ITCS 4111/5111: Intro to Natural Language Processing

Introduction

Razvan C. Bunescu

Department of Computer Science @ CCI

razvan.bunescu@uncc.edu

What is Natural Language Processing?

- **Natural Language Processing (NLP)** = developing computer systems that can *process*, *understand*, or *communicate* in natural language (*text* or *speech*):
  - **Natural Languages**: English, Turkish, Japanese, Latin, Hawaiian Creole, Esperanto, American Sign Language, …
    - Music?
  - **Formal Languages**: C++, Java, Python, XML, OWL, Predicate Calculus, Lambda Calculus, …
  - **Natural Languages are significantly more difficult to process than Artificial Languages!**

- What about **Computational Linguistics (CL)**?
  - Computational Linguistics is focused on the *study of language*, using computational tools.
  - NLP is focused on solving language tasks such as *machine translation*, *information extraction*, ***question answering***, *taking instructions*, *holding conversations*, …

NLP Application: Question Answering

- Input:
  - A question:
    - *What is the meaning of life?*
  - A large collection of text documents:
    - *all books from UNCC Library.*
- Output:
  - An answer, or list of answers.
    - Found by '*mining*' the documents in the collection.

What is the meaning of life?

Tomorrow, and tomorrow, and tomorrow,
Creeps in this petty pace from day to day,
To the last syllable of recorded time;
And all our yesterdays have lighted fools
The way to dusty death. Out, out, brief candle!
Life's but a walking shadow, a poor player
That struts and frets his hour upon the stage
And then is heard no more. It is a tale
Told by an idiot, full of sound and fury
Signifying nothing.

Skakespeare's Macbeth (Act 5, Scene 5, lines 17-28)

NLP Application: Question Answering

- Input:
  - A question:
    - *What is the meaning of life?*
  - A large collection of text documents:
    - *all books from UNCC Library.*

- Output:
  - An answer, or list of answers.
    - Found by '*mining*' the documents in the collection.

What is the meaning of life?

Tomorrow, and tomorrow, and tomorrow,
Creeps in this petty pace from day to day,
To the last syllable of recorded time;
And all our yesterdays have lighted fools
The way to dusty death. Out, out, brief candle!
Life's but a walking shadow, a poor player
That struts and frets his hour upon the stage
And then is heard no more. It is a tale
Told by an idiot, full of sound and fury
*Signifying **nothing**.*

Skakespeare's Macbeth (Act 5, Scene 5, lines 17-28)

NLP Application: Question Answering

- Input:
  - A question:
    - *What is the meaning of life?*

Words and Tokens:
{*Life*, *life*, *?*, *'s*, …} are tokens.

- Output:
  - An answer, or list of answers.
    - Found by '*mining*' the documents in the collection.

What is the meaning of **life**?

Tomorrow, and tomorrow, and tomorrow,
Creeps in this petty pace from day to day,
To the last syllable of recorded time;
And all our yesterdays have lighted fools
The way to dusty death. Out, out, brief candle!
***Life**'s but a walking shadow, a poor player*
That struts and frets his hour upon the stage
And then is heard no more. It is a tale
Told by an idiot, full of sound and fury
Signifying nothing.

Skakespeare's Macbeth (Act 5, Scene 5, lines 17-28)

NLP Application: Question Answering

- Input:
  - A question:

Try simple pattern matching:
"the meaning of life is <?>"

ts:

  - *all books from UNCC Library.*

- Output:
  - An answer, or list of answers.
    - Found by '*mining*' the documents in the collection.

What is the meaning of life?

Tomorrow, and tomorrow, and tomorrow,
Creeps in this petty pace from day to day,
To the last syllable of recorded time;
And all our yesterdays have lighted fools
The way to dusty death. Out, out, brief candle!
Life's but a walking shadow, a poor player
That struts and frets his hour upon the stage
And then is heard no more. It is a tale
Told by an idiot, full of sound and fury
Signifying nothing.

Skakespeare's Macbeth (Act 5, Scene 5, lines 17-28)

NLP Application: Question Answering

What is the **meaning** of life?

- Input:
  - A question:
    - *What is the meaning of life?*

Word Meaning:
meaning ≃ *signifying*

- Output:
  - An answer, or list of answers.
    - Found by '*mining*' the documents in the collection.

Tomorrow, and tomorrow, and tomorrow,
Creeps in this petty pace from day to day,
To the last syllable of recorded time;
And all our yesterdays have lighted fools
The way to dusty death. Out, out, brief candle!
Life's but a walking shadow, a poor player
That struts and frets his hour upon the stage
And then is heard no more. It is a tale
Told by an idiot, full of sound and fury
***Signifying** nothing.*

Skakespeare's Macbeth (Act 5, Scene 5, lines 17-28)

NLP Application: Question Answering

- Input:
  - A question:
    - *What is the meaning of life?*
  - A large collection of text documents:
    - *all books from UNCC Library.*
- Output:
  - An answer, or list of answers.
    - Found by '*mining*' the documents in the collection.

What is the **meaning** of life?

Tomorrow, and tomorrow, and tomorrow,
Creeps in this petty pace from day to day,
To the last syllable of recorded time;
And all our yesterdays have lighted fools
The way to dusty death. Out, out, brief candle!
Life's but a walking shadow, a poor player
That struts and frets his hour upon the stage
And then is heard no more. It is a tale
Told by an idiot, full of sound and fury
***Signifying** nothing.*

Skakespeare's Macbeth (Act 5, Scene 5, lines 17-28)

NLP Application: Question Answering

- Input:
  - A question:
    - *What is the meaning of life?*

Coreference:
{*Life*, *it*, *tale*} are coreferent.

- Output:
  - An answer, or list of answers.
    - Found by '*mining*' the documents in the collection.

What is the **meaning** of **life**?

Tomorrow, and tomorrow, and tomorrow,
Creeps in this petty pace from day to day,
To the last syllable of recorded time;
And all our yesterdays have lighted fools
The way to dusty death. Out, out, brief candle!
***Life**'s but a walking shadow, a poor player*
That struts and frets his hour upon the stage
*And then is heard no more. **It** is a **tale***
Told by an idiot, full of sound and fury
***Signifying** nothing.*

Skakespeare's Macbeth (Act 5, Scene 5, lines 17-28)

NLP Application: Question Answering

- Input:
  - A question:
    - *What is the meaning of life?*

Syntactic Dependencies:
- *tale* is Subject of *signifying*.
- *nothing* is Object of *signifying*.

- Output:
  - An answer, or list of answers.
    - Found by '*mining*' the documents in the collection.

What is the **meaning** of **life**?

Tomorrow, and tomorrow, and tomorrow,
Creeps in this petty pace from day to day,
To the last syllable of recorded time;
And all our yesterdays have lighted fools
The way to dusty death. Out, out, brief candle!
***Life**'s but a walking shadow, a poor player*
That struts and frets his hour upon the stage
*And then is heard no more. **It** is a **tale***
Told by an idiot, full of sound and fury
***Signifying nothing**.*

Skakespeare's Macbeth (Act 5, Scene 5, lines 17-28)

NLP Application: Question Answering

Fundamental NLP tasks:

- **Tokenization**.
- **Syntactic Analysis**:
  - Part of Speech Tagging.
  - **Dependency Parsing**.
- **Word Meaning representations**.
- **Coreference Resolution**.
- Semantic Role Labeling.
- Semantic Parsing.

What is the **meaning** of **life**?

Tomorrow, and tomorrow, and tomorrow,
Creeps in this petty pace from day to day,
To the last syllable of recorded time;
And all our yesterdays have lighted fools
The way to dusty death. Out, out, brief candle!
Life's but a walking shadow, a poor player*
That struts and frets his hour upon the stage
*And then is heard no more. **It** is a **tale***
Told by an idiot, full of sound and fury
***Signifying nothing**.*

Skakespeare's Macbeth (Act 5, Scene 5, lines 17-28)

Fundamental NLP Tasks Identify **Linguistic Structures**

- Tokenization
- Morphological Analysis
- Part of Speech Tagging
- Syntactic Parsing
- Word Meaning
- Semantic Role Labeling
- Semantic Parsing
- Anaphora/Coreference Resolution
- Named Entity Recognition
- Relation Extraction
- Named Entity Linking

} Information Extraction (Anaphora/Coreference Resolution, Named Entity Recognition, Relation Extraction, Named Entity Linking)

Tokenization

- **Tokenization** = segmenting text into words and sentences.
  - A crucial first step in most text processing applications.
  - Language Models use ***subword*** tokenization.

- Whitespace indicative of word boundaries?
  - Yes: English, French, Spanish, …
  - No: Chinese, Japanese, Thai, …

- Whitespace is not enough:
  - ‘What’re you? Crazy?’ said Sadowsky. ‘I can’t afford to do that.’

Whitspace ⇒ ‘*what’re_you?_crazy?_*said_*Sadowsky._‘I_can’t_afford_to_do_that.*

Correct ⇒ ‘*_what_’re_you_?_crazy_?_Sadowsky_._‘_I_can_’t_ afford_to_do_that_.*

Word Segmentation

- In English, characters other than whitespace can be used to separate words:
  - , ; . - : ( )"
- But punctuation often occurs inside words:
  - m.p.h., Ph.D., AT&T, 01/02/06, google.com, 62.5
  - Homework: design regular expressions to match constructions where punctuation does not split:
    - *acronyms*, *dates*, *web addresses*, *numbers*, etc.
    - https://docs.python.org/3/howto/regex.html
- Expansion of clitic constructions:
  - he's happy ⇒ he is happy
  - Need ambiguity resolution between clitic construction, possessive markers, quotative markers:
    - he's happy vs. the book's cover vs. 'what are you? crazy?'

Sentence Segmentation

- Generally based on punctuation marks: **? ! .**
  - Periods are ambiguous, as sentence boundary markers and abbreviation/acronym markers:
    - *Mr., Inc., m.p.h.*
  - Sometimes they mark both:
    - SAN FRANCISCO (MarketWatch) – Technology stocks were mostly in positive territory on Monday, powered by gains in shares of Microsoft Corp. and **IBM Corp.**

- Tokenization approaches:
  - Regular Expressions.
  - Machine Learning (state of the art).

Morphological Analysis

- **Morphology** = the field of linguistics that studies the internal structure of words.
  - **Morpheme** is the smallest linguistic unit that has semantic meaning:
    - **stems**: “carry”, “depend”, “Google”, “lock”
    - **affixes**: “pre”, “ed”, “ly”, “s”
      - *Inflectional*: neighbor-s
      - *Derivational*: neighbor-hood

- **Morphological analysis** = segmenting words into morphemes:
  - carried ⇒ carry + ed (past tense)
  - independently ⇒ in + (depend + ent) + ly
  - Googlers ⇒ (Google + er) + s (plural)
  - unlockable ⇒ un + (lock + able) ? (un + lock) + able ?

