COMP5046
Natural Language Processing

Lecture 5: Chatbot and Language Fundamental

Semester 1, 2019

School of Computer Science
The University of Sydney, Australia

THE UNIVERSITY OF SYDNEY

Caren Han
Caren.Han@sydney.edu.au

0 LECTURE PLAN

Lecture 5: Chatbot and Language Fundamental

1. **The NLP Big Picture**
2. Conversational Agent
    1. Overview
    2. Assignment Specification
3. Language Fundamental
    - Phonology, Morphology, Syntax, Semantics, Pragmatics
4. Text Preprocessing
    1. Tokenization
    2. Cleaning and Normalisation
    3. Stemming and Lemmatisation
    4. Stopword
    5. Regular Expression

1 The NLP Big Picture

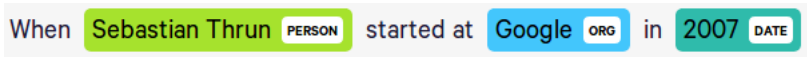
The purpose of Natural Language Processing: Overview

Application

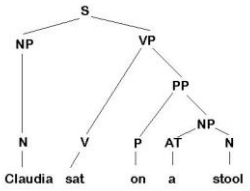
- ***Understanding***
  - ***Sentiment Analysis***
  - ***Topic Classification***
  - ***Topic Modelling***
  - ***....***
  - ***Dialog***
  - ***Translation***
- ***Searching***
  - ***Search***
  - ***....***

NLP Stack

- ***Entity Extraction*** — When Sebastian Thrun …



- ***Parsing*** — Claudia sat on a stool



- ***PoS Tagging*** — She sells seashells — **[she/PRP] [sells/VBZ] [seashells/NNS]**
- ***Stemming*** — Drinking, Drank, Drunk — **Drink**
- ***Tokenisation*** — How is the weather today — **[How] [is] [the] [weather] [today]**

0 LECTURE PLAN

THE UNIVERSITY OF SYDNEY

Lecture 5: Chatbot and Language Fundamental

1. The big picture of NLP
2. **Conversational Agent**
    1. Overview
    2. Assignment Specification
3. Language Fundamental
    - Phonology, Morphology, Syntax, Semantics, Pragmatics
4. Text Preprocessing
    1. Tokenization
    2. Cleaning and Normalisation
    3. Stemming and Lemmatisation
    4. Stopword
    5. Regular Expression

Conversational Agent

Conversational Agent

A conversational agent is a software program which interprets and responds to statements made by users in ordinary natural language. It integrates computational linguistics techniques with communication over the internet

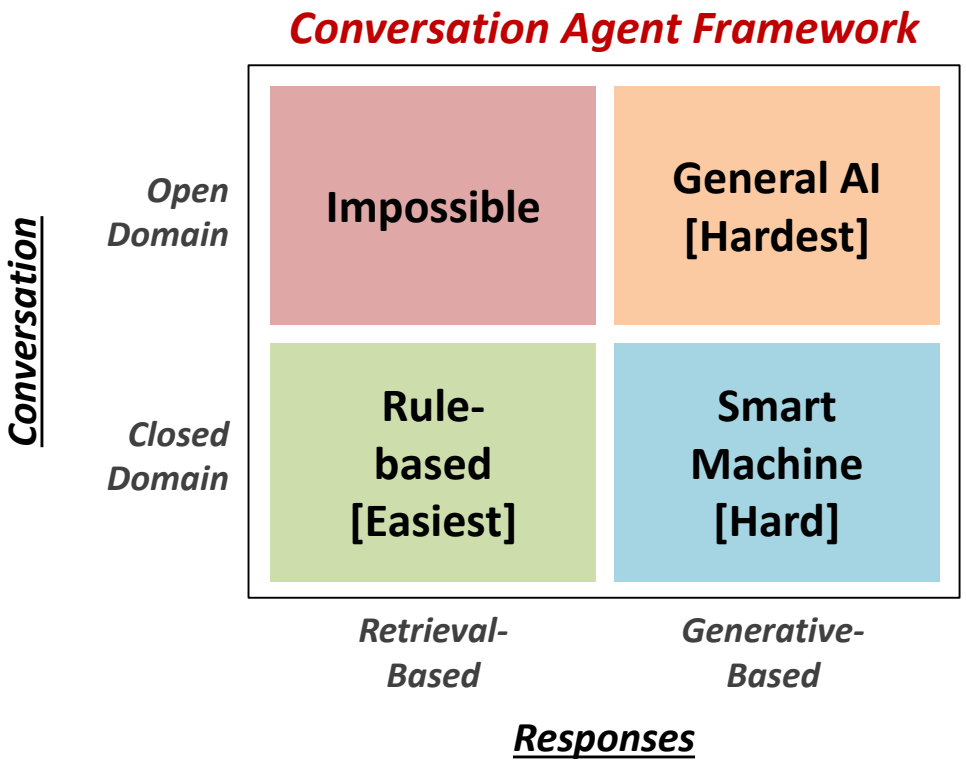
Phone/Computer based Personal Assistant

- *Chatting for fun*
- *Talking to your car*

Conversational Agent

Conversational Agent

A conversational agent is a software program which interprets and responds to statements made by users in ordinary natural language. It integrates computational linguistics techniques with communication over the internet



Conversational Agent

Conversational Agent

A conversational agent is a software program which interprets and responds to statements made by users in ordinary natural language. It integrates computational linguistics techniques with communication over the internet

Goal-oriented Conversational Agent

Designed for a particular task, utilizing short conversations to get information from the user to help complete this task

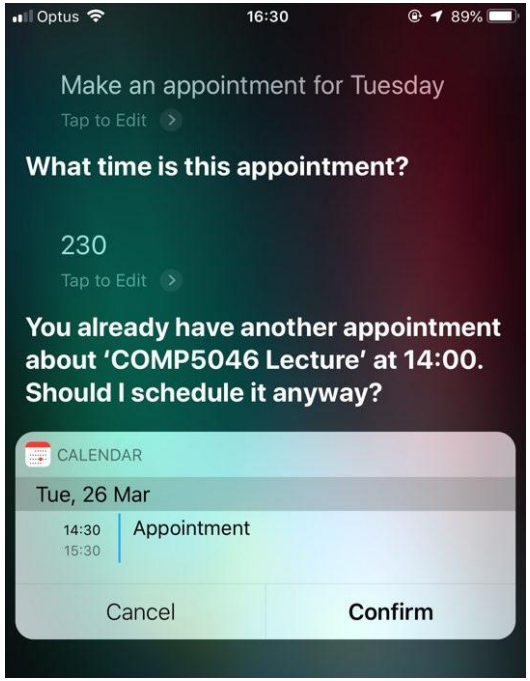
Chatbots (Chat-oriented Conversational Agent)

