

**COMP5046**  
**Natural Language Processing**

Lecture 5: Assignment1 and Language Fundamental

Dr. Caren Han  
Semester 1, 2021  
School of Computer Science,  
University of Sydney

1

**0 LECTURE PLAN**

**Lecture 5: Assignment1 and Language Fundamental**

1. RNN/LSTM, Dealing Context Review
2. Assignment 1 Discussion
3. Sentiment Analysis
  1. Sentiment Analysis Overview
  2. Assignment Specification
4. Language Fundamental
  - Phonology, Morphology, Syntax, Semantics, Pragmatics
5. Text Preprocessing
  1. Tokenization
  2. Cleaning and Normalisation
  3. Stemming and Lemmatisation
  4. Stopword
  5. Regular Expression

2

**1 RNN/LSTM Review**

Neural Network + Memory = Recurrent Neural Network

Legend:

- $\sigma$  Tanh function
- $h_{t-1}$  new hidden state
- $h_t$  previous hidden state
- $x_t$  input
- concatenation

Hidden Layer

Input Layer

$h_t = \tanh(W_{hh}h_{t-1} + W_{hx}x_t + b_h)$

New hidden state  $h_t$  is a function of previous state  $h_{t-1}$  and input  $x_t$  with parameters  $W$ .

3

**1 RNN/LSTM Review**

**LSTM (Long Short-Term Memory) – Forget Gate**

$f_t = \sigma(W_f[h_{t-1}, x_t] + b_f)$

**Decides what information should be thrown away or kept**  
Information from the **previous hidden state** and information from the **current input** is passed through the **sigmoid function**. Values come out between 0 and 1. The closer to 0 means to forget, and the closer to 1 means to keep.

4

1 RNN/LSTM Review

### LSTM (Long Short-Term Memory) – Input Gate

$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i)$   
 $\tilde{C}_t = \tanh(W_C[h_{t-1}, x_t] + b_C)$

1. Pass the previous hidden state and current input into a sigmoid function
2. Pass the hidden state and current input into the tanh function to squish values between -1 and 1 to help regulate the network
3. Multiply the tanh output with the sigmoid output

\*sigmoid output will decide which information is important to keep from the tanh output

5

1 RNN/LSTM Review

### LSTM (Long Short-Term Memory) – Cell States

$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$

- the cell state gets pointwise multiplied by the forget vector
- take the output from the input gate and do a pointwise addition which updates the cell state to new values that the neural network finds relevant
- That gives us our new cell state

6

1 RNN/LSTM Review

### LSTM (Long Short-Term Memory) – Output Gate

$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o)$   
 $h_t = o_t * \tanh(C_t)$

**decides what the next hidden state should be.**

- pass the previous hidden state and the current input into a sigmoid function
- pass the newly modified cell state to the tanh function
- multiply the tanh output with the sigmoid output to decide what information the hidden state should carry

7

1 Dealing Context: Review

### V to V' – Projection with Context (1)

Context      Data

8

1 Dealing Context: Review

V to V' – Projection with Context (2)

Context:  $c_1, c_2, c_3, c_4$

Data:  $x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9$

Equation:  $V \xrightarrow{W^C} I + \xrightarrow{W} II = V'$

9

1 Dealing Context: Review

V to V' with Context - Linear Algebra

[1 x 9] matrix:  $x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9$

[9 x 2] matrix:  $w_{1,1}, w_{1,2}, w_{2,1}, w_{2,2}, w_{3,1}, w_{3,2}, w_{4,1}, w_{4,2}, w_{5,1}, w_{5,2}, w_{6,1}, w_{6,2}, w_{7,1}, w_{7,2}, w_{8,1}, w_{8,2}, w_{9,1}, w_{9,2}$

[1 x 4] matrix:  $c_1, c_2, c_3, c_4$

[4 x 2] matrix:  $w^{c1,1}, w^{c1,2}, w^{c2,1}, w^{c2,2}, w^{c3,1}, w^{c3,2}, w^{c4,1}, w^{c4,2}$

[1 x 2] matrix:  $I, II$

10

1 Dealing Context: Review

V to V' with Context - Linear Algebra (Simplified)

[1 x (9+4)] matrix:  $x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, c_1, c_2, c_3, c_4$

[(9+4) x 2] matrix:  $w_{1,1}, w_{1,2}, w_{2,1}, w_{2,2}, w_{3,1}, w_{3,2}, w_{4,1}, w_{4,2}, w_{5,1}, w_{5,2}, w_{6,1}, w_{6,2}, w_{7,1}, w_{7,2}, w_{8,1}, w_{8,2}, w_{9,1}, w_{9,2}, w^{c1,1}, w^{c1,2}, w^{c2,1}, w^{c2,2}, w^{c3,1}, w^{c3,2}, w^{c4,1}, w^{c4,2}$