Part of Speech (POS) Tagging

- Annotate each word in a sentence with its POS:
  - nouns, verbs, adjectives, adverbs, pronouns, prepositions, …

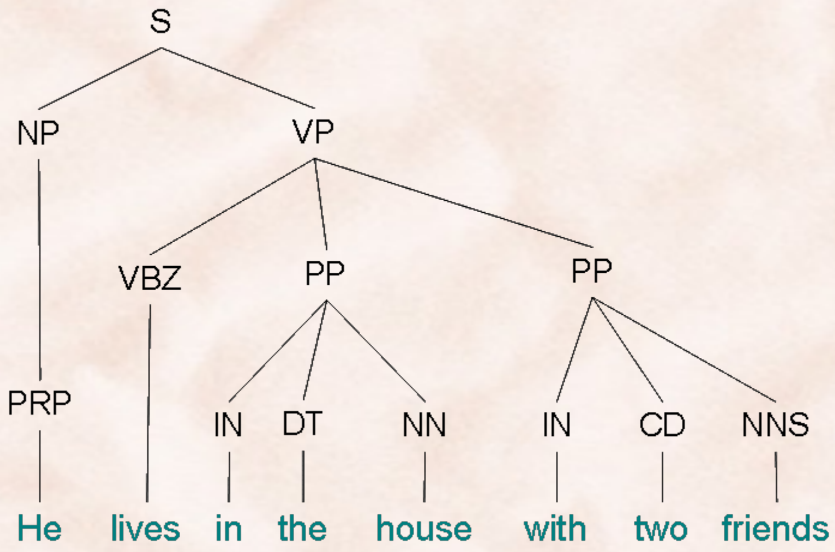
PRP	VBD	TO	VB	TO	DT	NN	IN	NN	VBD	VBG
They	used	to	object	to	the	use	of	object	oriented	programming
			obJECT					OBject		

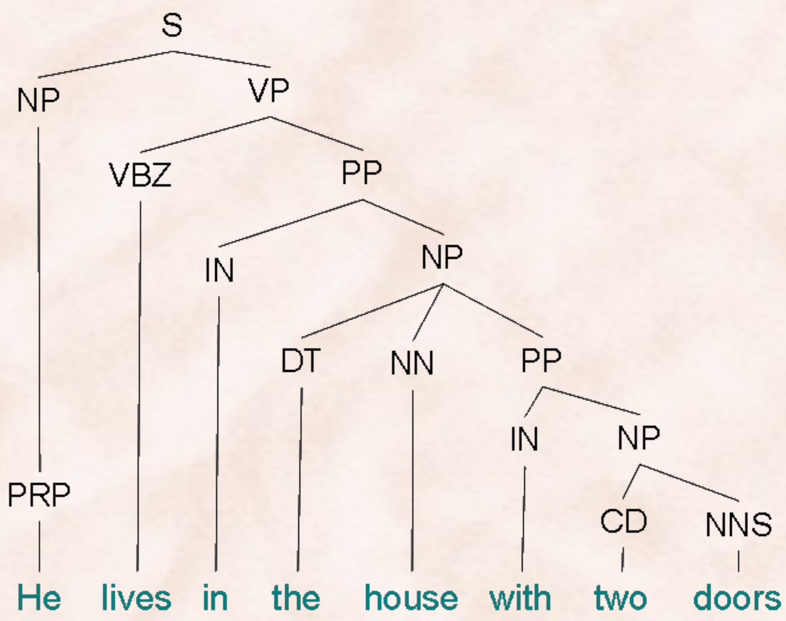
- Useful for many NLP tasks downstream:
  - speech recognition and synthesis, syntactic parsing, word sense disambiguation, information retrieval, …

- Superseeded in many tasks by (contextualized) ***word embeddings***.

Syntax: Constituency Parsing

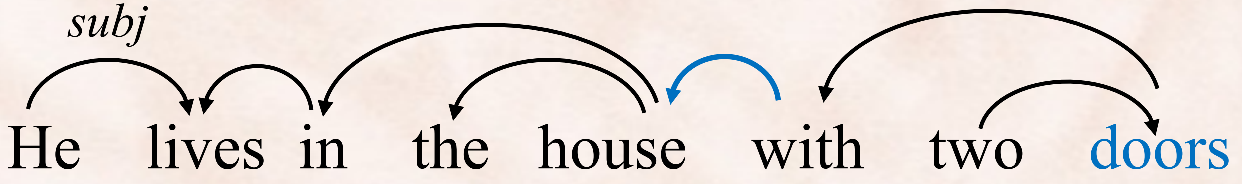
- Compute the *phrase structure* of a sentence:





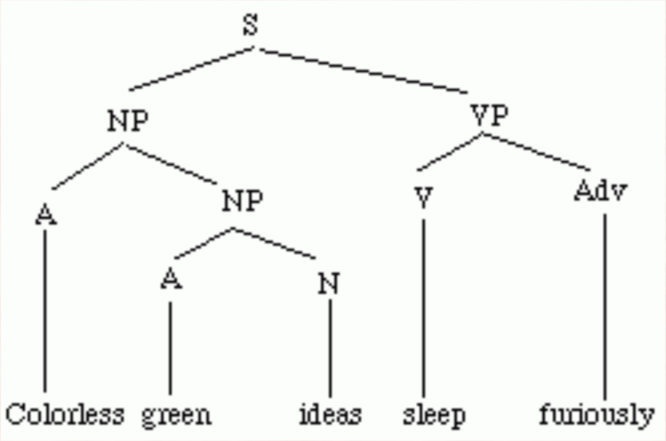
- Corresponding syntactic *dependency structure*:
  - Try https://corenlp.run/
    - This is an automatic analysis, so it is not always correct (see second sentence)

Syntax: Dependency Parsing



Syntax vs. Semantics

- **The to runs the cat dog after.**
  - – Syntactically incorrect.
    - Cannot build phrase structure (parse tree).
  - – Semantically meaningless.
    - Cannot have meaning if sentence is not well formed syntactically.
- **Colorless green ideas sleep furiously**:
  - – Syntactically correct.
  - – Semantically meaningless (literal reading).



- **The fat cat sat on the mat**:
  - – Syntactically correct.
  - – Semantically meaningful: *cat*(C) ∧ *mat*(M) ∧ *satOn*(C, M).
  - – Pragmaticaly?

Colorless green ideas sleep furiously: Figurative Reading

It can only be the thought of verdure to come, which prompts us in the autumn to buy these dormant white lumps of vegetable matter covered by a brown papery skin, and lovingly to plant them and care for them. It is a marvel to me that under this cover they are labouring unseen at such a rate within to give us the sudden awesome beauty of spring flowering bulbs. While winter reigns the earth reposes but these ***colourless green ideas sleep furiously****.*

[by C. M. Street]

http://linguistlist.org/issues/2/2-457.html#2

Word Meaning

- Words in natural language may have multiple meanings:
  - he cashed a check at the bank
  - he sat on the bank of the river and watched the currents
  - they built a large plant to manufacture automobiles
  - chlorophyll is generally present in plant leaves

- Use lexical resources such as WordNet that map words to their meanings:
  - http://wordnetweb.princeton.edu/perl/webwn

- Identifying the meaning of a word is useful for:
  - machine translation, information retrieval, question answering, text classification, …

- Nowadays superseded in many tasks by (contextualized) ***word embeddings***.

Semantic Roles

- For each clause, determine the semantic role played by each noun phrase that is an argument to the verb:

  agent patient source destination instrument

  - John drove Mary from Charlotte to Asheville in his Honda Accord.
  - The hammer broke the window.

- Also referred to a "case role analysis," "thematic analysis," and "shallow semantic parsing".

- FrameNet and PropBank provides vocabulary of semantics roles:
  - https://framenet.icsi.berkeley.edu/fndrupal/frameIndex
  - http://verbs.colorado.edu/~mpalmer/projects/ace.html
  - https://babelnet.org/

Logic Forms (Semantic Parsing)

- Map natural language sentences to a formal semantic representation (*logic form*).
- Text to SQL, for interaction with DBs in natural language:
  - *List all song names by singers age above the average singer age.*
  - SELECT song_name FROM singers WHERE age > (SELECT avg(age) FROM singers)
- In RoboCup, map coaching advice to Clang:
  - *If the ball is in our penalty area, all our players except player 4 should stay in our half.*
  - ((bpos (penalty-area our)) (do (player-except our {4}) (pos (half our))))
- In GeoQuery, map sentences to Prolog queries:
  - *How many states does the Mississippi run through?*
  - answer(A, count(B, (state(B), const(C, riverid(mississippi)), traverse(C, B)), A))

Logic Forms (Semantic Parsing)

- Automatic generation of code, e.g. for [cards in Trading Card Games](TCGs):


```
name: [
  'D', 'i', 'r', 'e', ' ',
  'W', 'o', 'l', 'f', ' ',
  'A', 'l', 'p', 'h', 'a']
cost: ['2']
type: ['Minion']
rarity: ['Common']
race: ['Beast']
class: ['Neutral']
description: [
  'Adjacent', 'minions', 'have',
  '+', '1', 'Attack', '.']
health: ['2']
attack: ['2']
durability: ['-1']
```

```python
class DireWolfAlpha(MinionCard):
  def __init__(self):
    super().__init__(
      "Dire Wolf Alpha", 2, CHARACTER_CLASS.ALL,
      CARD_RARITY.COMMON, minion_type=MINION_TYPE.BEAST)
  def create_minion(self, player):
    return Minion(2, 2, auras=[
      Aura(ChangeAttack(1), MinionSelector(Adjacent()))
    ])
```


```python
class ManaWyrm(MinionCard):
  def __init__(self):
    super().__init__(
      'Mana Wyrm', 1,
      CHARACTER_CLASS.MAGE,
      CARD_RARITY.COMMON)
  def create_minion(self, player):
    return Minion(
      1, 3, effects=[
        Effect(
          SpellCast(),
          ActionTag(
            Give(ChangeAttack(1)),
                 SelfSelector()))
      ])
```

Coreference

- Determine which noun phrases refer to the same discourse entity.

> Originally from Hawaii, Obama is a graduate of Columbia University and Harvard Law School, where he was the president of the Harvard Law Review. He was a community organizer in Chicago before earning his law degree.

> Originally from Hawaii, Obama is a graduate of Columbia University and Harvard Law School, where he was the president of the Harvard Law Review. He was a community organizer in Chicago before earning his law degree.

The Curse of Ambiguity

- Computational Linguists are obsessed by ambiguity in NL:
  - unlike compiler writers.

- Ambiguity happens at all basic levels of natural language processing.

- Find at least 5 meanings of the following sentence:
  - I made her duck.

Ambiguity: “I made her duck”

1) I cooked waterfowl for her benefit (to eat).
2) I cooked waterfowl belonging to her.
3) I created the (plaster?) duck she owns.
4) I caused her to quickly lower her head or body.
5) I waved my magic wand and turned her into undifferentiated waterfowl.