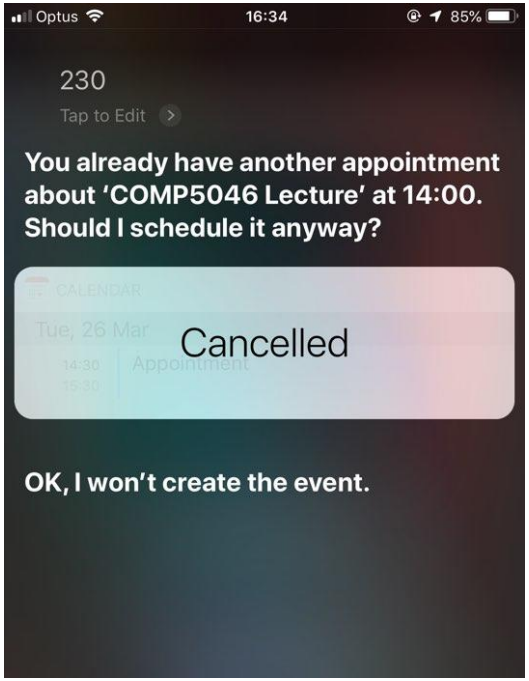
Designed to handle full conversations, mimicking the unstructured flow of a human to human conversation

Conversational Agent

Goal-oriented Conversational Agent

Designed for a particular task, utilizing short conversations to get information from the user to help complete this task



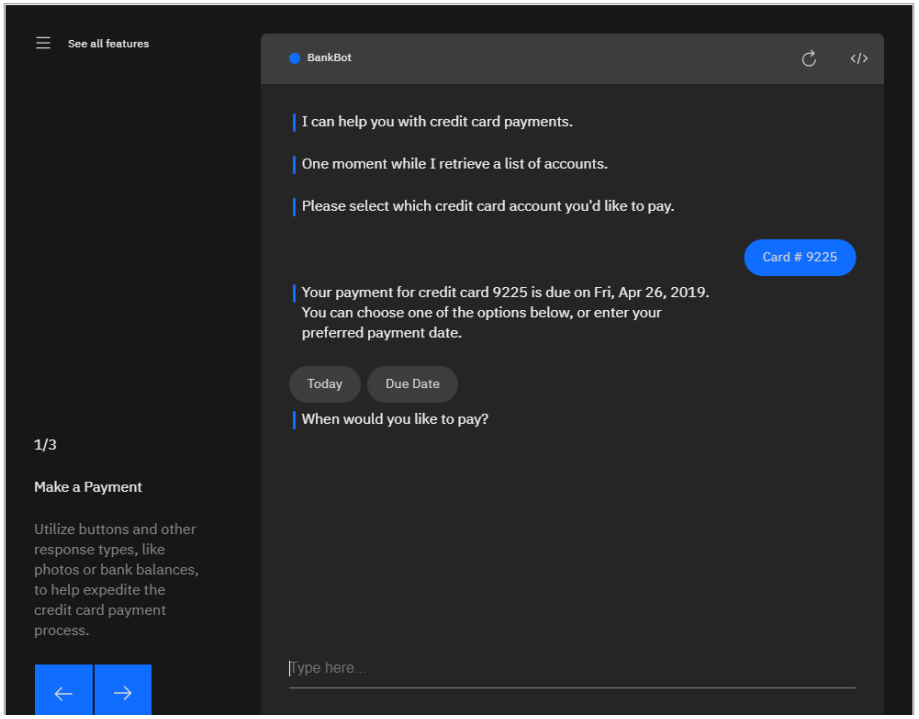


Apple Siri

Conversational Agent

Goal-oriented Conversational Agent

Designed for a particular task, utilizing short conversations to get information from the user to help complete this task



IBM Watson BankBot

Conversational Agent

Goal-oriented Conversational Agent

Frame-based Approach

- Based on a "**domain ontology**"
  A knowledge structure representing user intentions

- One or more Frame
  Each a collection of **slots**
  Each slot having a **value**
  A set of **slots**, to be filled with information of a given **type**
  Each associated with a **question** to the user

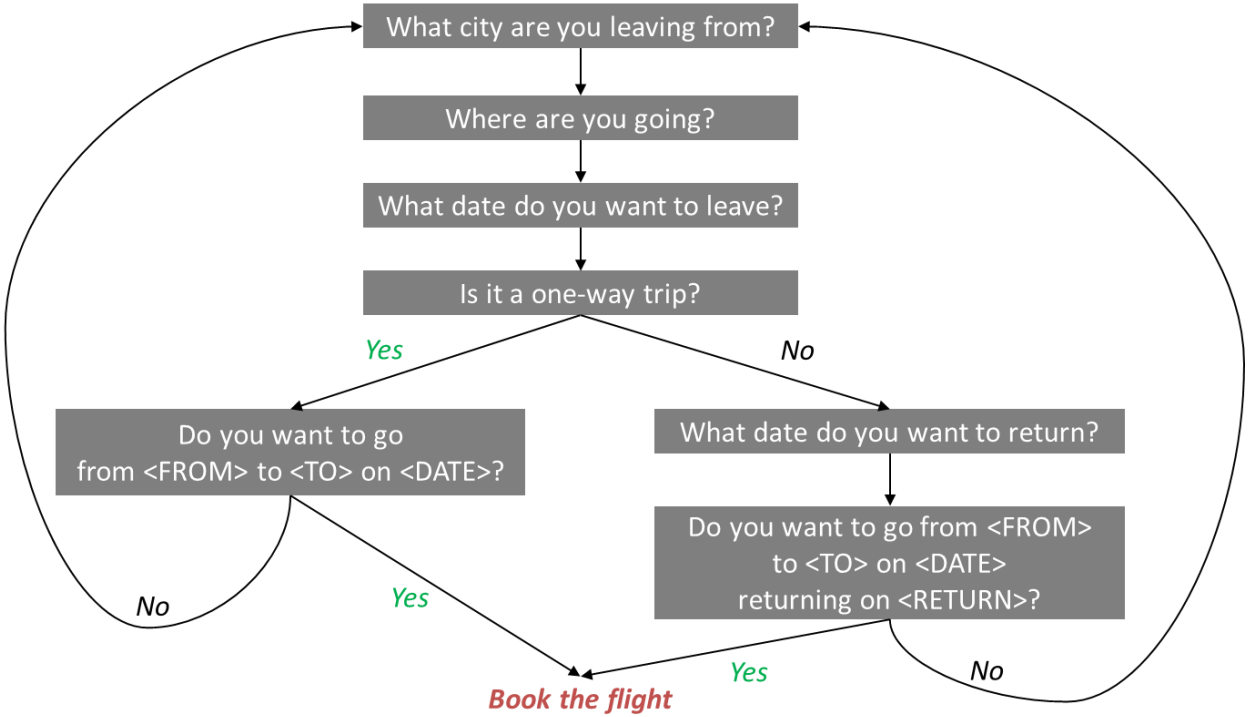
Slot	*Type*	*Question*
ORIGIN	*city*	*What city are you leaving from?*
DEST	*city*	*Where are you going?*
DEPT DATE	*date*	*What day would you like to leave?*
DEPT TIME	*time*	*What time would you like to leave?*
AIRLINE	*line*	*What is your preferred airline?*

Conversational Agent

Goal-oriented Conversational Agent

Dialogue is structured in a sequence of predetermined utterance

- Ask the user for a departure city
- Ask for a destination city
- Ask for a time
- Ask whether the trip is round--trip or not



Conversational Agent

Goal-oriented Conversational Agent

- System completely controls the conversation with the user.
- It asks the user a series of questions
- Ignoring (or misinterpreting) anything the user says that is not a direct answer to the system's questions

Conversational Agent

Dialogue Initiative

Systems that control conversation like this are:
system initiative or ***single initiative***

Initiative: who has control of conversation

In normal human to human dialogue,
initiative shifts back and forth between participants

2 Conversational Agent

THE UNIVERSITY OF SYDNEY

System Initiative

System completely controls the conversation

- ***Simple to build***
- ***User always knows what they can say next***
- ***System always knows what user can say next***
- ***Good for Very Simple tasks (entering a credit card, booking a flight)***

- ***Too limited: does not generate any new text, they just pick a response from a fixed set***
- ***A lot of hard coded rules have to be written so not much intelligent***

Conversational Agent

System Initiative: Issue

"Hi, I'd like to fly from Sydney Tuesday morning; I want a flight from Melbourne to Perth one way leaving after 5 p.m. on Wednesday."