[1 x 2] matrix:  $I, II$

11

1 Dealing Context: Review

$V \rightarrow V' \rightarrow 1$

$V = \begin{bmatrix} 10 & 2 & 8 \\ 2 & 15 & 3 \\ 5 & 1 & 5 \end{bmatrix}$

$V' = \begin{bmatrix} ? \\ ? \\ ? \end{bmatrix}$

$1 = ?$

Neural network diagram: 9 input nodes, 2 hidden nodes, 1 output node.

12

**0 LECTURE PLAN**

**Lecture 5: Assignment1 and Language Fundamental**

1. RNN/LSTM Review
2. **Assignment 1 Discussion**
3. Sentiment Analysis
  1. Sentiment Analysis Overview
  2. Assignment Specification
4. Language Fundamental
  - Phonology, Morphology, Syntax, Semantics, Pragmatics
5. Text Preprocessing
  1. Tokenization
  2. Cleaning and Normalisation
  3. Stemming and Lemmatisation
  4. Stopword
  5. Regular Expression

13

**2 Assignment 1 Discussion**

$V_s \rightarrow V's \rightarrow V'$

Reflect Data 2 (the present) and the past Data

Reflect Data 3 (the present) and the past Data

Reflect all information

14

**2 Assignment 1 Discussion**

$V_s \rightarrow V's \rightarrow V'$

Reflect Data 2 (the present) and the past Data

Reflect Data 3 (the present) and the past Data

Reflect all information

15

**2 Assignment 1 Discussion**

**RNN**

**N to 1 Task**

Positive

Neg

softmax

$x_1$   $x_2$   $x_3$   $x_4$   $x_5$

$h_0$   $h_1$   $h_2$   $h_3$   $h_4$   $h_5$

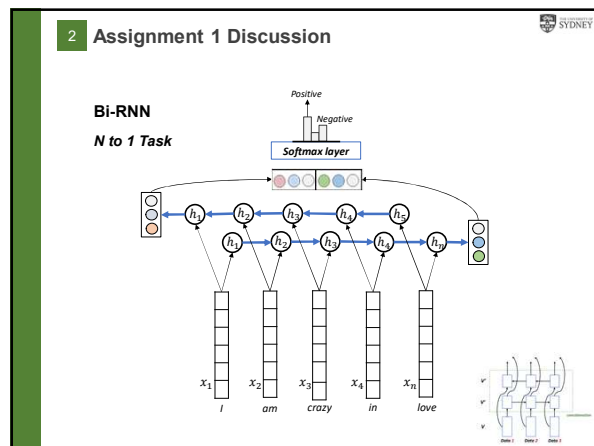
$W_{hh}$   $W_{hh}$   $W_{hh}$   $W_{hh}$