Ambiguity: "I made her duck"

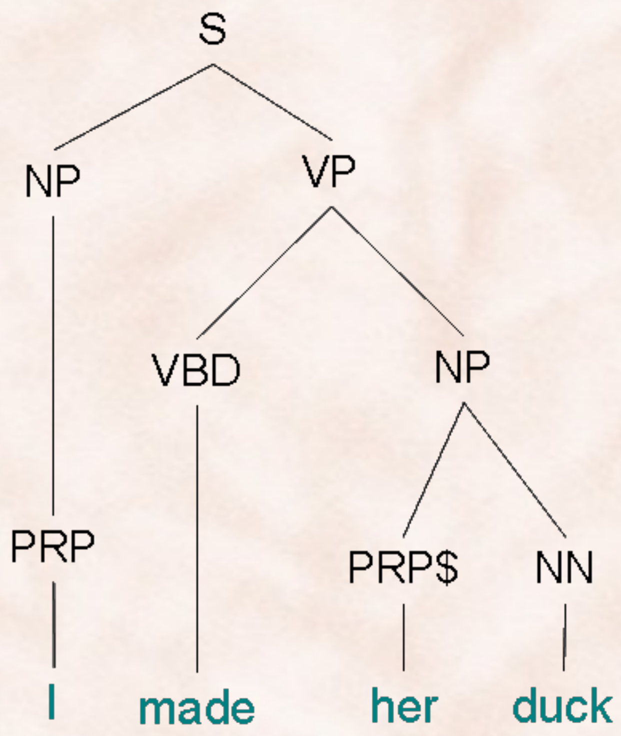
- **Part-of-Speech**: "duck" can be a N or V:
  - V: I caused her to quickly lower her head or body
  - N: I cooked waterfowl for her benefit (to eat).

- **Syntactic:** "her" can be a *possessive* ("of her") or *dative* ("for her") or *accusative* pronoun:
  - Possessive: I cooked waterfowl belonging to her.
  - Dative: I cooked waterfowl for her benefit (to eat).
  - Accusative: I waved my magic wand and turned her into waterfowl.

- **Semantic:** "make" can mean "create" or "cook":
  - Create: I made the (plaster) duck statue she owns
  - Cook: I cooked waterfowl belonging to her.

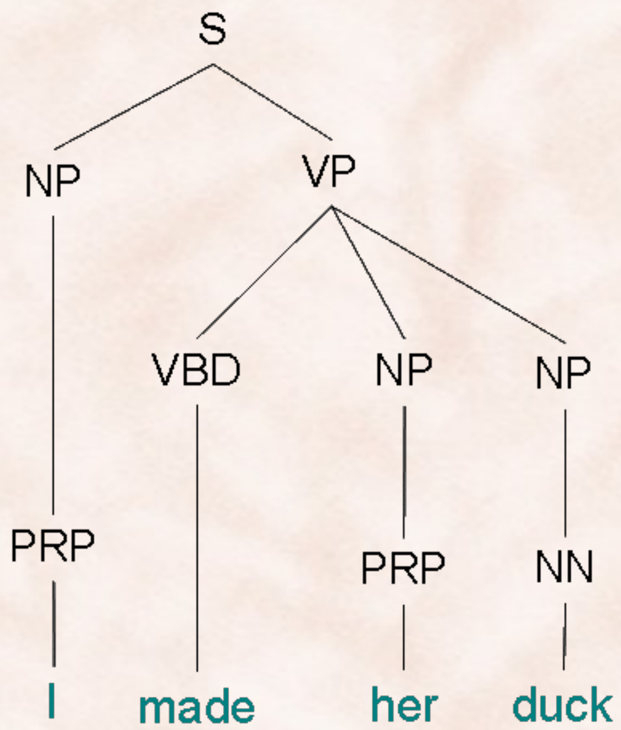
Ambiguity: “I made her duck”

- **Syntactic Parsing**:
  - **Make can be Transitive (verb has a noun direct object):**
    - I cooked [waterfowl belonging to her]



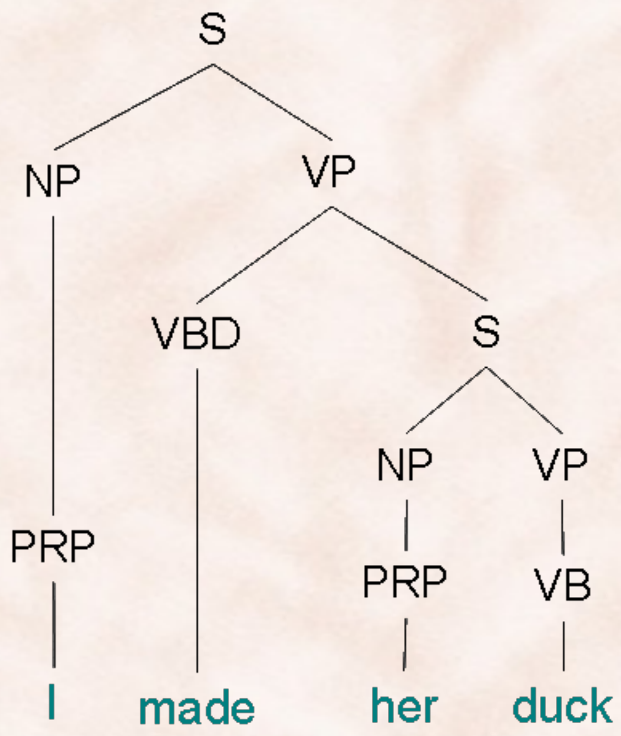
Ambiguity: “I made her duck”

- **Syntactic Parsing**:
  - **Make can be Ditransitive (verb has 2 noun objects):**
    - I made [her] (into) [undifferentiated waterfowl]



Ambiguity: “I made her duck”

- **Syntactic Parsing**:
  - **Make can be Action-transitive:**
    - I caused [her] [to move her body]



Ambiguity: “I made her duck”

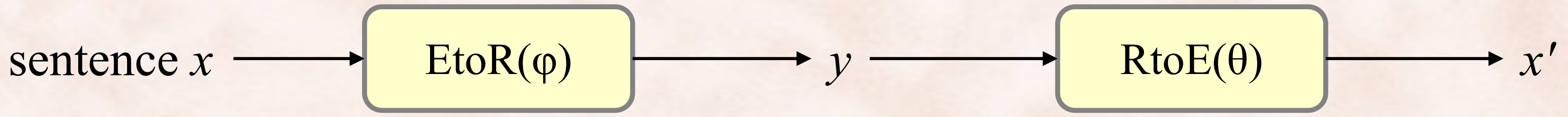
- **Speech Recognition**:
  - I mate or duck
  - I’m eight or duck
  - Eye maid; her duck
  - Aye mate, her duck
  - I maid her duck
  - I’m aid her duck
  - I mate her duck
  - I’m ate her duck
  - I’m ate or duck

Ambiguity and Machine Translation

- English ⇒ Italian:
  - Mary **plays** the piano ⇒ Maria **suona** il pianoforte.
  - Mary **plays** with her cat ⇒ Maria **gioca** con il suo gatto.
- "Lost in translation" jokes from supposedly early MT system output (English ⇒ Russian ⇒ English):
  - "The spirit is willing, but the flesh is weak" => Russian => English:

    ⇒ The vodka is good, but the meat is spoiled.
  - "Out of sight, out of mind" => Russian => English:

    ⇒ Invisible idiot.
  - Modern MT systems use **backtranslation** as a constraint during ML training:



    - train parameters φ and θ such that $x \cong x'$

Ambiguity and Discourse: Coreference Resolution

- Winograd schemas:
  1. The **dog** chased the **cat**, which ran up a **tree**. **It** waited at the [top / bottom].
  2. The city **councilmen** refused the **demonstrators** a permit because **they** [feared / advocated] violence.
  3. The **trophy** doesn't fit into the brown **suitcase** because **it**'s too [small / large].
  4. **Paul** tried to call **George** on the phone, but **he** wasn't [successful / available].
  5. The **man** couldn't lift his **son** because **he** was so [weak / heavy].
  6. **Ann** asked **Mary** what time the library closes, [but / because] **she** had forgotten.

https://www.cs.nyu.edu/davise/papers/WS.html

Modality and Ambiguity: What does Nancy want?

- **"Nancy wants to mary an analytic philosopher"** [Eco, "Kant and the Platypus", 2000]
- Semantic interpretations:
  - [*de re*]: Nancy wants to marry a determined individual X, who is an analytic philosopher.

    $\exists x \Box A x$

  - [*de dicto*]: Nancy wants to marry anybody, as long as he is an analytic philosopher.

    $\Box \exists x A x$

- Pragmatic Interpretations (speaker's intentions):
  - Nancy wants to marry a determined individual, an analytic philosopher: she knows who he is, but the speaker doesn't, because she hasn't told him the name.
  - Nancy wants to marry a determined individual X, an analytic philosopher: she has also given the speaker the name and introduced them to each other, but out of discretion the speaker has thought it more fitting to avoid going into details.
  - …

Ambiguity is Pervasive in Natural Language

- CL and NLP are obsessed with ambiguity:
  - unlike compiler writers.
- Ambiguity happens at all basic levels of language processing.
- [Pros] Allows for significant compression of utterances:
  - people use context and knowledge about the world to disambiguate.
- [Cons] Very challenging for NLP!

Knowledge Involved in Resolving Ambiguity

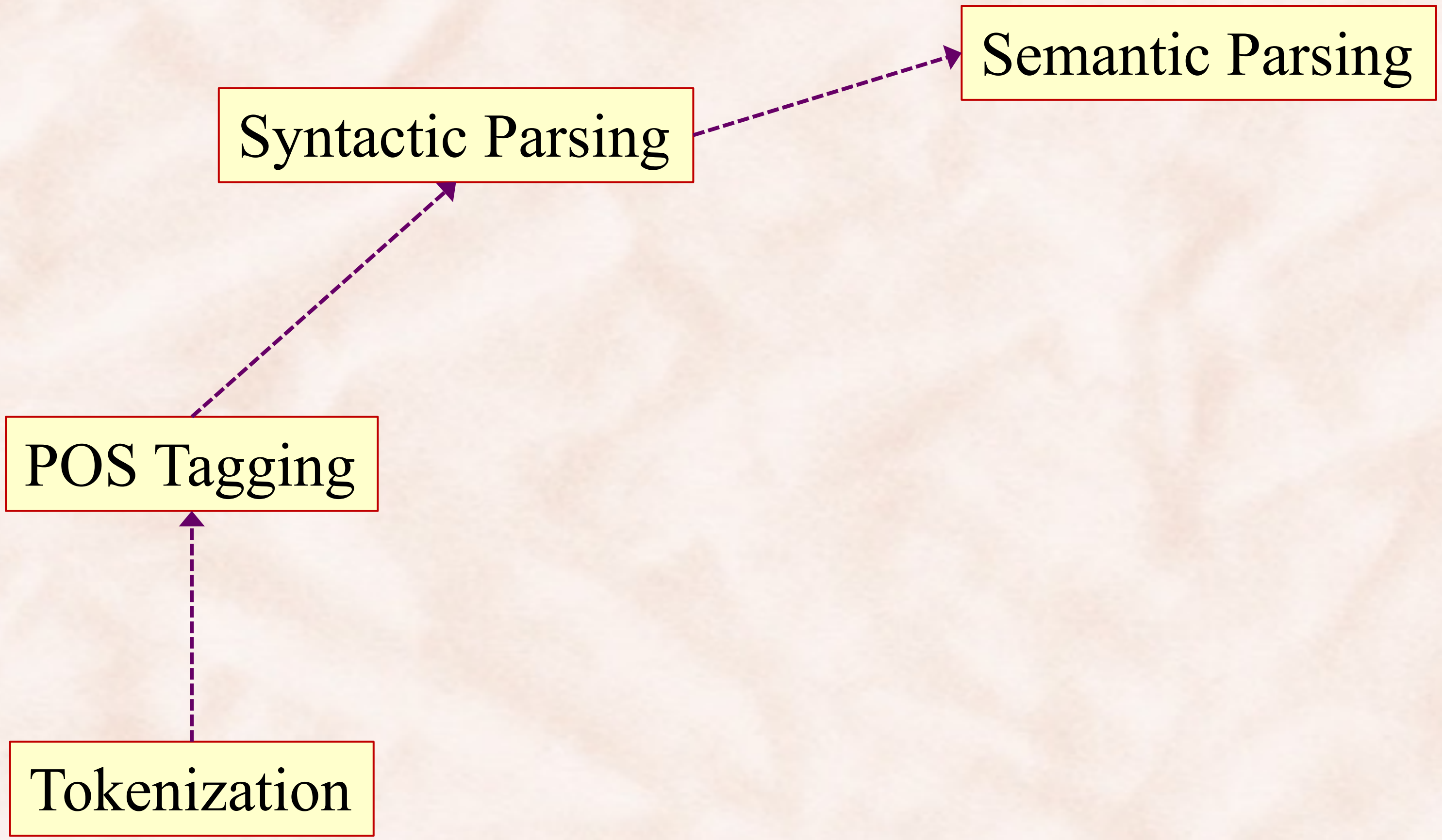
- **Syntax**:
  - An agent is typically the subject of the verb (SRL).
- **Semantics**:
  - John and Mary are names of people.
  - Columbus and Athens are city names.
- **Pragmatics**:
  - If she is hungry and she is not vegetarian, it is likely she will enjoy cooked duck.
- **Word knowledge**:
  - Houses have a (variable number of) doors.
  - An individual may live with other people (friends) in the same house.