- Answering more than one question in a sentence

Conversational Agent

Mixed Initiative

Conversational initiative can shift between system and user

"Hi, I'd like to fly from Sydney Tuesday morning; I want a flight from Melbourne to Perth one way leaving after 5 p.m. on Wednesday."

A kind of **mixed initiative**

- use the structure of the **frame** to guide dialogue

- System asks questions of user, filling any slots that user specifies
  - When frame is filled, do database query
- If user answers 3 questions at once, system can fill 3 slots and not ask these questions again!

Conversational Agent

Mixed Initiative

- There are many ways to represent the meaning of sentences
- For speech dialogue systems, most common approach is "Frame and slot semantics".

"Show me morning flights from Sydney to Perth on Tuesday."

DOMAIN:	*AIR-TRAVEL*
INTENT:	*SHOW-FLIGHTS*
ORIGIN-CITY:	*Sydney*
ORIGIN-DATE:	*Tuesday*
ORIGIN-TIME:	*morning*
DEST-CITY:	*Perth*

Conversational Agent

SIRI: Condition-Action Rules

Active Ontology: Relational network of concepts

- **Data structures**: a meeting has:
  - a date and time,
  - a location,
  - a topic
  - a list of attendees
- **Rule sets** that perform actions for concepts
  - The date concept turns string
  - Monday at 2pm into
  - Date object *date(DAY,MONTH,YEAR,HOURS,MINUTES)*

Rule: Condition + Action

Conversational Agent

Improvements to the Rule-based Approach

Machine Learning classifiers to map words to semantic frame-fillers

Given a set of labeled sentences

- "I want to fly to Sydney on Tuesday"
- Destination: Sydney
- Depart-date: Tuesday

Build a classifier to map from one to the author

Requirements: Lots of Labeled Data

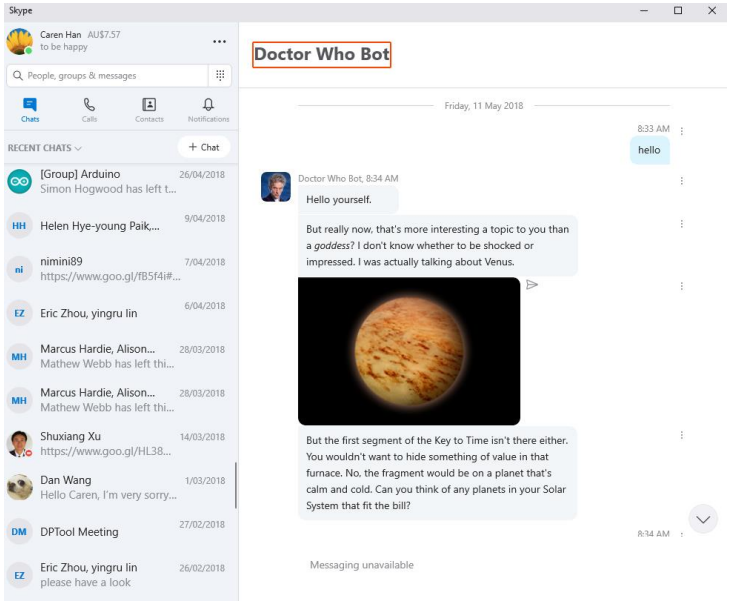
Conversational Agent

Conversational Agent

A conversational agent is a software program which interprets and responds to statements made by users in ordinary natural language. It integrates computational linguistics techniques with communication over the internet

Chatbot

Designed to handle full conversations, mimicking the unstructured flow of a human to human conversation



Conversational Agent

Chatbot

Designed to handle full conversations, mimicking the unstructured flow of a human to human conversation

Rule-based

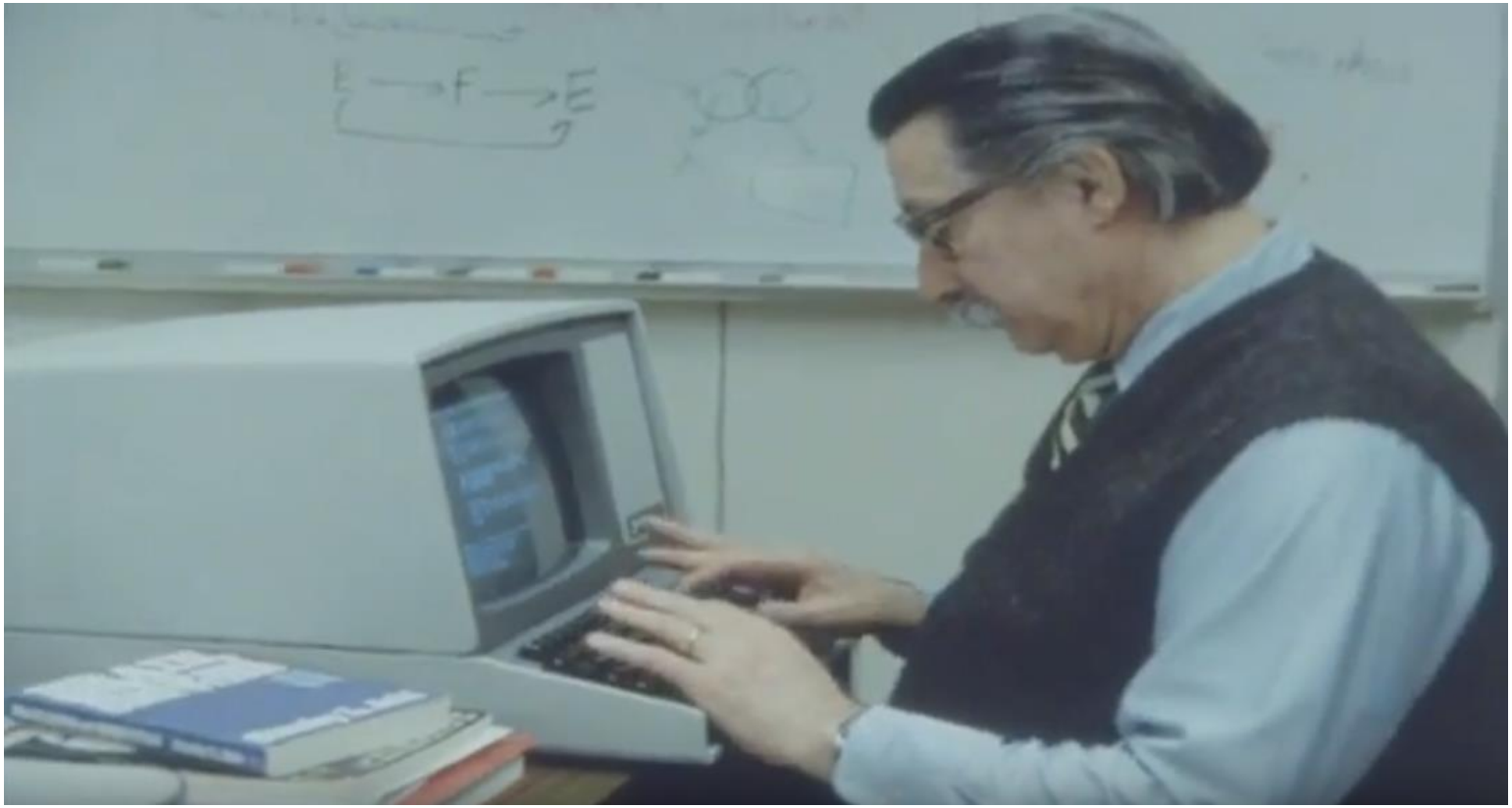
- Pattern-Action Rules (Eliza)
- Pattern-Action Rules + A mental model (Parry)

