

COMP5046

Natural Language Processing

Lecture 5: Assignment1 and Language Fundamental

Dr. Caren Han

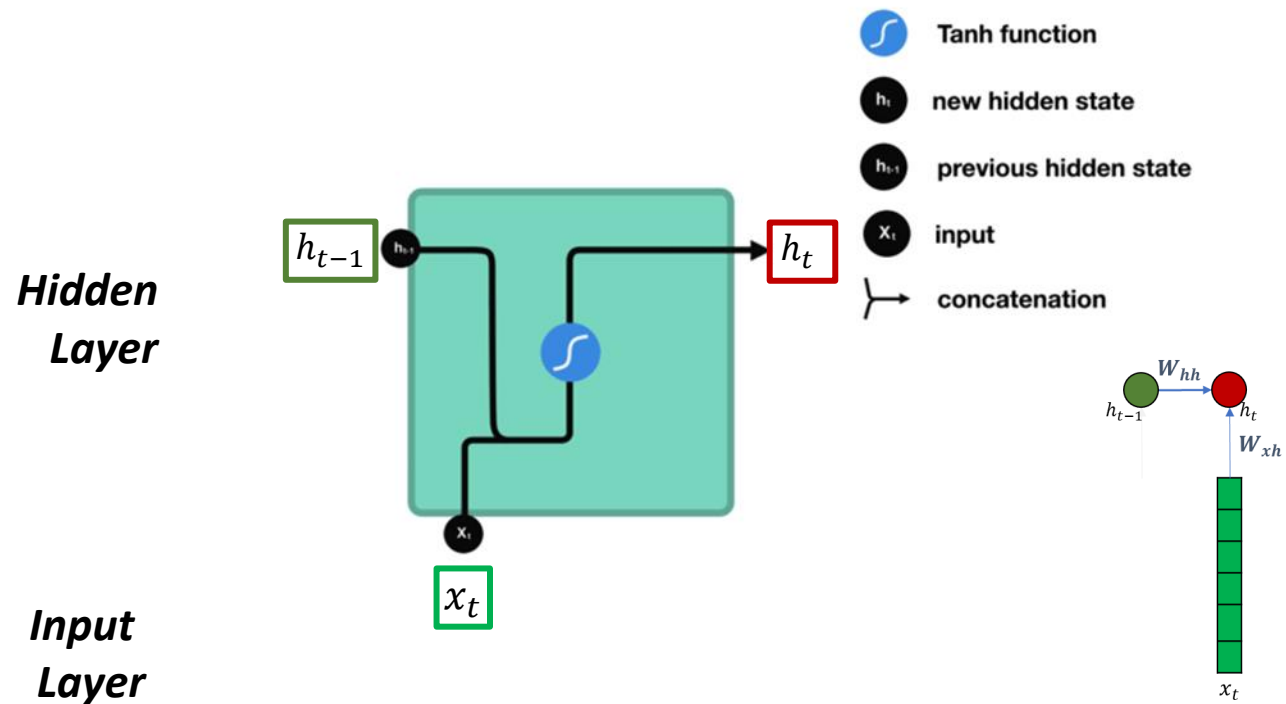
*Semester 1, 2021
School of Computer Science,
University of Sydney*



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

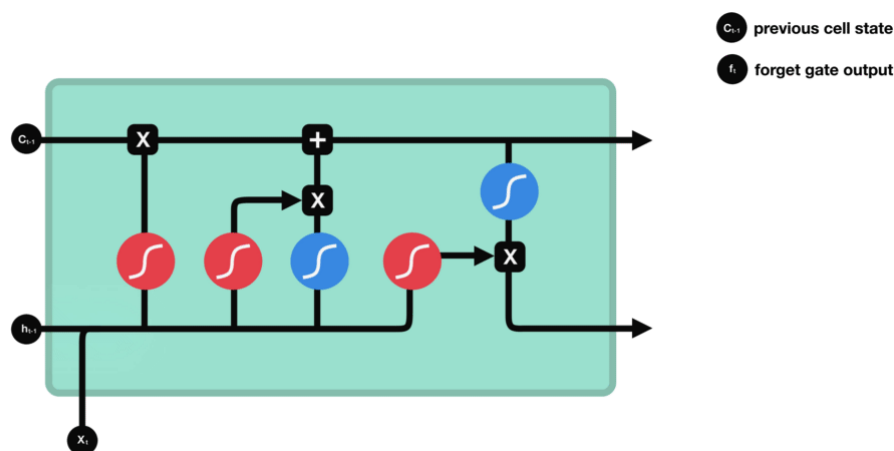
Neural Network + Memory = Recurrent Neural Network



$$\boxed{h_t} = \boxed{\tanh}(W_{hh} \underbrace{h_{t-1}}_{\text{Previous state}} + \underbrace{W_{xh} x_t}_{\text{input}} + \underbrace{b_h}_{\text{bias}})$$