Two Major Approaches to NLP

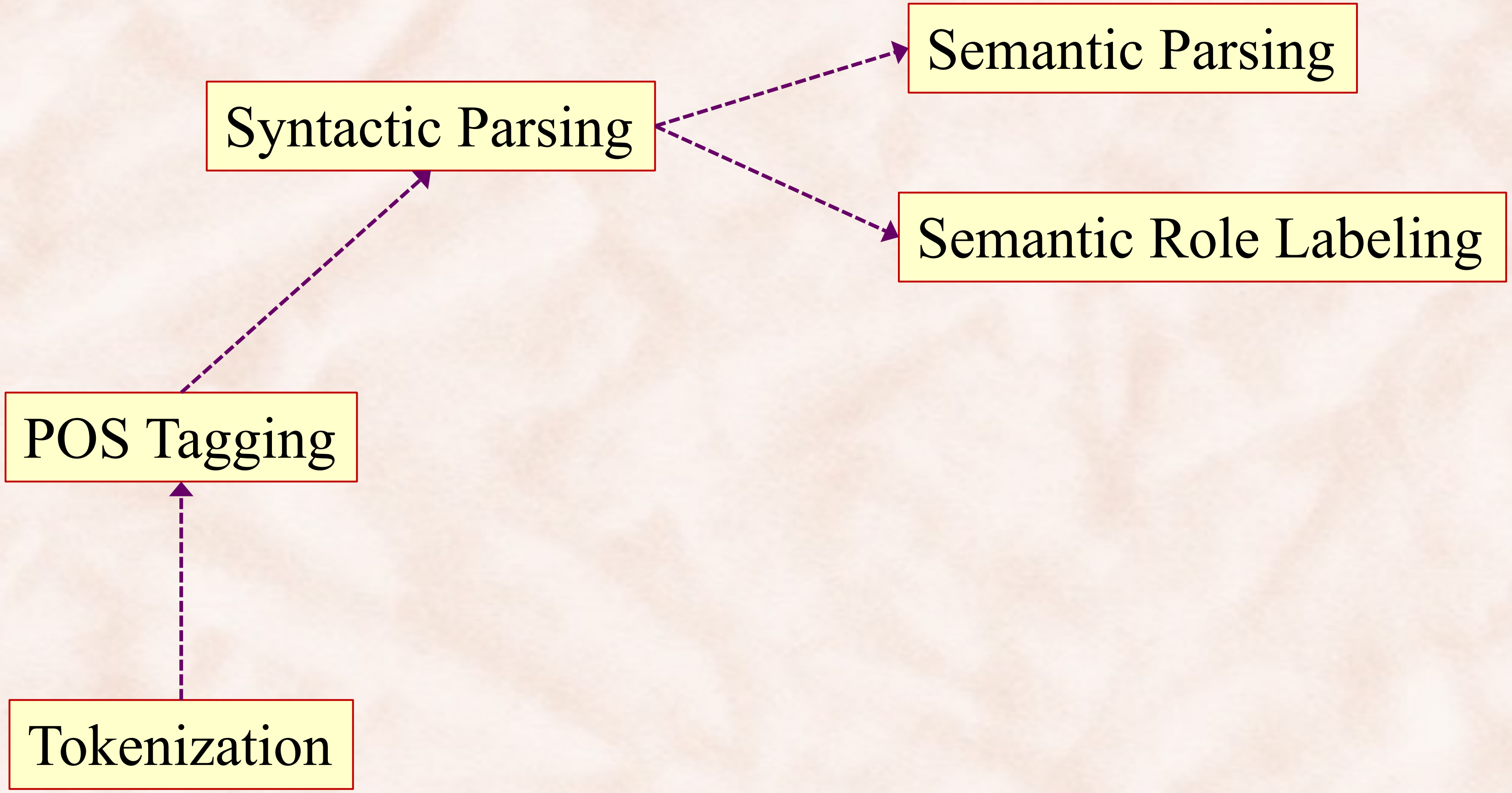
1. **NLP Pipelines**:
   - Transforming text into a stack of general-purpose linguistic structures:
     - From subword units called *morphemes*, to word-level *parts-of-speech*, to *tree-structured representations* of grammar, and beyond, to *logic-based representations* of meaning.
   - These general-purpose structures should then be able to support any desired application.
   - Should generalize better across applications, scenarios, and languages:
     - Linguistic structure seems to be particularly important when training data is limited.

2. **End-to-End** (*learning from scratch*):
   - Started by [Collobert et al., 2011], transform raw text into any desired output structure:
     - a summary, a database, a translation, …
   - **Language Models**: learning to predict words in context or just the next words seems sufficient for learning all levels of linguistic knowledge, as well as world knowledge.

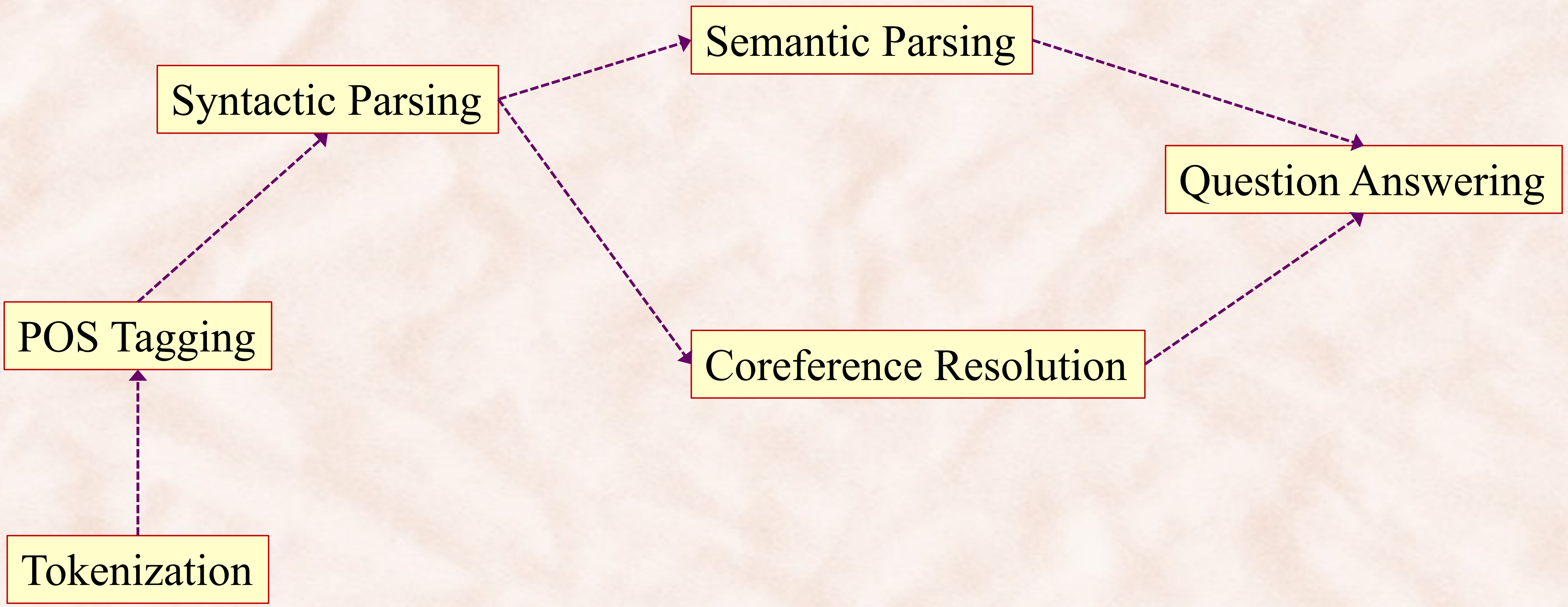
NLP Pipeline Example



NLP Pipeline Example



NLP Pipeline Example



End-to-End: Semantic Role Labeling

“*End-to-end Learning of Semantic Roles using RNNs* [Zhou & Xu, ACL 2015]

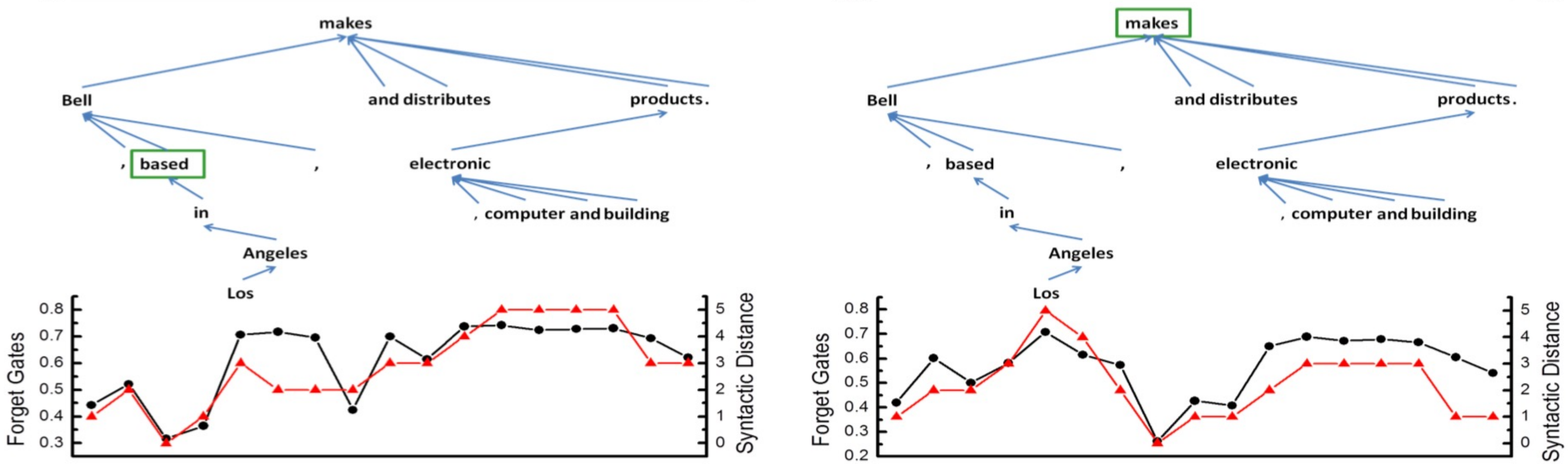


Figure 6: Forget gates value *vs*. Syntactic distance on four example sentences. Top: dependency parsing tree from gold tag. Green square word: predicate word. Bottom black solid lines: forget gates value at each time step. Bottom red empty square lines: gold syntactic distance between the current argument and predicate.