$W_{xh}$   $W_{xh}$   $W_{xh}$   $W_{xh}$   $W_{xh}$

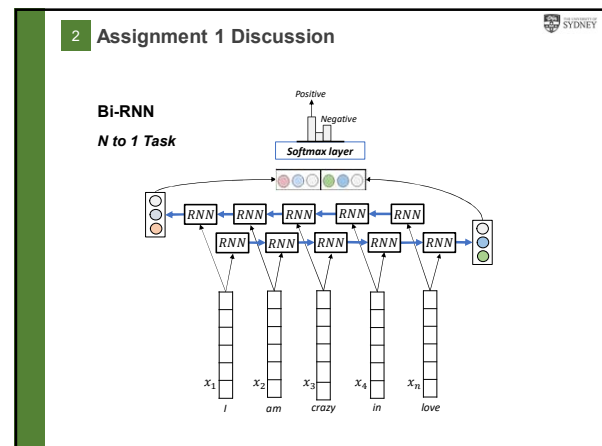
$l$   $am$   $crazy$   $in$   $love$

$\tanh$

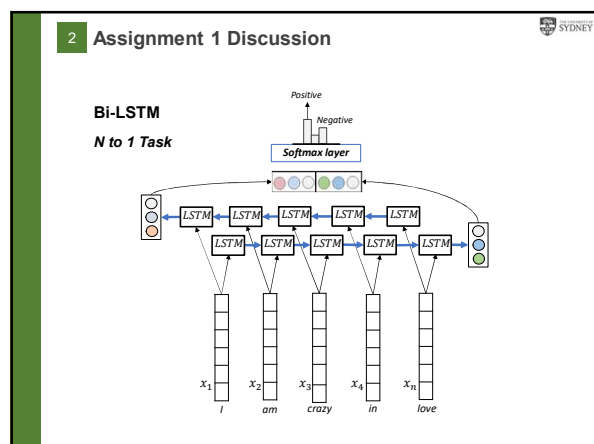
16



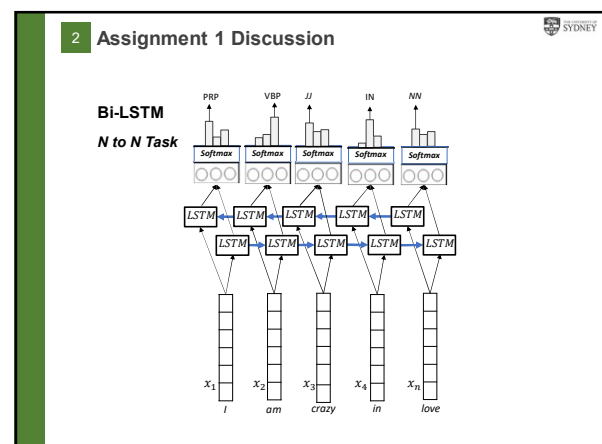
17



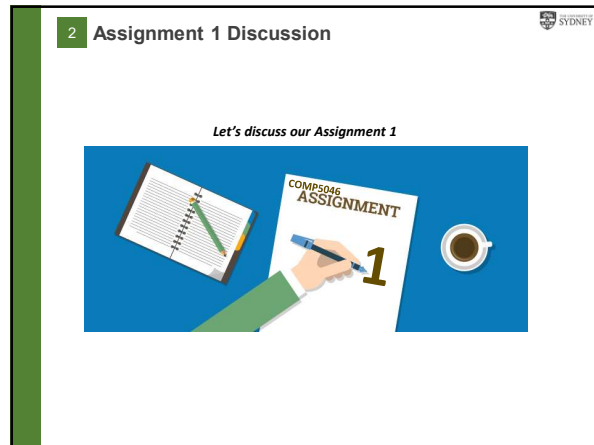
18



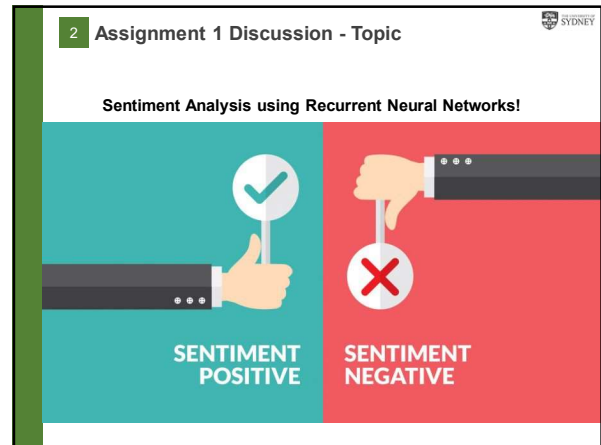
19



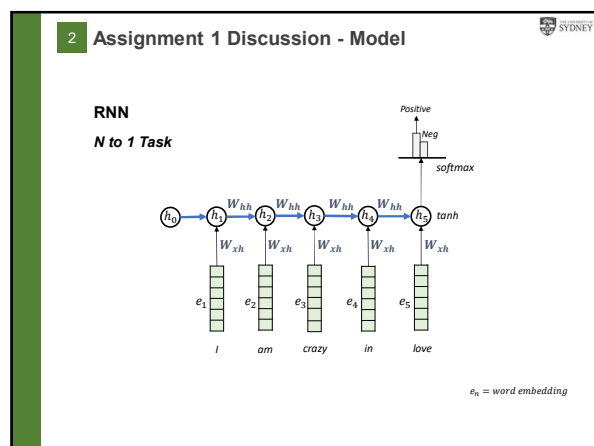
20



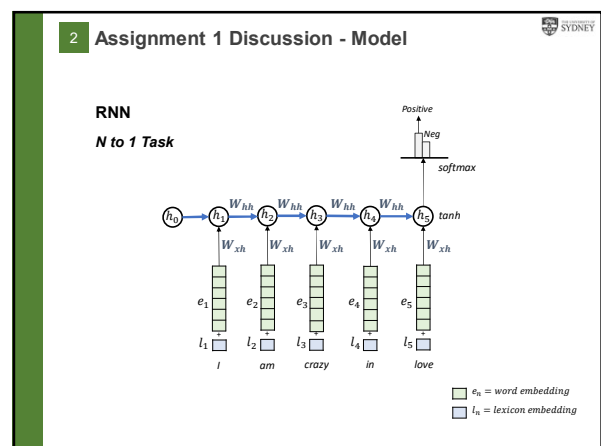
21



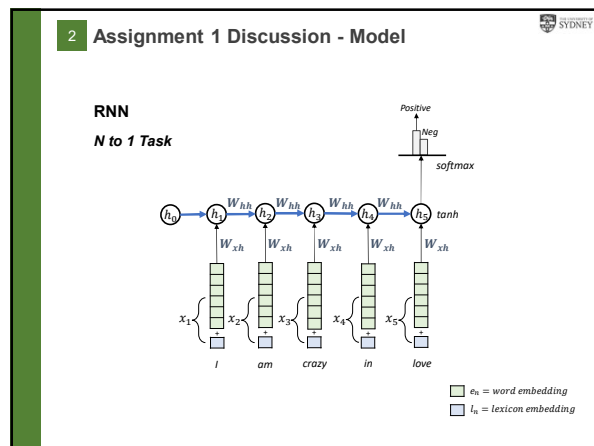
22



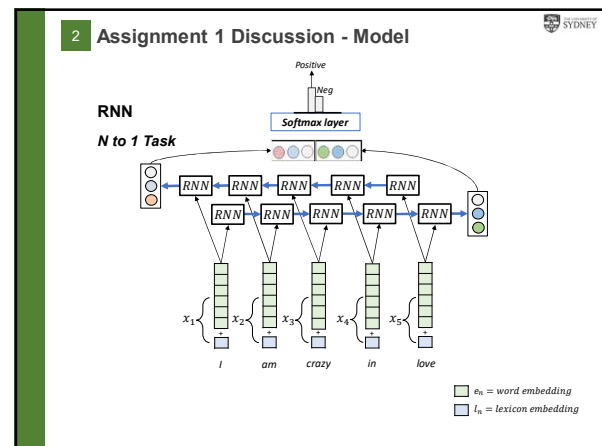
23



24



25



26

**2 Assignment 1 Discussion - Model**

Assignment 1 Specification can be found in <https://github.com/usydnlp/COMP5046>

27

**0 LECTURE PLAN**

**Lecture 5: Assignment1 and Language Fundamental**

1. RNN/LSTM, Dealing Context Review
2. Assignment 1 Discussion
3. Sentiment Analysis
  1. Sentiment Analysis Overview
  2. Assignment Specification
4. Language Fundamental
  - Phonology, Morphology, Syntax, Semantics, Pragmatics
5. Text Preprocessing
  1. Tokenization
  2. Cleaning and Normalisation
  3. Stemming and Lemmatisation
  4. Stopword
  5. Regular Expression

28

### 3 The NLP Big Picture

**The purpose of Natural Language Processing: Overview**

The diagram illustrates the NLP Big Picture, showing the relationship between Applications and the NLP Stack.

**Applications:**

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

**NLP Stack:**

- Entity Extraction:** When Sebastian Thrun ... When Sebastian Thrun **was** started at **Google** in **2007** **was**
- 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]

29

### 3 Sentiment Analysis

**Movie Review – Positive or Negative**

The screenshot shows a movie review interface with sentiment analysis overlays. The reviews are categorized as Positive (thumbs up) or Negative (thumbs down).

**Reviews:**

- Positive:** "Best for the best movie I have ever seen" (10/10)
- Positive:** "A classic piece of unforgettable film-making." (10/10)
- Negative:** "Terribly Overrated" (2/10)
- Negative:** "After a big buildup, a real disappointment" (2/10)

**Too easy?** 😊

30

### 3 Sentiment Analysis

**What is Sentiment Analysis?**

The screenshot shows a movie review interface with sentiment analysis overlays. The reviews are categorized as Positive (thumbs up) or Negative (thumbs down).