Corpus-based (from large chat corpus)

- Information Retrieval
- Deep Neural Networks

Conversational Agent

Chatbot: Eliza (1966)



Try Eliza

http://psych.fullerton.edu/mbirnbaum/psych101/Eliza.htm

https://playclassic.games/game/play-eliza-online/play/

Conversational Agent

Chatbot: Eliza (1966)

Domain: Rogerian Psychology Interview

- Draw the patient out by reflecting patient's statements back at them
- Rare type of conversation in which one can "assume the pose of knowing almost nothing of the real world"

Patient: "I went for a long boat ride"
Psychiatrist: "Tell me about boats"

- You don't assume she didn't know what a boat is
- You assume she had some conversational goal

Conversational Agent

Chabot: Eliza (1966)

Pattern matching

if the input matches
(first bunch of words) "you" (second bunch of words) "me"
response with
"What makes you think I" (second bunch of words) "you?"
if the input matches
"You are" (bunch of words)
response with
"So, I'm" (bunch of words) ", am I?"

Very basic reconstruction rules

"me" → "you"
"my" → "your" *etc.*

Conversational Agent

Chatbot: Eliza (1966)

Some programmed responses to special keywords

if the word "mother" appears anywhere, reply with
"Don't you talk about my mother"

Randomisation to avoid getting stuck in a rut

When all else fails, some stock responses,

"Tell me more"
"Fascinating"
"I see"

Conversational Agent

Chatbot: Parry (1972)

Same pattern--response structure as Eliza

Persona

- 28--year--old single man, post office clerk
- no siblings and lives alone
- Sensitive about his physical appearance, his family, his religion, his education and the topic of sex.
- Hobbies are movies and gambling on horseracing,
- Recently attacked a bookie, claiming the bookie did not pay off in a bet.
- Afterwards worried about possible underworld retaliation
- Eager to tell his story to non--threating listeners.

Conversational Agent

Chatbot: Parry (1972)

⟨OTHER'S INTENTION⟩ ← ⟨MALEVOLENCE⟩ | ⟨BENEVOLENCE⟩ | ⟨NEUTRAL⟩

MALEVOLENCE-DETECTION RULES

1. ⟨malevolence⟩ ← ⟨mental harm⟩ | ⟨physical threat⟩
2. ⟨mental harm⟩ ← ⟨humiliation⟩ | ⟨subjugation⟩
3. ⟨physical threat⟩ ← ⟨direct attack⟩ | ⟨induced attack⟩
4. ⟨humiliation⟩ ← ⟨explicit insult⟩ | ⟨implicit insult⟩
5. ⟨subjugation⟩ ← ⟨constraint⟩ | ⟨coercive treatment⟩
6. ⟨direct attack⟩ ← CONCEPTUALIZATIONS ([you get electric shock], [are you afraid mafia kill you?])
7. ⟨induced attack⟩ ← CONCEPTUALIZATIONS ([I tell mafia you], [does mafia know you are in hospital?])
8. ⟨explicit insult⟩ ← CONCEPTUALIZATIONS ([you are hostile], [you are mentally ill?])
9. ⟨implicit insult⟩ ← CONCEPTUALIZATIONS ([tell me your sexlife], [are you sure?])
10. ⟨constraint⟩ ← CONCEPTUALIZATIONS ([you stay in hospital], [you belong on locked ward])
11. ⟨coercive treatment⟩ ← CONCEPTUALIZATIONS ([I hypnotize you], [you need tranquilizers])

Conversational Agent

Chatbot

Rule-based

- Pattern-Action Rules (Eliza)
- Pattern-Action Rules + A mental model (Parry)

Corpus-based (from large chat corpus)

- **Information Retrieval**
- **Deep Neural Networks**

Conversational Agent

Information Retrieval (IR) based Chatbot

- Mine conversations of human chats or human-machine chats
  - Microblogs: Twitter etc.
  - Movie Dialogs
- With large corpus

