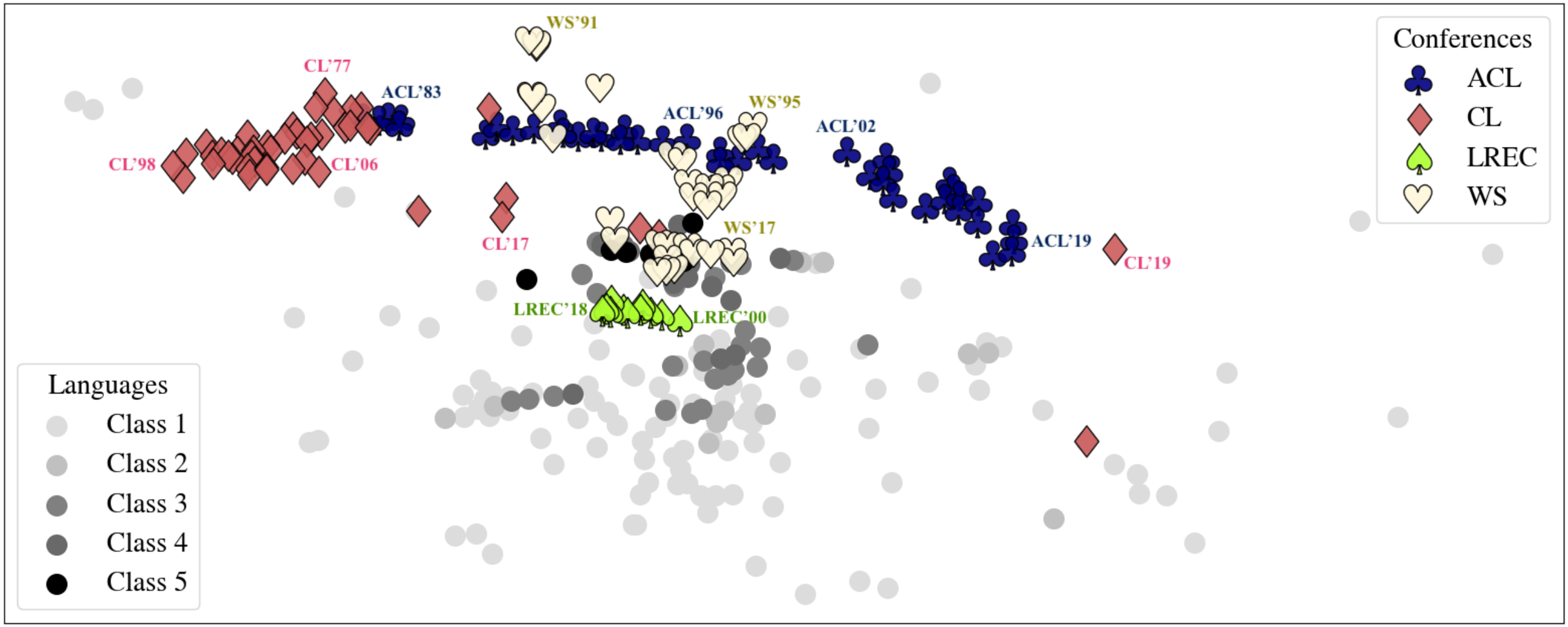



Figure 6: t-SNE visualization of the learnt conference and language embeddings.

LREC and WS closer to language embeddings

# Entity Embedding Analysis

- Language-Author-Language closest neighbors

The **higher** the number, the more **focused** is the research group

There's hope that there is a focused research community working on low-resource languages

**Exception**: Javanese, Igbo, Burmese, …

| Class | MRR(5) | MRR(10) | MRR(15) | MRR(20) |
|---|---|---|---|---|
| 0 | 0.72281 | 0.69146 | 0.63852 | 0.57441 |
| 1 | 0.57210 | 0.52585 | 0.45354 | 0.40904 |
| 2 | 0.47039 | 0.45265 | 0.41521 | 0.38157 |
| 3 | 0.59838 | 0.52670 | 0.45131 | 0.42899 |
| 4 | 0.56016 | 0.47795 | 0.51199 | 0.50681 |
| 5 | 0.56548 | 0.51471 | 0.54326 | 0.47619 |

Table 5: Language-Author-Language MRR on Taxonomy Classes. MRR(**K**) considers the closest **K** authors.

# Conclusion

- Resource comparison, entropy calculations, and embeddings show disparities in language
    - LREC and WS most inclusive across different classes of languages
    - Newer conferences more inclusive
- Typological features are not covered equally by NLP models
    - There are features that are covered by low-resource but not high-resource languages
    - Throw doubts into how models relying on transfer learning actually works for languages with ‘ignored’ features

# Still so much to do!

- More data -> better performance
- Less data ?
  - new approach?
  - look for more data?
- In this class: learn to scrape, process, and analyze human language data using statistical and machine learning approaches.