**Reviews:**

- Positive:** "Best for the best movie I have ever seen" (10/10)
- Positive:** "A classic piece of unforgettable film-making." (10/10)
- Negative:** "Terribly Overrated" (2/10)
- Negative:** "After a big buildup, a real disappointment" (2/10)

**Examples of Sentiment Analysis:**

- 100% predictable:** "So obvious that the ship would sink." (0 of 3 people found this review helpful)
- One Star:** "There were no wolves in the movie." (0 of 3 people found this review helpful)
- The snowman keeps falling apart:** "The snowman keeps falling apart." (5 of 12 people found this review helpful)

31

### 3 Sentiment Analysis

**What is Sentiment Analysis?**

"Sentiment analysis is the operation of **understanding the intent or emotion behind a given piece of text**. It is part of text classification, but it is useful for extracting structured information"

**Different Names of a 'Sentiment Analysis'**


- Opinion extraction
- Opinion mining
- Sentiment mining
- Subjectivity analysis

32



### 3 Sentiment Analysis

#### Sentiment Analysis



Customer reviews

4.6 out of 5

13,551 customer ratings

5 star 79%

4 star 9%

3 star 2%

2 star 1%

1 star 4%

Write a customer review

See all 24 reviews

33

### 3 Sentiment Analysis

#### What is Sentiment Analysis?

Emotion, Mood, Interpersonal stances, **Attitude**, Personality traits

*Typology of Affective States (Scherer et al. 2006)*

**Attitudes**

Enduring, affectively colored beliefs, **dispositions towards objects/persons**

- liking, loving, hating, valuing, desiring

Scherer, K., Dan, E., & Frijol, A. (2006). What determines a feeling's position in affective space? A case for appraisal. *Cognition & Emotion*, 20(1), 93-113.

34

### 3 Sentiment Analysis

#### Sentiment Analysis: Examples

##### Apple iPhone 7 - 128GB - Rose Gold (Unlocked)

4.6 out of 5

30 product ratings

Write a review

Aspects

92% Excellent

84% Longevity

51% Performance

Most relevant reviews

Excellent phone

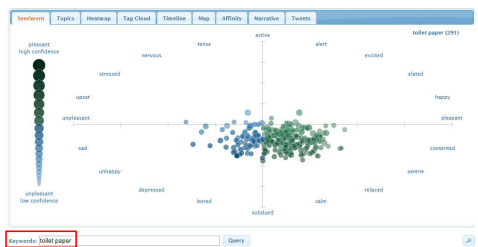
Really good for price

Good practical iPhone

35

### 3 Sentiment Analysis

#### Sentiment Analysis: Sentiment viz



Keywords: toilet paper

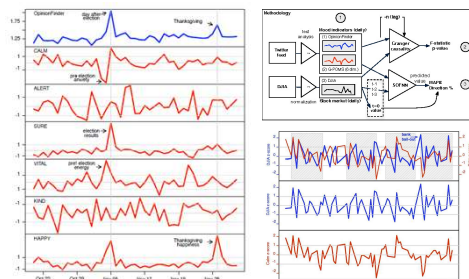
[https://www.csc2.ncsu.edu/faculty/healey/tweet\\_vis/tweet\\_app/](https://www.csc2.ncsu.edu/faculty/healey/tweet_vis/tweet_app/)

36

### 3 Sentiment Analysis

#### Sentiment Analysis: Examples

Twitter mood predicts the stock market (Bollen et al. 2011)



37

### 3 Sentiment Analysis

#### Sentiment Analysis Tasks

- **Movie:** Is this review positive or negative?
- **Products:** what do people think about the new phone?
- **Public sentiment:** how is consumer confidence? Is despair increasing?
- **Politics:** what do people think about this candidate or issue?
- **Prediction:** predict election outcomes or market trends from sentiment

38

### 3 Sentiment Analysis

#### What will be considered to analyse sentiment

**Sentiment analysis = the detection of Attitudes**

Enduring, affectively colored beliefs, dispositions towards objects/persons

#### Main Factors

- **Target Object:** an entity that can be a product, person, event, organisation, or topic (e.g. iPhone)
- **Attribute:** an object usually has two types of attributes
  - Components (e.g. touch screen, battery)
  - Properties (e.g. size, weight, colour, voice quality)
  - Explicit and implicit attributes:
    - Explicit attributes: appearing in the attitude (e.g. "the battery life of this phone was not long")
    - Implicit attributes: not appearing in the attitude (e.g. "this phone is too expensive" – the property price)
- **Attitude Holder:** the person or organisation that expresses the opinion (e.g. my mother was mad with me)
- **Type of attitude:** positive, negative, or neutral or set of types (e.g. happy)
- **Time:** the time that expresses the opinion

39

### 3 Sentiment Analysis

#### What is Sentiment Analysis?

- **Basic Task:** Is the attitude of this text positive or negative?



- **More complex task:** Rank the attitude of this text from 1 to 5

Likert Scale (1 to 5)


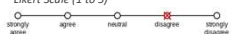


- **Advanced task:** Detect the target, source, or complex attitude types

40

### 3 Sentiment Analysis

**What is Sentiment Analysis?**

- Basic Task: Is the attitude of this text **positive or negative**?  