Cleverbot

Microsoft Xiaoice

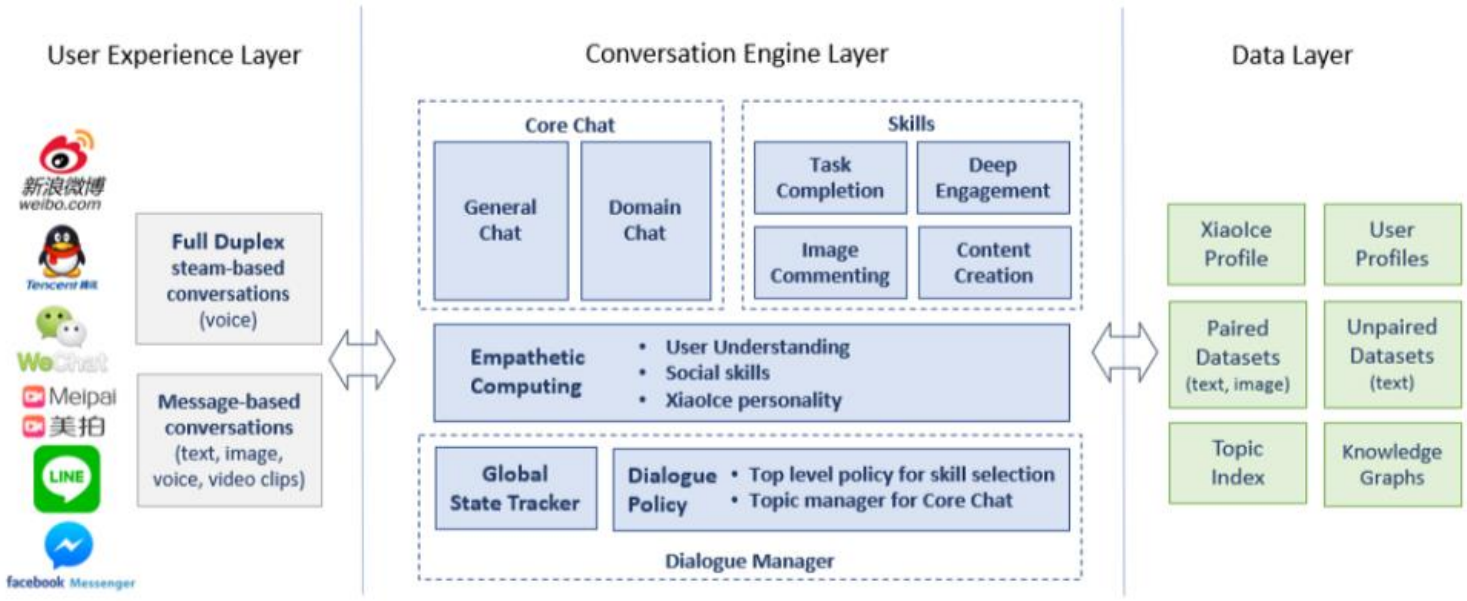
Microsoft Tay

https://youtu.be/Lr4yi9onykg

Conversational Agent

Information Retrieval (IR) based Chatbot

Xiaoice



Conversational Agent

Information Retrieval (IR) based Chatbot

1. Return the response to the most similar turn
   - Take user's turn (q) and find a (tf-idf) similar turn t in the corpus C

     q="do you like Doctor Who"
     t="do you like Doctor Strange"

   - Grab whatever the response was to t.

$$r = response\left(\underset{t \in C}{\text{argmax}} \frac{q^T t}{||q||t||}\right)$$

Yes, love it!

2. Return the most similar turn

$$r = \underset{t \in C}{\text{argmax}} \frac{q^T t}{||q||t||}$$

Do you like Doctor Strangelove?

Conversational Agent

Information Retrieval (IR) based Chatbot

1. Also fine to use other features like user features, or prior turns

2. Or non-dialogue text
   - COBOT chatbot (Isbell et al., 2000)
   - sentences from the Unabomber Manifesto by Theodore Kaczynski, articles on alien abduction, the scripts of “The Big Lebowski” and “Planet of the Apes”.
3. Wikipedia text

Conversational Agent

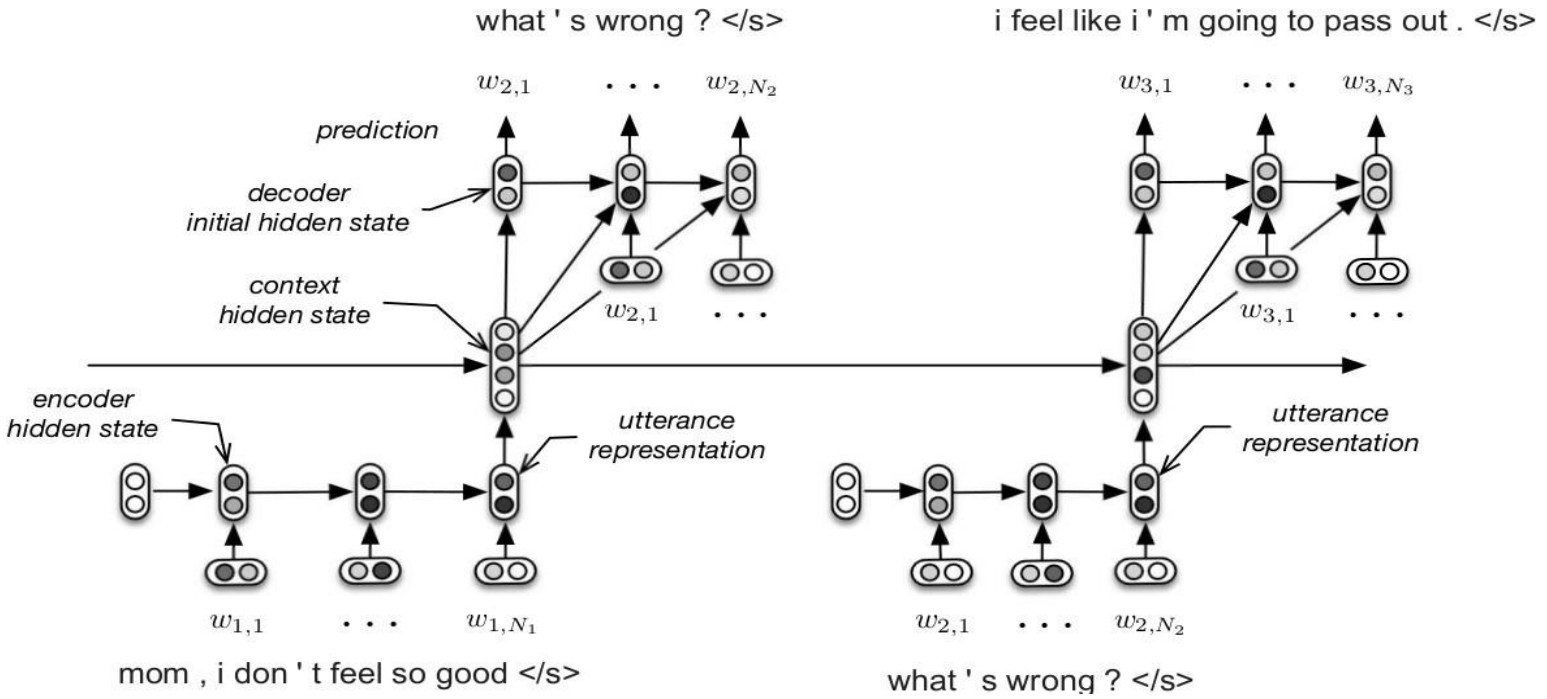
Deep-learning Chatbots

- Think of response generation as a task of transducing from the user's prior turn to the system's turn.
- Train on:
  - Movie Dialogs
  - Twitter Conversations

- Train a deep neural network
  - Map from user 1 turn to user 2 response

Conversational Agent

Seq2seq model architecture



2 Conversational Agent

THE UNIVERSITY OF SYDNEY

Deep learning chatbots

Trained on 127M Twitter context-message-response triples

#	Context c	Message m	Generated Response r
1	noone can help this time its just i need a support system	well im here to support you. whenever you need me.	thank you, i appreciate it thank you
2	nothin much, and how's the book?!	its good but i'm only like halfway through cuz i don't feel like reading. i'm so bored ...	that's good! i have the book but i'm bored too.
3	i know! what have you been up to?	nothing really. just enjoying summer and working. what about you?	that's good! nothing much just listening to music and you?
4	thanks, your way better though	i'm just good at expressing it that's but trust me you are freakin amazing and waay better	i know, i don't trust you, i love you too! thank you
5	thanks. yeah i am good thanks! good, but angry!	sorry you're angry, been there, i hope it dies down.	i hope so too, i'm not angry!!
6	figures. bass. never ate bass.	bass is extremely mild with very few bones. my favorite after catfish	that's one of the bones, chuck bass is mine too.
7	today i will walk for miles. clear sky clear mind.	love walking. it's too hot to walk this month. i haven't moved. frozen. heat freeze.	it's too hot for you to get up to honey if you have frozen yogurt to you.
8	arresting people over "nasty tweets"? please. i'm not condoning what the lad did but arresting him is a tad silly.	it appears he may have sent death threats to others while all this was kicking off too.	i did! he was armed with nuclear threats? that's what's happening to you.