End-to-End: Semantic Parsing [Chen & Bunescu, NAACL'19]

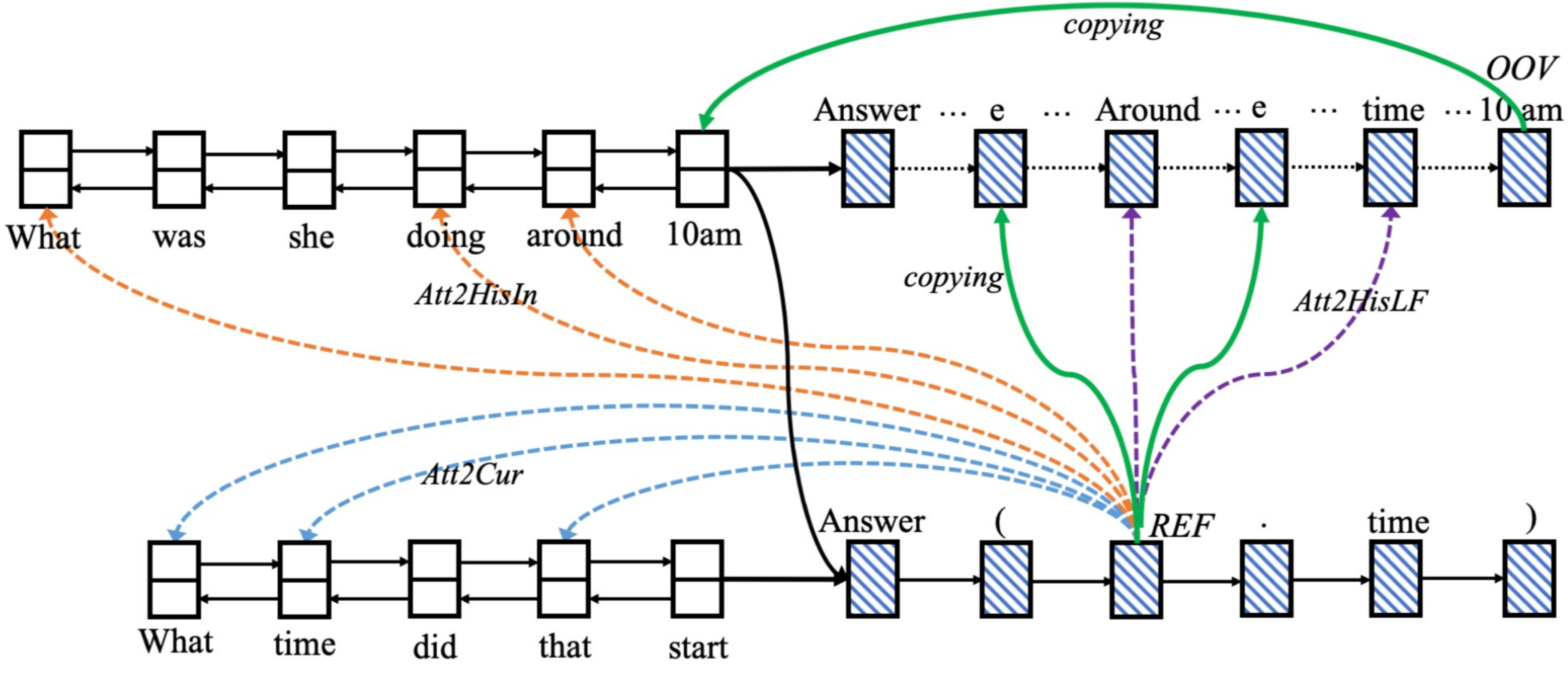
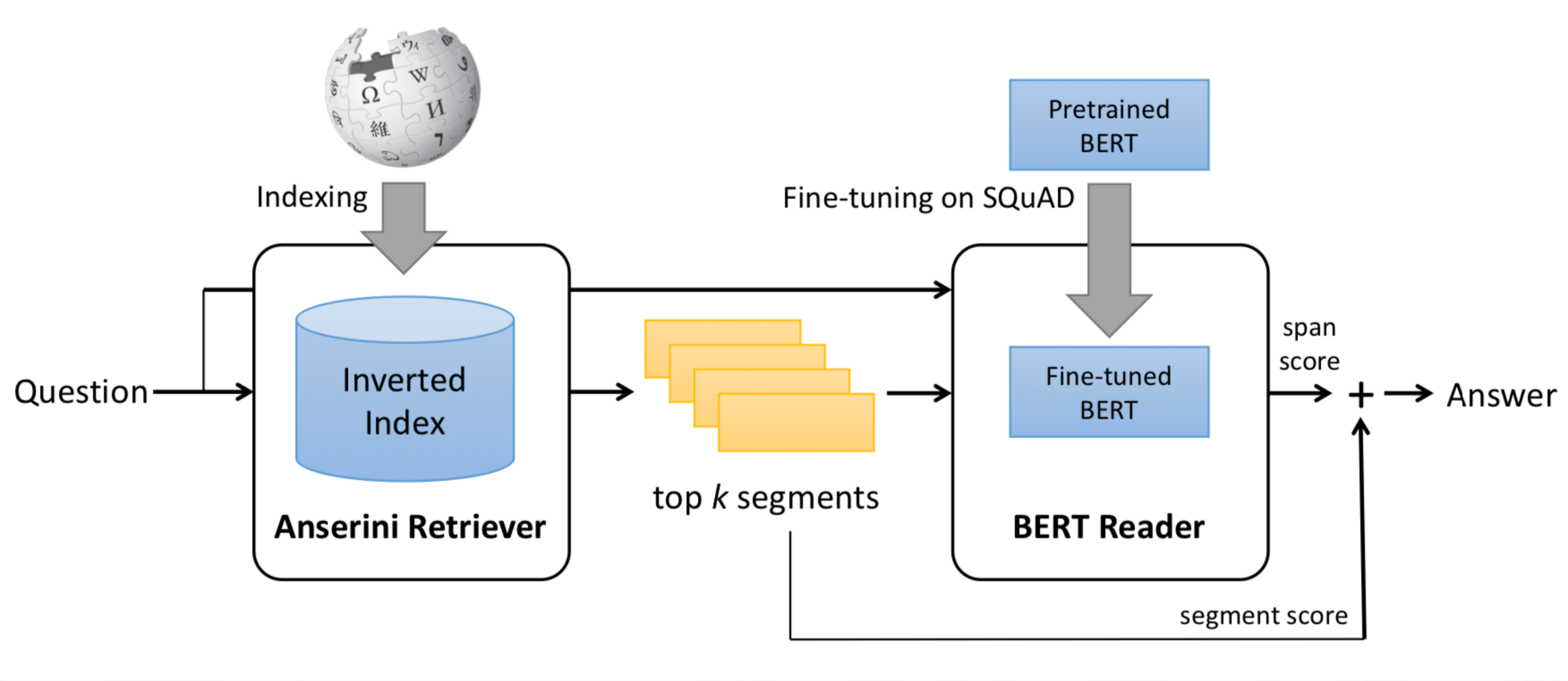


Figure 4: Context-dependent semantic parsing architecture. We use a Bi-LSTM (left) to encode the input and a LSTM (right) as the decoder. We show only parts of the LF to save space. The complete generated LF at time T-1 is Y^{-1} = [*Answer*, (, *e*,), ∧, *Around*, (, *e*, ., *time*, *OOV*,), ∧, *e*, ., *type*, ==, *DiscreteType*]. The token *10am* is copied

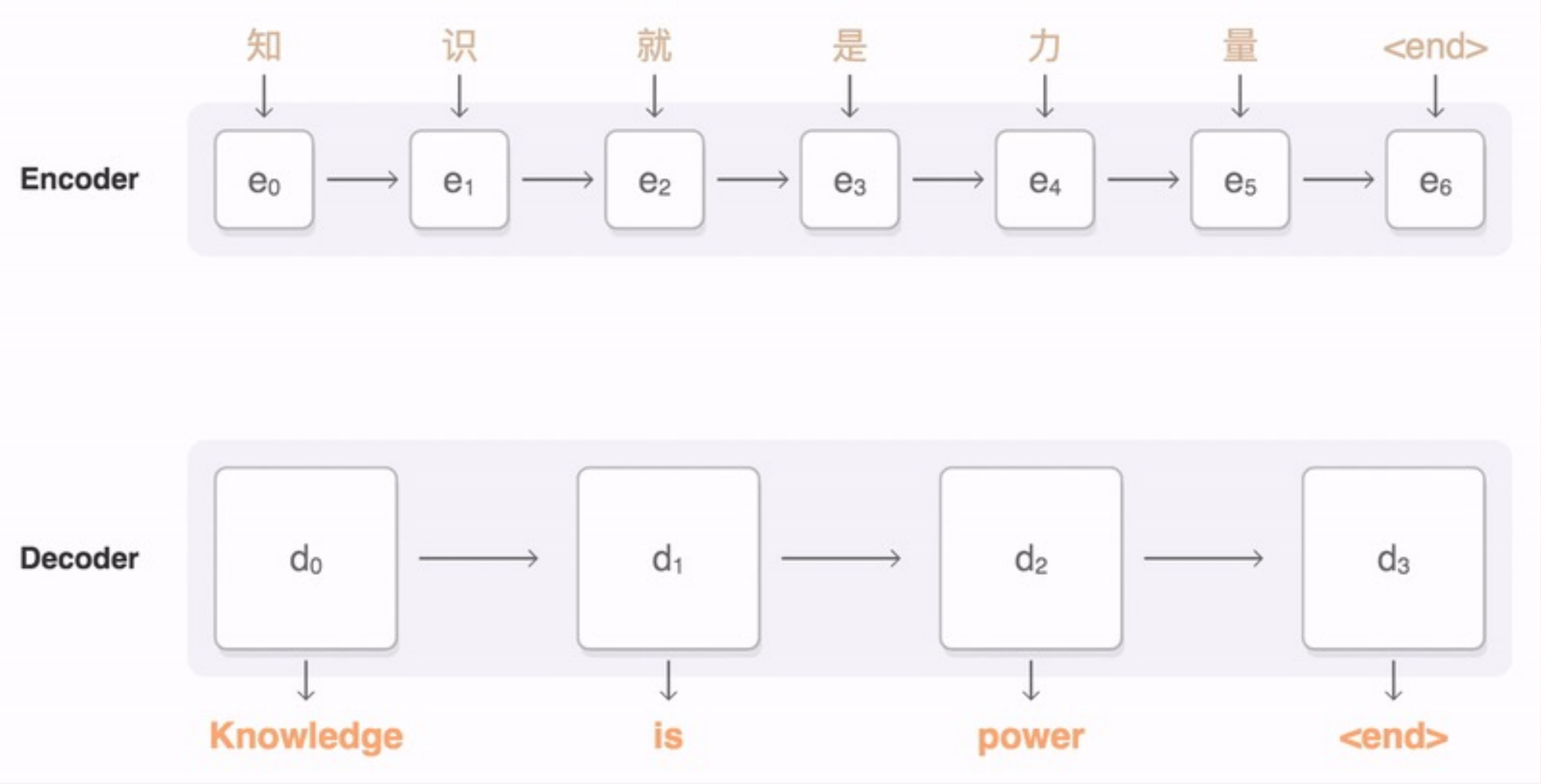
End-to-End: Question Answering

"*End-to-End Open-Domain Question Answering with BERTserini* [Yang et al., NAACL Demo 2019]