- More complex task: Rank the attitude of this text from 1 to 5  
**Likert Scale (1 to 5)**  

- Advanced task: Detect the target, source, or complex attitude types

41

### 3 Sentiment Analysis

**Finding aspect/attribute/target of sentiment**

**Title: Sharp, Solid, but Harder to Hold than iPhone 7**  
 - By Tristan on March 13, 2017

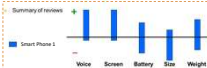
"my thoughts on the iPhone 7 are:  
 1) Retina display is awesome. Everything looks more defined and sharper. There is much color and clarity out there... or should I say, in those digital images and videos... needless to say, the camera as well captures great images."  
 ...."

**Attribute based Summary**


- Attribute 1: display
  - Positive
    1. Retina display is awesome
    2. There is much color and clarity out there
    3. —
- Attribute 2: camera
  - Positive
    1. the camera as well captures great images.
    2. —

**Attribute based Visualisation**

Summary of reviews



Comparison of reviews



42

### 3 Sentiment Analysis

**Features Vectors: a bird's eye view**

- Word ngrams (up to 4), skip ngrams w/ 1 missing word
- Character ngrams up to 5
- All caps: number of words in capitals
- Number of continuous punctuation marks, either exclamation or question or mixed. Also whether last char contains one of these.
- Presence of emoticons

**Classify your Sentiment is a classification problem**

- Typically people have used **Naïve Bayes** or **Support Vector Machines (SVM)** in the past [Mohammad et al. 2013]
- Artificial Neural Nets** are also becoming more popular now [Nogueira dos Santos & Gatti, 2014]

43

### 3 Sentiment Analysis

**Useful Sentiment Lexicons**

Name	Details
The General Inquirer <a href="http://www.cis.upenn.edu/~c3/c3/GeneralInquirer.html">http://www.cis.upenn.edu/~c3/c3/GeneralInquirer.html</a>	Categories <ul style="list-style-type: none"> <li>Positive (1515 words) and Negative (2281 words)</li> <li>Strong vs Weak, Active vs Passive, Overstated versus Understated</li> <li>Pleasure, Pain, Virtue, Vice, Motivation, Cognitive Orientation, etc</li> </ul> Free to use
LIWC Linguistic Inquiry and Word Count <a href="http://www.uconn.edu/liwc/">http://www.uconn.edu/liwc/</a>	2300 words and less than 70 classes <ul style="list-style-type: none"> <li>Affective Processes               <ul style="list-style-type: none"> <li>negative emotion (bad, weird, hate, problem, tough)</li> <li>positive emotion (love, nice, sweet)</li> </ul> </li> <li>Cognitive Processes               <ul style="list-style-type: none"> <li>Tentative (maybe, perhaps, guess), Inhibition (block, constraint)</li> <li>Pronouns, Negation (no, never), Quantifiers (few, many)</li> </ul> </li> </ul> \$30 or \$90 fee
MPQA Subjectivity Cues Lexicon <a href="http://www.cba.hawaii.edu/~mccoy/subjectivity/">http://www.cba.hawaii.edu/~mccoy/subjectivity/</a>	Each word annotated for intensity (strong, weak) 6885 words from 8221 lemmas <ul style="list-style-type: none"> <li>2718 positive</li> <li>4912 negative</li> </ul> GNU GPL (widely-used free software license)
Opinion Lexicon <a href="http://www.cba.hawaii.edu/~mccoy/opinion-lexicon-english.txt">http://www.cba.hawaii.edu/~mccoy/opinion-lexicon-english.txt</a>	6786 words <ul style="list-style-type: none"> <li>2006 positive/ 4783 negative</li> </ul> Free to use
SentWordNet <a href="http://www.net-net.net/">http://www.net-net.net/</a>	All WordNet synsets automatically annotated for degrees of positivity, negativity, and neutrality/objectiveness <ul style="list-style-type: none"> <li>[estimable(1,3)] "may be computed or estimated"</li> <li>Pos 0 Neg 0 Obj 1</li> <li>[estimable(1,1)] "deserving of respect or high regard"</li> <li>Pos .75 Neg 0 Obj .25</li> </ul> Free to use

44

### 3 Sentiment Analysis

Can you build the sentiment lexicon by yourself?