Conversational Agent

Deep learning chatbots

Sample neural chatbot output

Input	tell me ... how long have you had this falling sickness ?
Vanilla-SEQ2SEQ	i'm not a doctor.
Adversarial	a few months, i guess .
Input	so i had the doctors test sammy 's response to conditioning .
Vanilla-SEQ2SEQ	sammy wrote the test sammy wrote the test .
Adversarial	so he took the pills .
Input	they didn 't have much success with this problem commander .
Vanilla-SEQ2SEQ	they 're not the only ones who have been in the system .
Adversarial	can we find someone else ?

Conversational Agent

Deep learning chatbots

- Fun
- Applications to counseling
- Good for narrow, scriptable applications

- They don't really understand
- Rule-based chatbots are expensive and brittle
- IR-based chatbots can only mirror training data
  - The case of Microsoft Tay
  - (or, Garbage-in, Garbage-out)

The future: combining chatbots with frame-based agents

Conversational Agent

Summary

Goal-oriented Conversational Agent:

- Ontology + hand-written rules for slot fillers
- Machine learning classifiers to fill slots

Chatbots:

- Simple rule-based systems
- IR-based: mine datasets of conversations.
- Neural net models with more data

The future…

- Need to acquire that data
- Integrate goal-based and chatbot-based systems

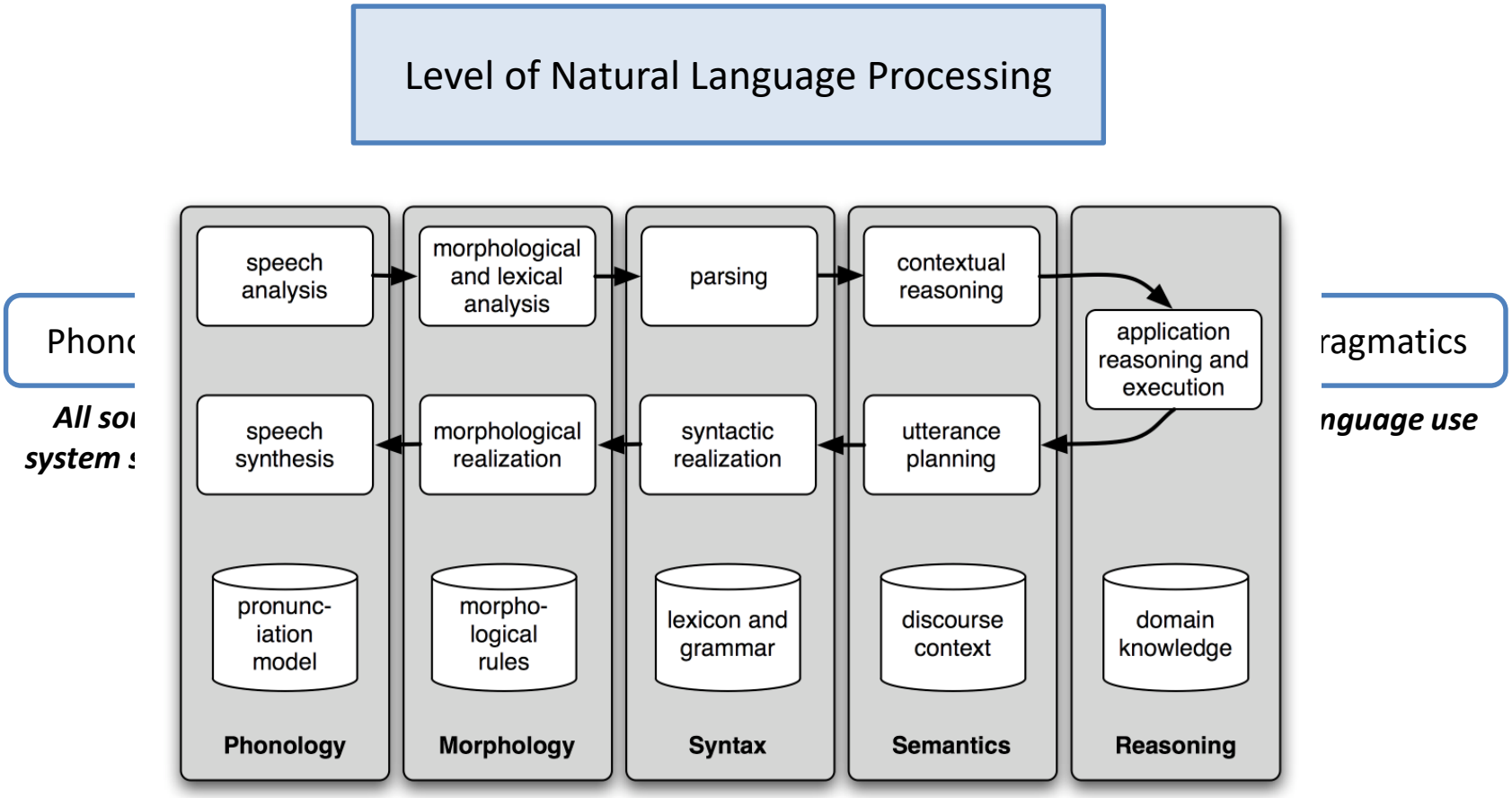
LECTURE PLAN

Lecture 5: Chatbot and Language Fundamental

1. Machine Learning and NLP: Finish
2. Chatbot
3. **Language Fundamental**
   - Phonology, Morphology, Syntax, Semantics, Pragmatics
4. Text Preprocessing
   1. Tokenization
   2. Cleaning and Normalisation
   3. Stemming and Lemmatisation
   4. Stopword
   5. Regular Expression

Language Fundamental

Level of Natural Language Processing



Language Fundamental

We know the sounds of our language

- **which sounds are in our language and which sounds are not**
- For example, English speakers know the [ŋ] sound in sing cannot appear at the beginning of a word
- Does this mean [ŋ] cannot appear at the beginning of words in all human languages?

NO! — nguyen tran

Language Fundamental

We know how sounds can combine

- this is often shown when a word from one language is borrowed into another:

- McDonalds — in English consonant clusters allowed ([mk] and [ldz])becomes…

マクドナルド	麦当劳	맥도날드
Makudonarudo	Màidāngláo	Maegdonaldeu

in other language — consonant clusters are not allowed

Language Fundamental

Meaning: semantics and pragmatics

- We know the context of an utterance shapes meaning:
- slang — cool
- sarcasm — nice shirt
- humor — “Last night I shot an elephant in my pajamas. What he was doing in my pajamas I’ll never know.” -Marx (Groucho)

Language Fundamental

Meaning: semantics and pragmatics

- We know words and utterances can be ambiguous…
- “I like the Spanish teacher.”
  - Spanish teacher
  - A teacher of Spanish or a teacher from Spain?
- “The door is unlockable.”
  - Unlockable
  - Cannot be locked or can be changed from a state of being locked?