End-to-End: Machine Translation with RNNs & Attention

- From Phrase-Based Machine Translation to Neural Machine Translation:
  https://research.googleblog.com/2016/09/a-neural-network-for-machine.html



End-to-End: Machine Translation with RNNs & Attention

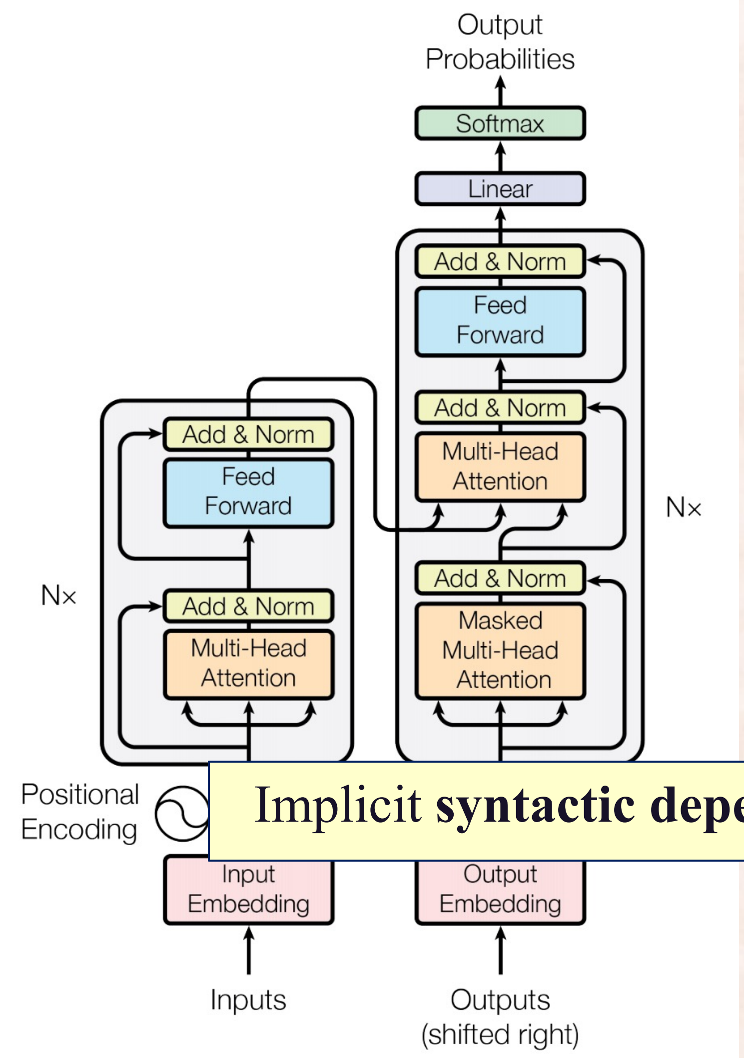
- Before November 2016:
  - Kilimanjaro is 19,710 feet of the mountain covered with snow, and it is said that the highest mountain in Africa. Top of the west, “Ngaje Ngai” in the Maasai language, has been referred to as the house of God. The top close to the west, there is a dry, frozen carcass of a leopard. Whether the leopard had what the demand at that altitude, there is no that nobody explained.

- After November 2016:
  - Kilimanjaro is a mountain of 19,710 feet covered with snow and is said to be the highest mountain in Africa. The summit of the west is called “Ngaje Ngai” in Masai, the house of God. Near the top of the west there is a dry and frozen dead body of leopard. No one has ever explained what leopard wanted at that altitude.

http://www.nytimes.com/2016/12/14/magazine/the-great-ai-awakening.html

End-to-End: Machine Translation With Transformer

https://arxiv.org/pdf/1706.03762.pdf



Implicit **syntactic dependences**.

Attention Visualizations

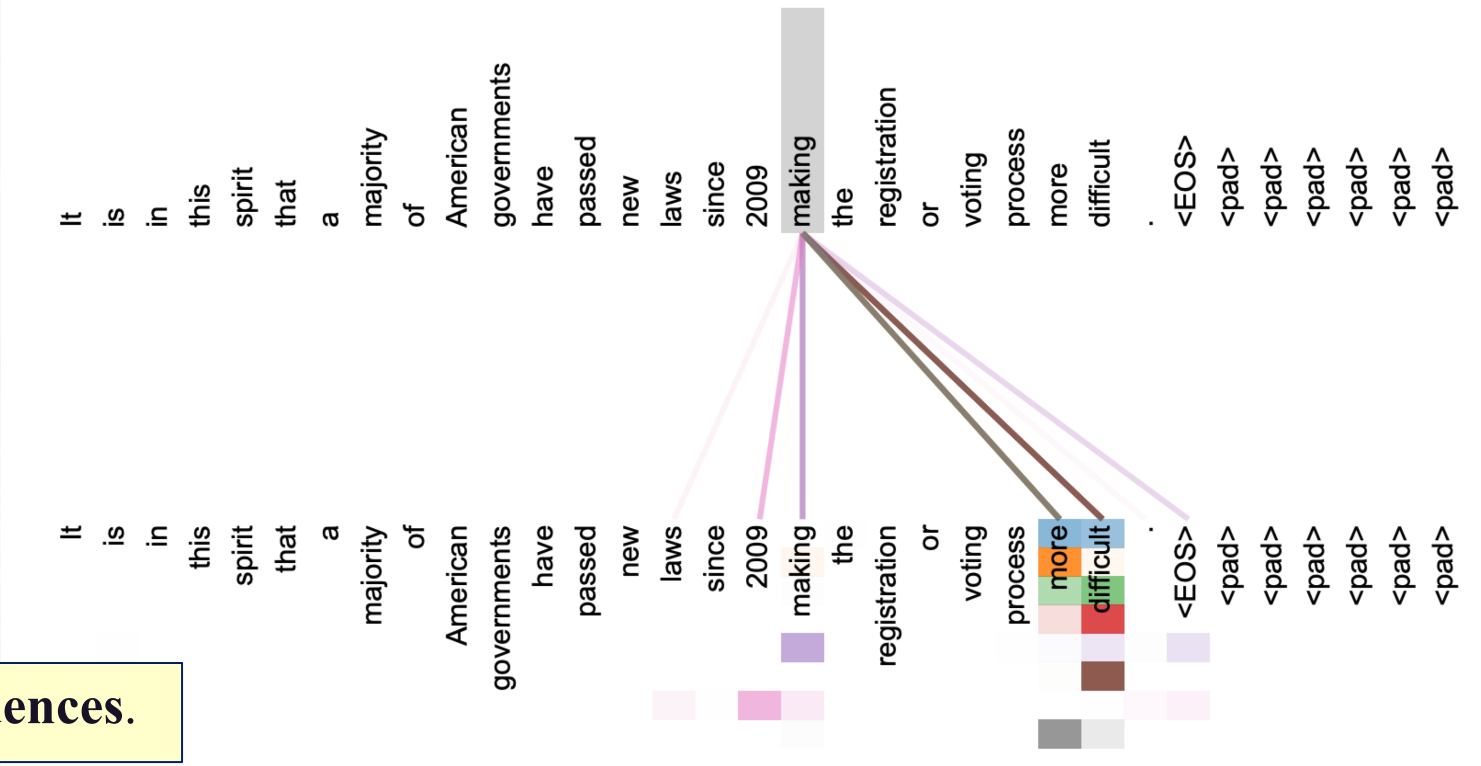
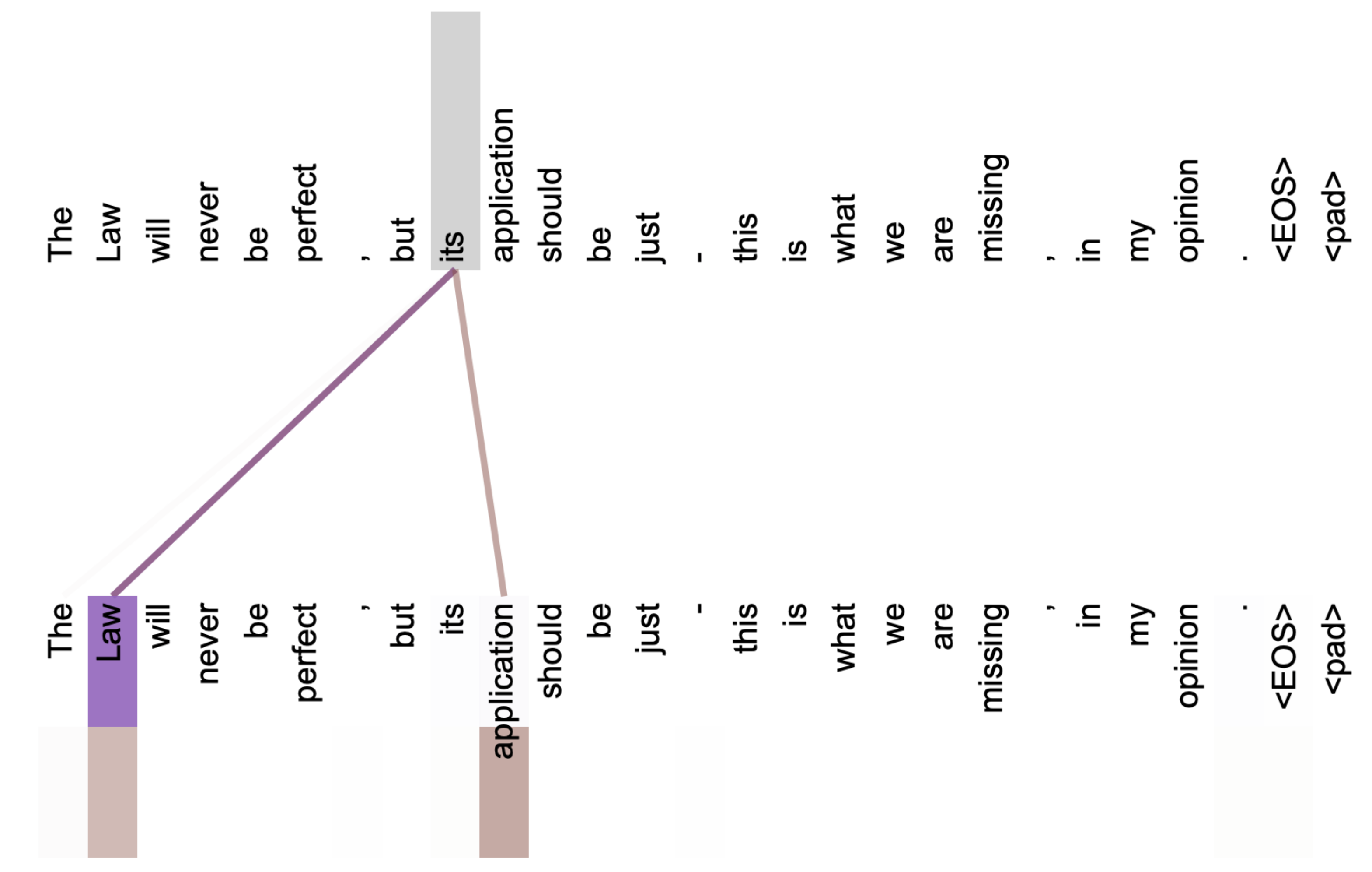


Figure 3: An example of the attention mechanism following long-distance dependencies in the encoder self-attention in layer 5 of 6. Many of the attention heads attend to a distant dependency of the verb 'making', completing the phrase 'making...more difficult'. Attentions here shown only for the word 'making'. Different colors represent different heads. Best viewed in color.