**Bootstrap style: Semi-supervised learning of lexicons**

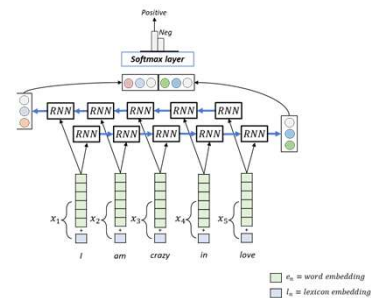
- Use a small amount of information
- A few labeled examples
- A few hand-built patterns
- Bootstrapping a lexicon

[amazonmechanical turk](#)

45

### 3 Sentiment Analysis

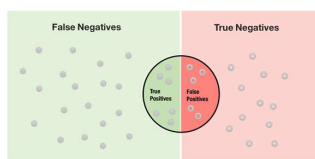
**Assignment 1: Sentiment Analysis**



46

### 3 Sentiment Analysis

**Assignment 1: Sentiment Analysis**



$$\begin{aligned} \text{precision} &= \frac{TP}{TP + FP} \\ \text{recall} &= \frac{TP}{TP + FN} \\ F1 &= \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \\ \text{accuracy} &= \frac{TP + TN}{TP + FN + FP} \\ \text{specificity} &= \frac{TN}{TN + FP} \end{aligned}$$

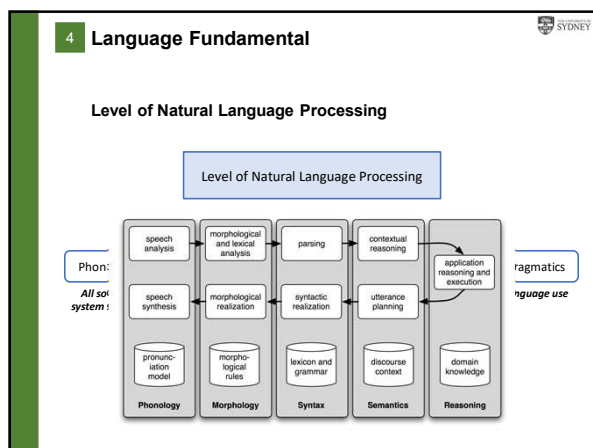
47

### 0 LECTURE PLAN

**Lecture 5: Assignment1 and Language Fundamental**

1. RNN/LSTM, Dealing Context Review
2. Assignment 1 Discussion
3. Sentiment Analysis
  1. Sentiment Analysis Overview
  2. Assignment Specification
4. **Language Fundamental**
  - Phonology, Morphology, Syntax, Semantics, Pragmatics
5. Text Preprocessing
  1. Tokenization
  2. Cleaning and Normalisation
  3. Stemming and Lemmatisation
  4. Stopword
  5. Regular Expression

48



49

4 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) does not appear at the beginning of a word
- Does this mean that [ŋ] cannot appear at the beginning of words in all human languages?

NO! — Nguyen Tran      NO! — Andrew Ng

50

4 Language Fundamental

We know how sounds can combine

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

51

4 Language Fundamental

Morphology: Pieces of words

- A field of linguistics focused on the study of the **forms and formation of words in a language**
- Words in a language consist of one element or elements of meaning which are **morphemes**
  - Morphemes** are the pieces of words: bases, roots and affixes (pre-fix, suffix).

The diagram shows the morphological structure of the word "unacceptable":

- un-** (Prefix, a type of affix): A group of letters placed before the root word.
- accept** (Root word): The central morpheme, the key element.
- able** (Suffix, a type of affix): A group of letters placed after the root word.

The word "unacceptable" is shown as the combination of these three parts.

52

**4 Language Fundamental**

**Morphology: Pieces of words**

- A field of linguistics focused on the study of the *forms and formation of words in a language*
- Words in a language consist of one element or elements of meaning which are **morphemes**
  - Morphemes** are the pieces of words: bases, roots and affixes.
- walk walked walking walks walk walk -ed walk -ing walk -s

53

**4 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?’

54

**0 LECTURE PLAN**

**Lecture 5: Assignment1 and Language Fundamental**

- RNN/LSTM, Dealing Context Review
- Assignment 1 Discussion
- Sentiment Analysis
  - Sentiment Analysis Overview
  - Assignment Specification
- Language Fundamental
  - Phonology, Morphology, Syntax, Semantics, Pragmatics
- Text Preprocessing**
  - Tokenization
  - Cleaning and Normalisation
  - Stemming and Lemmatisation
  - Stopword
  - Regular Expression

55

**5 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

56

## 5 Text Preprocessing

### How many words?

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

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

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

57

## 5 Text Preprocessing

### How many words?

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

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

58

## 5 Text Preprocessing

### Tokenization: language issues

- French**
  - L'ensemble → one token or two?
    - L ? L' ? Le ?
  - Want l'ensemble to match with un ensemble
    - Until 2003, Google cannot make this work
- German noun compounds are not segmented**
  - Lebensversicherungsgesellschaftsangestellter*
  - 'life insurance company employee'
  - German information retrieval needs *compound splitter*