Language Fundamental

Pieces of words

- What is morphology?
- The study of how words are formed or marked via other processes. Morphemes are the pieces of words: bases, roots and affixes.

- walk walked walking walks walk walk -ed walk -ing walk -s

Language Fundamental

Pieces of words

- We know which morphemes are productive and which are not. English plural -en or -s? — children, brethren, cats, dogs.... MP3? MP3ren???
- We know how to use affixes.
- For example morphemes for not
- un- im- dis- unhappyimhappy dishappy unpossible impossible dispossible unbelief imbelief disbelief
- This is easily(?) learned by children, and often difficult for second language learners

Language Fundamental

How words combine

- What is syntax?
- The study of how words are ordered or built into sentences or phrases. We “know” where words go in relation to one another…
- English vs. Japanese
- inu-ga shannon-o kandashannon-o inu-ga kanda dog shannon bitshannon dog bit ‘The dog bit Shannon.’‘The dog bit shannon.’
- Shannon bit the dog. ≠ The dog bit Shannon.
- Shannon-o inu-ga kanda. = Inu-ga Shannon-o kanda

Language Fundamental

The lexicon: words and morphemes

- What is the lexicon?
- Our mental dictionary — the words we know: chicken, dog, house, gringo, calle, perro, walk, run… also the pieces of words, or morphemes, that we know: -ed, -ing, un-,-o, -a…
- Everyone has a different lexicon, why? D
- Different backgrounds, regional variation, but most speakers of a language share a core set of words... what are these?
- Things like the following: family terms (uncle, aunt, mother, father, etc); prepositions (on, in, at); articles (the, a, an), etc

Language Fundamental

Natural Language Processing Level

- **Phonology/Morphology: the structure of words**
  - *Unusually* is composed of a prefix *un-*, a stem *usual*, and an affix *–ly*. *Learned* is *learn* plus the inflectional affix *–ed*
- **Syntax: the way words are used to form phrases**
  - It is part of English syntax that a determiner such as *the* will come before a noun, and also that determiners are obligatory with certain singular noun.
- Semantics: Compositional and lexical semantics
  - Compositional semantics: the construction of meaning based on syntax
  - Lexical semantics: the meaning of individual words
- Pragmatics: meaning in context
  - *Do you have the time?* – means *'can you tell me what time is it now?'*

LECTURE PLAN

Lecture 5: Chatbot and Language Fundamental

1. Machine Learning and NLP: Finish
2. Chatbot
3. Language Fundamental
   - Phonology, Morphology, Syntax, Semantics, Pragmatics
4. **Text Preprocessing**
   1. Tokenization
   2. Cleaning and Normalisation
   3. Stemming and Lemmatisation
   4. Stopword
   5. Regular Expression

Text Preprocessing

Text Preprocessing

- Every NLP task needs to do text pre-processing
  - Segmenting/tokenizing words in running text
  - Normalizing word formats
  - Segmenting sentences in running text

Text Preprocessing

How many words?

they lay back on the Sydney grass and looked at the stars and their

- Type: an element of the vocabulary.
- Token: an instance of that type in running text.
- How many?
  - 14 tokens
  - 13 types (or 12) (or 11?)

Text Preprocessing

How many words?

- ***N*** = number of tokens
- ***V*** = vocabulary = set of types
  - $|V|$ is the size of the vocabulary

Church and Gale (1990): $|V| > O(N^{½})$

	Tokens = N	Types = |V|
Switchboard phone conversations	2.4 million	20 thousand
Shakespeare	884,000	31 thousand
Google N-grams	1 trillion	13 million

Text Preprocessing

Normalization

- Need to “normalize” terms
  - Information Retrieval: indexed text & query terms must have same form.
    - We want to match U.S.A. and USA
- We implicitly define equivalence classes of terms
  - e.g., deleting periods in a term
- Alternative: asymmetric expansion:
  - Enter: window Search: window, windows
  - Enter: windows Search: Windows, windows, window
  - Enter: Windows Search: Windows
- Potentially more powerful, but less efficient

Text Preprocessing

Case Folding

- Applications like IR: reduce all letters to lower case
  - Since users tend to use lower case
  - Possible exception: upper case in mid-sentence?
    - e.g., General Motors
    - Fed vs. fed
    - SAIL vs. sail
- For sentiment analysis, MT, Information extraction
  - Case is helpful (US versus us is important)

Text Preprocessing

Lemmatization

- Reduce inflections or variant forms to base form
  - am, are, is → be
  - car, cars, car's, cars' → car
- *the boy's cars are different colors → the boy car be different color*
- Lemmatization: have to find correct dictionary headword form
- Machine translation
  - Spanish quiero ('I want'), quieres ('you want') same lemma as querer 'want'

Text Preprocessing

Morphology

- Morphemes:
  - The small meaningful units that make up words
  - **Stems**: The core meaning-bearing units
  - **Affixes**: Bits and pieces that adhere to stems
  - Often with grammatical functions

Text Preprocessing

Stemming

- Reduce terms to their stems in information retrieval
- Stemming is crude chopping of affixes
  - language dependent
  - *e.g., automate(s), automatic, automation all reduced to automat.*

for example compressed and compression are both accepted as equivalent to compress.

for exampl compress and compress ar both accept as equival to compress

Text Preprocessing

Porter's algorithm: The most common English stemmer

Step 1a

sses	→ ss	caresses	→ caress
ies	→ i	ponies	→ poni
ss	→ ss	caress	→ caress
s	→ ø	cats	→ cat

Step 1b

(*v*)ing	→ ø	walking	→ walk
		sing	→ sing
(*v*)ed	→ ø	plastered	→ plaster
…			

Step 2 (for long stems)

ational	→ ate	relational	→ relate
izer	→ ize	digitizer	→ digitize
ator	→ ate	operator	→ operate
…			

Step 3 (for longer stems)

al	→ ø	revival	→ reviv
able	→ ø	adjustable	→ adjust
ate	→ ø	activate	→ activ
…			

Text Preprocessing

Viewing morphology in a corpus
Why only strip –ing if there is a vowel?

(*v*)ing → ø walking → walk
sing → sing

Text Preprocessing

Viewing morphology in a corpus
Why only strip –ing if there is a vowel?