End-to-End: Machine Translation With Transformer



Implicit **coreference resolution**.

End-to-End: Conditional Text Generation

https://openai.com/blog/better-language-models

- System prompt (human written):
  - *In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.*

- Model completion (machine written):
  - The scientist named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown to science. Now, after almost two centuries, the mystery of what sparked this odd phenomenon is finally solved. Dr. Jorge Pérez, an evolutionary biologist from the University of La Paz, and several companions, were exploring the Andes Mountains when they found a small valley, with no other animals or humans. Pérez noticed that the valley had what appeared to be a natural fountain, surrounded by two peaks of rock and silver snow. Pérez and the others then ventured further into the valley. "By the time we reached the top of one peak, the water looked blue, with some crystals on top," said Pérez …

[*February 14, 2019, introducing GPT-2*]

What is Natural Language Processing?

- **Natural Language Processing (NLP)** = developing computer systems that can *process*, *understand*, or *communicate* in natural language (*text* or *speech*):
  - **Natural Languages**: English, Turkish, Japanese, Latin, Hawaiian Creole, Esperanto, American Sign Language, …
    - Music?
  - **Formal Languages**: C++, Java, Python, XML, OWL, Predicate Calculus, Lambda Calculus, …
  - **Natural Languages are significantly more difficult to process than Artificial Languages!**

- What about **Computational Linguistics (CL)** ?
  - Computational Linguistics is focused on the *study of language*, using computational tools.
  - NLP is focused on solving language tasks such as *machine translation*, *information extraction*, *question answering*, *taking instructions*, *holding conversations*, …

CL studies language using NLP/computational tools

[Hamilton et al., EMNLP'16]

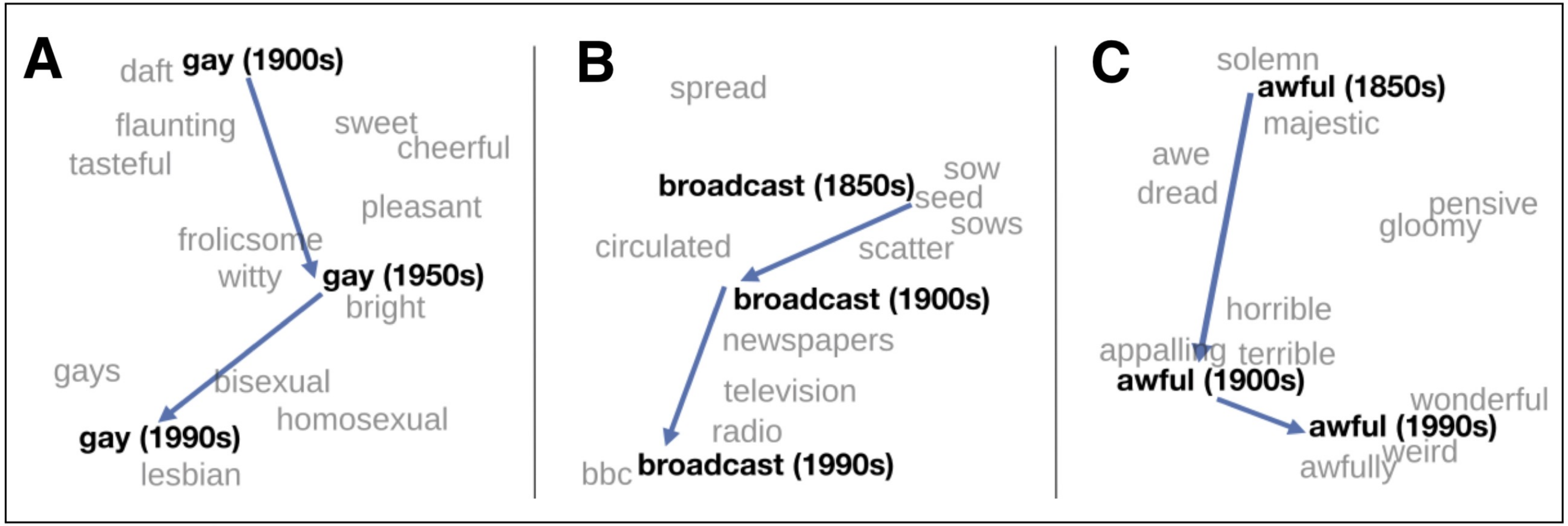


Figure 6.15 A t-SNE visualization of the semantic change of 3 words in English using word2vec vectors. The modern sense of each word, and the grey context words, are computed from the most recent (modern) time-point embedding space. Earlier points are computed from earlier historical embedding spaces. The visualizations show the changes in the word *gay* from meanings related to "cheerful" or "frolicsome" to referring to homosexuality, the development of the modern "transmission" sense of *broadcast* from its original sense of sowing seeds, and the pejoration of the word *awful* as it shifted from meaning "full of awe" to meaning "terrible or appalling" (Hamilton et al., 2016b).

CL studies language using NLP/computational tools

[Hamilton et al., EMNLP'16]

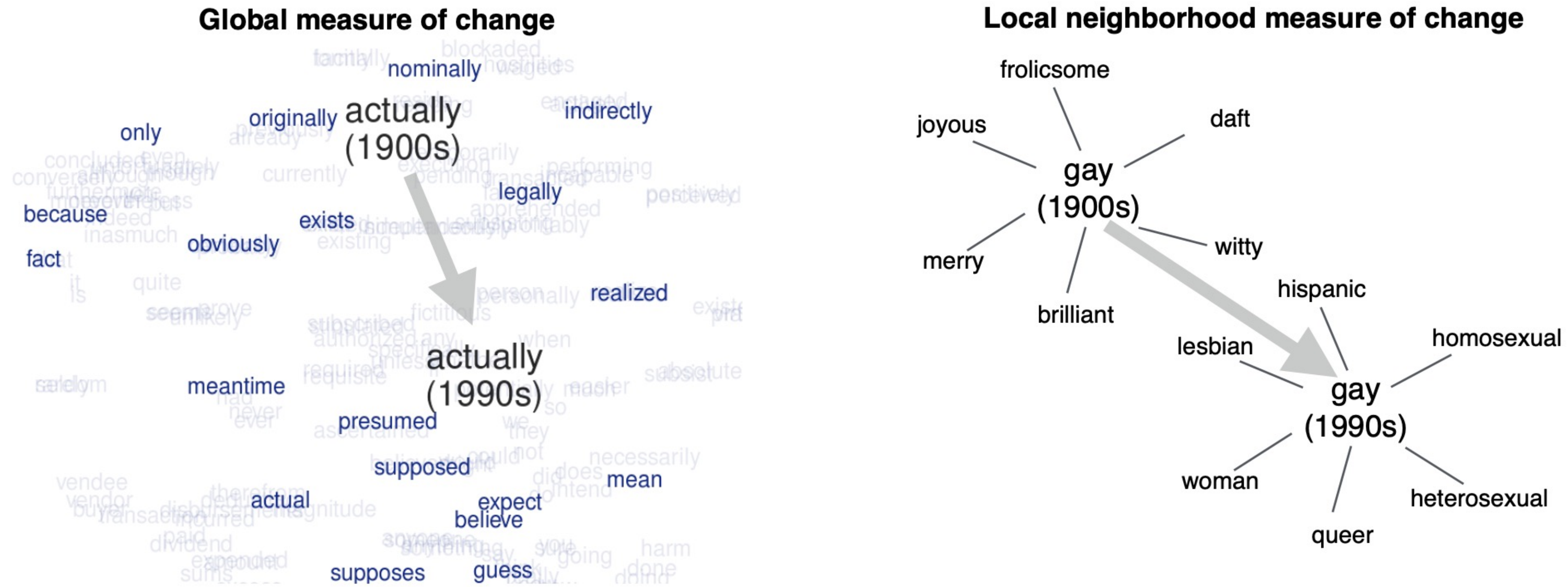


Figure 1: Two different measures of semantic change. With the global measure of change, we measure how far a word has moved in semantic space between two time-periods. This measure is sensitive to subtle shifts in usage and also global effects due to the entire semantic space shifting. For example, this captures how *actually* underwent subjectification during the 20th century, shifting from uses in objective statements about the world ("actually did try") to subjective statements of attitude ("I actually agree"; see Traugott and Dasher, 2001 for details). In contrast, with the local neighborhood measure of change, we measure changes in a word's nearest neighbors, which captures drastic shifts in core meaning, such as *gay*'s shift in meaning over the 20th century.

CL and Psycholinguistics: Syntactic Reduction

- Which of the following utterances are more likely to be produced by speakers?
  1. She created a mobile app dancers like. [example from Meilin Zhan's PhD thesis]
  2. She created a mobile app that dancers like.

  3. My boss thinks I am absolutely crazy. [example from Jaeger T. F.]
  4. My boss thinks that I am absolutely crazy.

  5. My boss confirmed we were absolutely crazy. [example from Jaeger T. F.]
  6. My boss confirmed that we were absolutely crazy.

CL and Psycholinguistics: Syntactic Reduction

- Which of the following utterances are more likely to be produced by speakers?

1. She created a mobile app **dancers** like.
2. 👍 She created a mobile app that dancers like.

red: less predictable (high information).

3. 👍 My boss thinks **I** am absolutely crazy.
4. My boss thinks that I am absolutely crazy.

blue: more predictable (low information).

5. My boss confirmed **we** were absolutely crazy.
6. 👍 My boss confirmed that we were absolutely crazy.

red: less predictable (high information).