59

## 5 Text Preprocessing

### Tokenization: language issues

- Chinese has no spaces between words:
  - 悉尼大学位于澳大利亚悉尼
  - 悉尼大学 位于 澳大利亚 悉尼
  - University of Sydney is located in Sydney, Australia
- Further complicated in Japanese, with multiple alphabets intermingled
  - Dates/amounts in multiple formats

フォーチュン500社は情報不足のため時間あた\$500K(約6,000万円)

Katakana      Hiragana      Kanji      Romaji

60

## 5 Text Preprocessing

### Tokenization: language issues

- Arabic (or Hebrew) is basically written right to left, but with certain items like numbers written left to right
- Words are separated, but letter forms within a word form complex ligatures

← → ← → ← start  
استقلت الجزائر في سنة 1962 بعد 132 عام من الاحتلال الفرنسي.

- 'Algeria achieved its independence in 1962 after 132 years of French occupation.'
- With Unicode, the order of characters in files matches the conceptual order, and the reversal of displayed characters is handled by the rendering system.

61

## 5 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

62

## 5 Text Preprocessing

### Case Folding

- Applications like IR: **convert 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, Machine Translation, Information extraction
  - Case is helpful (US versus us is important)

63

## 5 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'

64



## 5 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

65

## 5 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.



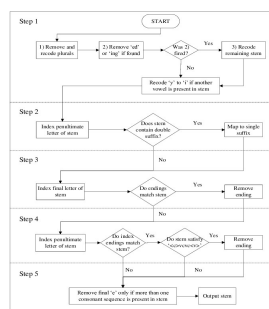
for exampl compress and  
compress ar both accept  
as equival to compress

66

## 5 Text Preprocessing

### Porter's algorithm: The most common English stemmer

#### Porter Stemming Algorithm



67

## 5 Text Preprocessing

### Dealing with complex morphology is sometimes necessary

- Some languages requires complex morpheme segmentation
  - Turkish
  - Uygarlastıramadıklarımızdanmissinizcasına
  - '(behaving) as if you are among those whom we could not civilize'
  - Uygar 'civilized' + las 'become'
    - + tir 'cause' + ama 'not able'
    - + dik 'past' + lar 'plural'
    - + imiz 'p1pl' + dan 'abl'
    - + mis 'past' + siniz '2pl' + casına 'as if'

68

## 5 Text Preprocessing

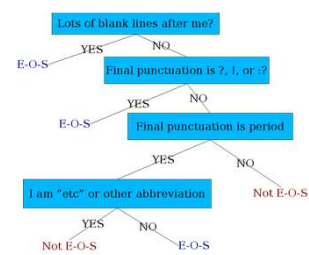
### Sentence Segmentation

- !, ? 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

69

## 5 Text Preprocessing

### Sentence Segmentation using a Decision Tree



70

## 5 Text Preprocessing

### Implementing Decision Trees or other classifiers

- 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
- As features that could be exploited by any kind of classifier
  - Logistic regression
  - SVM
  - Neural Nets
  - etc.

71

## 5 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



72

5

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
[a-z]	A lower case letter
[0-9]	A single digit

73

5

Text Preprocessing



Regular Expressions: Negation in Disjunction

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


Pattern	Matches
[^A-Z]	Not an upper case letter
[^Ss]	Neither 'S' nor 's'
[^e^]	Neither e nor ^
a^b	The pattern 'a caret b'

- Caret means negation only when showing as the first symbol in []

74

5


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	



75


5

Text Preprocessing



Regular Expressions: ? \* + .

Pattern	Matches
colou?r	Optional previous char
oo*h!	0 or more of previous char
o+h!	1 or more of previous char
baa+	
beg.n	




Stephen C Kleene

Kleene \*, Kleene +

76

5

Text Preprocessing




Regular Expressions: Anchors ^ \$

Pattern	Matches
<code>^[A-Z]</code>	<u>P</u> alo Alto
<code>^[^A-Za-z]</code>	<u>"</u> Hello"
<code>\.\$</code>	The end <u>.</u>
<code>.\$</code>	The end <u>?</u> The end! <u>.</u>

77

5

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

78

/

Reference



Reference

- Serban, Iulian V., Alessandro Sordani, Yoshua Bengio, Aaron Courville, and Joelle Pineau. 2015. "Building End-To-End Dialogue Systems Using Generative Hierarchical Neural Network Models."

79