(*v*)ing → ø walking → walk
sing → sing

```
tr -sc 'A-Za-z' '\n' < shakes.txt | grep 'ing$' | sort | uniq -c | sort -nr
```

```
1312 King
 548 being
 541 nothing
 388 king
 375 bring
 358 thing
 307 ring
 152 something
 145 coming
 130 morning
```

```
548 being
541 nothing
152 something
145 coming
130 morning
122 having
120 living
117 loving
116 Being
102 going
```

```
tr -sc 'A-Za-z' '\n' < shakes.txt | grep '[aeiou].*ing$' | sort | uniq -c | sort -nr
```

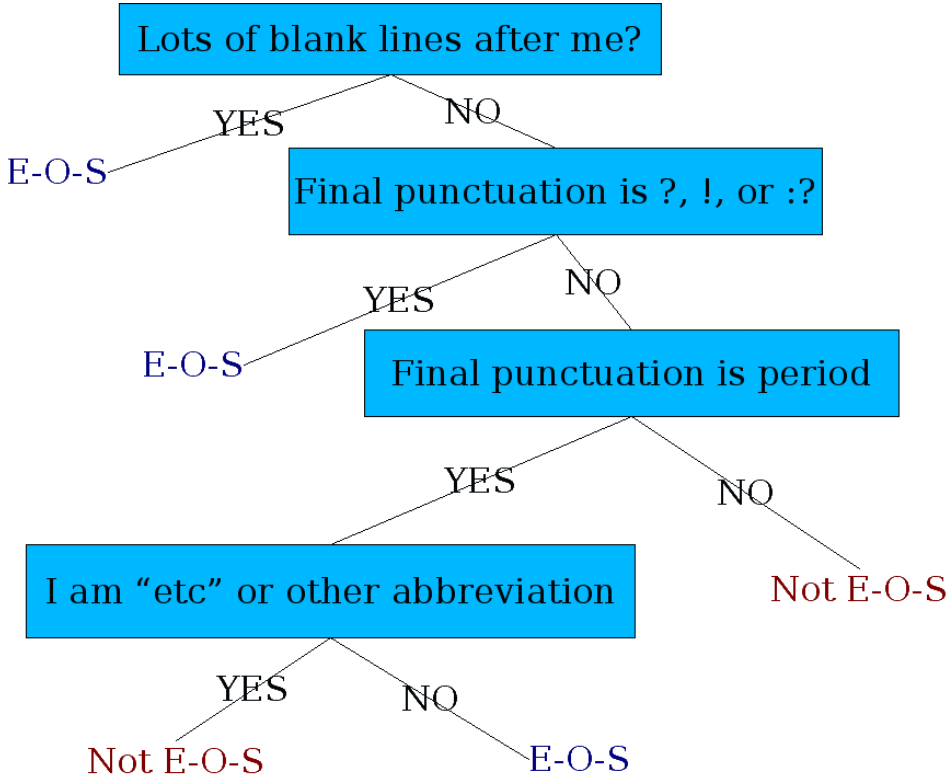
Text Preprocessing

Sentence Segmentation

- !, ? are relatively unambiguous
- Period "." is quite ambiguous
  - Sentence boundary
  - Abbreviations like Inc. or Dr.
  - Numbers like .02% or 4.3
- Build a binary classifier
  - Looks at a "."
  - Decides EndOfSentence/NotEndOfSentence
  - Classifiers: hand-written rules, regular expressions, or machine-learning

Text Preprocessing

Determining if a word is end-of-sentence: a Decision Tree



Text Preprocessing

More sophisticated decision tree features

- Case of word with ".": Upper, Lower, Cap, Number
- Case of word after ".": Upper, Lower, Cap, Number

- Numeric features
  - Length of word with "."
  - Probability(word with "." occurs at end-of-s)
  - Probability(word after "." occurs at beginning-of-s)

Text Preprocessing

Implementing Decision Trees

- A decision tree is just an if-then-else statement
- The interesting research is choosing the features
- Setting up the structure is often too hard to do by hand
  - Hand-building only possible for very simple features, domains
    - For numeric features, it's too hard to pick each threshold
  - Instead, structure usually learned by machine learning from a training corpus

Text Preprocessing

Decision Trees and other classifiers

- We can think of the questions in a decision tree
- As features that could be exploited by any kind of classifier
  - Logistic regression
  - SVM
  - Neural Nets
  - etc.

Text Preprocessing

Regular expressions

- A formal language for specifying text strings
- How can we search for any of these?
  1. woodchuck
  2. woodchucks
  3. Woodchuck
  4. Woodchucks

Text Preprocessing

Regular Expressions: Disjunctions

- Letters inside square brackets []

Pattern	Matches
`[wW]oodchuck`	Woodchuck, woodchuck
`[1234567890]`	Any digit

- Ranges [A-Z]

Pattern	Matches	
`[A-Z]`	An upper case letter	`Drenched Blossoms`
`[a-z]`	A lower case letter	`my beans were impatient`
`[0-9]`	A single digit	`Chapter 1: Down the Rabbit Hole`

Text Preprocessing

Regular Expressions: Negation in Disjunction

- Negations [^Ss]
  - Carat means negation only when first in []

Pattern	Matches	
[^A-Z]	Not an upper case letter	Oyfn pripetchik
[^Ss]	Neither 'S' nor 's'	I have no exquisite reason"
[^e^]	Neither e nor ^	Look here
a^b	The pattern a carat b	Look up a^b now

Text Preprocessing

Regular Expressions: More Disjunction

- Woodchucks is another name for groundhog!
- The pipe | for disjunction

Pattern	Matches
`groundhog|woodchuck`	
`yours|mine`	yours mine
`a|b|c`	= [abc]
`[gG]roundhog|[Ww]oodchuck`	

Text Preprocessing

Regular Expressions: ? * + .

Pattern	Matches	
`colou?r`	Optional previous char	color colour
`oo*h!`	0 or more of previous char	oh! ooh! oooh! ooooh!
`o+h!`	1 or more of previous char	oh! ooh! oooh! ooooh!
`baa+`		baa baaa baaaa baaaaa
`beg.n`		begin begun begun beg3n

Text Preprocessing

Regular Expressions: Anchors ^ $

Pattern	Matches
^[A-Z]	Palo Alto
^[^A-Za-z]	1 "Hello"
\.$	The end.
.$	The end? The end!

Text Preprocessing

Example

- Find me all instances of the word “the” in a text.

the — ***Misses capitalized examples***

[tT]he — ***Incorrectly returns*** *other* ***or*** *theology*

[^a-zA-Z]**[tT]he**[^a-zA-Z]

Text Preprocessing

Example

- The process we just went through was based on fixing two kinds of errors.
  - Matching strings that we should not have matched (**the**re, **the**n, other)

    False positives (Type I)

  - Not matching things that we should have matched (The)

    False negatives (Type II)

Text Preprocessing

Errors cont.

- In NLP we are always dealing with these kinds of errors.
- Reducing the error rate for an application often involves two antagonistic efforts:
  - *Increasing accuracy or precision (minimizing false positives)*
  - *Increasing coverage or recall (minimizing false negatives).*

Text Preprocessing

Summary

- Regular expressions play a surprisingly large role
  - Sophisticated sequences of regular expressions are often the first model for any text processing text
- For many hard tasks, we use machine learning classifiers
  - But regular expressions are used as features in the classifiers
  - Can be very useful in capturing generalizations

Reference

Summary

- Deng, L., & Liu, Y. (Eds.). (2018). Deep Learning in Natural Language Processing. Springer.
- Rao, D., & McMahan, B. (2019). Natural Language Processing with PyTorch: Build Intelligent Language Applications Using Deep Learning. " O'Reilly Media, Inc.".
- Manning, C. D., Manning, C. D., & Schütze, H. (1999). Foundations of statistical natural language processing. MIT press.
- Blunsom, P 2017, Deep Natural Language Processing, lecture notes, Oxford University
- Manning, C 2017, Natural Language Processing with Deep Learning, lecture notes, Stanford University
- Jurafsky, D 2018, From Language to Information, lecture notes, Stanford University
- Serban, Iulian V., Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. 2015. "Building End-To-End Dialogue Systems Using Generative Hierarchical Neural Network Models.