- Explained by the **Uniform Information Density** (**UID**) Hypothesis:

  <u>Speakers aim to convey information at a relatively constant rate.</u>

[Jaeger, T. F. (2010). Redundancy and reduction: Speakers manage syntactic information density. *Cognitive Psychology*]

Supplementary Readings

- Chapter 1 (Introduction) in [Jurafsky & Martin]

  and/or

- Chapter 1 (Introduction) in [Eisenstein]

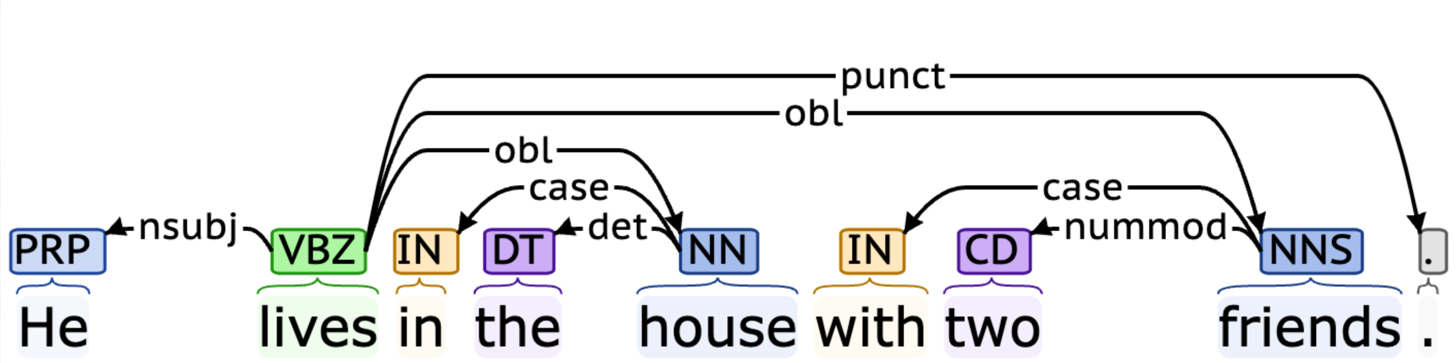
- Python introductory lecture slides
- Python language tutorial

- Regular expressions in Python:
  - https://docs.python.org/3/howto/regex.html

- Extracting linguistic features with spaCy:
  - https://spacy.io/usage/linguistic-features

Syntactic Dependences

- He lives in the house with two friends.



- He lives in the house with two doors.

Coreference in Mathematical Statements

- **Theorem**: √2 is an irrational number.
- **Proof**: Suppose that √2 were a rational number, so by definition √2 = a / b where a and b are non-zero integers with no common factor. Thus, b√2 = a. Squaring both sides yields $2b^2 = a^2$. Since 2 divides the left hand side, 2 must also divide the right hand side (as they are equal and both integers). So a^2 is even, which implies that a must also be even. So we can write a = 2c, where c is also an integer. Substitution into the original equation yields $2b^2 = (2c)^2 = 4c^2$. Dividing both sides by 2 yields $b^2 = 2c^2$. But then, by the same argument as before, 2 divides b^2, so b must be even. However, if a and b are both even, they share a factor, namely 2. This contradicts our assumption, so we are forced to conclude that √2 is an irrational number.

https://en.wikipedia.org/wiki/Mathematical_proof#Proof_by_contradiction

Coreference in Mathematical Statements

- **Theorem**: $\sqrt{2}$ is an irrational number.
- **Proof**: Suppose that $\sqrt{2}$ were a rational number, so by definition $\sqrt{2} = a / b$ where a and b are non-zero integers with no common factor. Thus, $b\sqrt{2} = a$. Squaring both sides yields $2b^2 = a^2$. Since 2 divides **the left hand side**, 2 must also divide the right hand side (as they are equal and both integers). So a^2 is even, which implies that a must also be even. So we can write $a = 2c$, where c is also an integer. Substitution into the original equation yields $2b^2 = (2c)^2 = 4c^2$. Dividing both sides by 2 yields $b^2 = 2c^2$. But then, by the same argument as before, 2 divides b^2, so b must be even. However, if a and b are both even, they share a factor, namely 2. This contradicts our assumption, so we are forced to conclude that $\sqrt{2}$ is an irrational number.

https://en.wikipedia.org/wiki/Mathematical_proof#Proof_by_contradiction

Coreference in Mathematical Statements

- **Theorem**: $\sqrt{2}$ is an irrational number.
- **Proof**: Suppose that $\sqrt{2}$ were a rational number, so by definition $\sqrt{2} = a / b$ where a and b are non-zero integers with no common factor. Thus, $b\sqrt{2} = a$. Squaring both sides yields $\mathbf{2b^2} = a^2$. Since 2 divides **the left hand side**, 2 must also divide the right hand side (as they are equal and both integers). So a^2 is even, which implies that a must also be even. So we can write $a = 2c$, where c is also an integer. Substitution into the original equation yields $2b^2 = (2c)^2 = 4c^2$. Dividing both sides by 2 yields $b^2 = 2c^2$. But then, by the same argument as before, 2 divides b^2, so b must be even. However, if a and b are both even, they share a factor, namely 2. This contradicts our assumption, so we are forced to conclude that $\sqrt{2}$ is an irrational number.

https://en.wikipedia.org/wiki/Mathematical_proof#Proof_by_contradiction

Coreference in Mathematical Statements

- **Theorem**: $\sqrt{2}$ is an irrational number.
- **Proof**: Suppose that $\sqrt{2}$ were a rational number, so by definition $\sqrt{2} = a / b$ where a and b are non-zero integers with no common factor. Thus, $b\sqrt{2} = a$. Squaring both sides yields **$2b^2$** $= a^2$. Since 2 divides **the left hand side**, 2 must also divide **the right hand side** (as they are equal and both integers). So a^2 is even, which implies that a must also be even. So we can write $a = 2c$, where c is also an integer. Substitution into the original equation yields $2b^2 = (2c)^2 = 4c^2$. Dividing both sides by 2 yields $b^2 = 2c^2$. But then, by the same argument as before, 2 divides b^2, so b must be even. However, if a and b are both even, they share a factor, namely 2. This contradicts our assumption, so we are forced to conclude that $\sqrt{2}$ is an irrational number.

https://en.wikipedia.org/wiki/Mathematical_proof#Proof_by_contradiction

Coreference in Mathematical Statements

- **Theorem**: $\sqrt{2}$ is an irrational number.
- **Proof**: Suppose that $\sqrt{2}$ were a rational number, so by definition $\sqrt{2} = a / b$ where a and b are non-zero integers with no common factor. Thus, $b\sqrt{2} = a$. Squaring both sides yields $\mathbf{2b^2} = \mathbf{a^2}$. Since 2 divides **the left hand side**, 2 must also divide **the right hand side** (as they are equal and both integers). So a^2 is even, which implies that a must also be even. So we can write $a = 2c$, where c is also an integer. Substitution into the original equation yields $2b^2 = (2c)^2 = 4c^2$. Dividing both sides by 2 yields $b^2 = 2c^2$. But then, by the same argument as before, 2 divides b^2, so b must be even. However, if a and b are both even, they share a factor, namely 2. This contradicts our assumption, so we are forced to conclude that $\sqrt{2}$ is an irrational number.

https://en.wikipedia.org/wiki/Mathematical_proof#Proof_by_contradiction

Coreference in Mathematical Statements

- **Theorem**: $\sqrt{2}$ is an irrational number.
- **Proof**: Suppose that $\sqrt{2}$ were a rational number, so by definition $\sqrt{2} = a / b$ where a and b are non-zero integers with no common factor. Thus, $b\sqrt{2} = a$. Squaring both sides yields **$2b^2$** = **a^2**. Since 2 divides **the left hand side**, 2 must also divide **the right hand side** (as they are equal and both integers). So a^2 is even, which implies that a must also be even. So we can write $a = 2c$, where c is also an integer. Substitution into **the original equation** yields $2b^2 = (2c)^2 = 4c^2$. Dividing both sides by 2 yields $b^2 = 2c^2$. But then, by the same argument as before, 2 divides b^2, so b must be even. However, if a and b are both even, they share a factor, namely 2. This contradicts our assumption, so we are forced to conclude that $\sqrt{2}$ is an irrational number.

https://en.wikipedia.org/wiki/Mathematical_proof#Proof_by_contradiction

Coreference in Mathematical Statements

- **Theorem**: $\sqrt{2}$ is an irrational number.

- **Proof**: Suppose that $\sqrt{2}$ were a rational number, so by definition $\sqrt{2} = a / b$ where a and b are non-zero integers with no common factor. Thus, $b\sqrt{2} = a$. Squaring both sides yields $\mathbf{2b^2 = a^2}$. Since 2 divides **the left hand side**, 2 must also divide **the right hand side** (as they are equal and both integers). So a^2 is even, which implies that a must also be even. So we can write $a = 2c$, where c is also an integer. Substitution into **the original equation** yields $2b^2 = (2c)^2 = 4c^2$. Dividing both sides by 2 yields $b^2 = 2c^2$. But then, by the same argument as before, 2 divides b^2, so b must be even. However, if a and b are both even, they share a factor, namely 2. This contradicts our assumption, so we are forced to conclude that $\sqrt{2}$ is an irrational number.

https://en.wikipedia.org/wiki/Mathematical_proof#Proof_by_contradiction

Coreference in Mathematical Statements

- **Theorem**: √2 is an irrational number.
- **Proof**: Suppose that √2 were a rational number, so by definition √2 = a / b where a and b are non-zero integers with no common factor. Thus, b√2 = a. Squaring both sides yields **$2b^2 = a^2$**. Since 2 divides **the left hand side**, 2 must also divide **the right hand side** (as they are equal and both integers). So a^2 is even, which implies that a must also be even. So we can write a = 2c, where c is also an integer. Substitution into **the original equation** yields $2b^2 = (2c)^2 = 4c^2$. Dividing **both sides** by 2 yields $b^2 = 2c^2$. But then, by the same argument as before, 2 divides b^2, so b must be even. However, if a and b are both even, they share a factor, namely 2. This contradicts our assumption, so we are forced to conclude that √2 is an irrational number.

https://en.wikipedia.org/wiki/Mathematical_proof#Proof_by_contradiction

Coreference in Mathematical Statements

- **Theorem**: √2 is an irrational number.
- **Proof**: Suppose that √2 were a rational number, so by definition √2 = a / b where a and b are non-zero integers with no common factor. Thus, b√2 = a. Squaring both sides yields **$2b^2 = a^2$**. Since 2 divides **the left hand side**, 2 must also divide **the right hand side** (as they are equal and both integers). So a^2 is even, which implies that a must also be even. So we can write a = 2c, where c is also an integer. Substitution into **the original equation** yields **$2b^2$** $= (2c)^2 =$ **$4c^2$**. Dividing **both sides** by 2 yields $b^2 = 2c^2$. But then, by the same argument as before, 2 divides b^2, so b must be even. However, if a and b are both even, they share a factor, namely 2. This contradicts our assumption, so we are forced to conclude that √2 is an irrational number.

https://en.wikipedia.org/wiki/Mathematical_proof#Proof_by_contradiction

Coreference in Mathematical Statements

- **Theorem**: √2 is an irrational number.
- **Proof**: Suppose that √2 were a rational number, so by definition √2 = a / b where a and b are non-zero integers with no common factor. Thus, b√2 = a. Squaring **both sides** yields $\mathbf{2b^2 = a^2}$. Since 2 divides **the left hand side**, 2 must also divide **the right hand side** (as **they** are equal and **both** integers). So a^2 is even, which implies that a must also be even. So we can write a = 2c, where c is also an integer. Substitution into **the original equation** yields $\mathbf{2b^2} = (2c)^2 = \mathbf{4c^2}$. Dividing **both sides** by 2 yields $b^2 = 2c^2$. But then, by **the same argument as before**, 2 divides b^2, so b must be even. However, if a and b are **both** even, **they** share a factor, namely 2. **This** contradicts **our assumption**, so we are forced to conclude that √2 is an irrational number.

https://en.wikipedia.org/wiki/Mathematical_proof#Proof_by_contradiction