New hidden state *A function with parameters W*

LSTM (Long Short-Term Memory) – Forget Gate

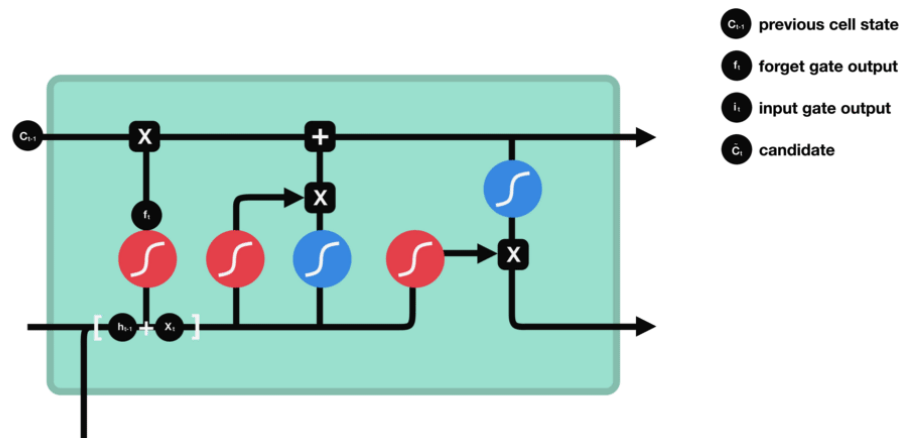


$$f_t = \sigma(W_f[h_{t-1}, x_t] + b_f)$$

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.

LSTM (Long Short-Term Memory) – Input Gate

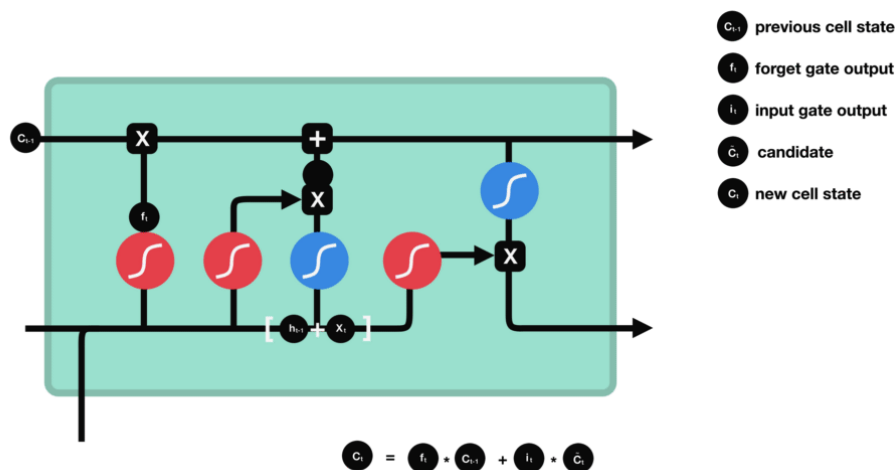


$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i)$$

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

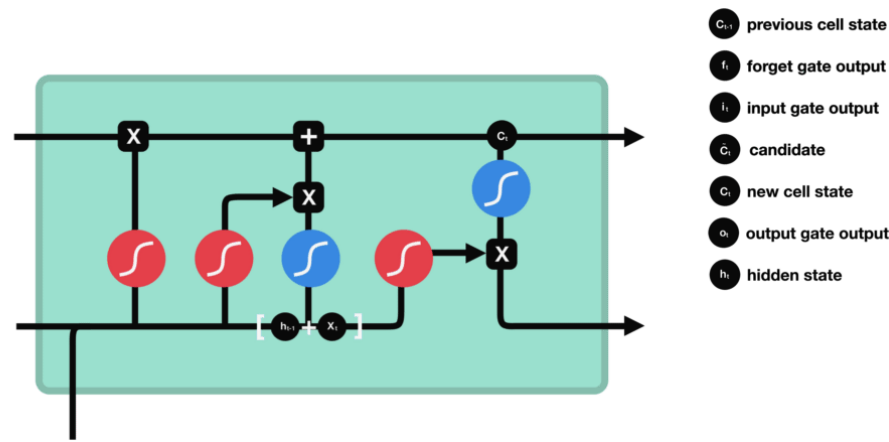
1. Pass the previous hidden state and current input into a sigmoid function
 2. Pass the hidden state and current input into the tanh function to squish values between -1 and 1 to help regulate the network
 3. Multiply the tanh output with the sigmoid output
- *sigmoid output will decide which information is important to keep from the tanh output*

LSTM (Long Short-Term Memory) – Cell States



- 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

LSTM (Long Short-Term Memory) – Output Gate



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

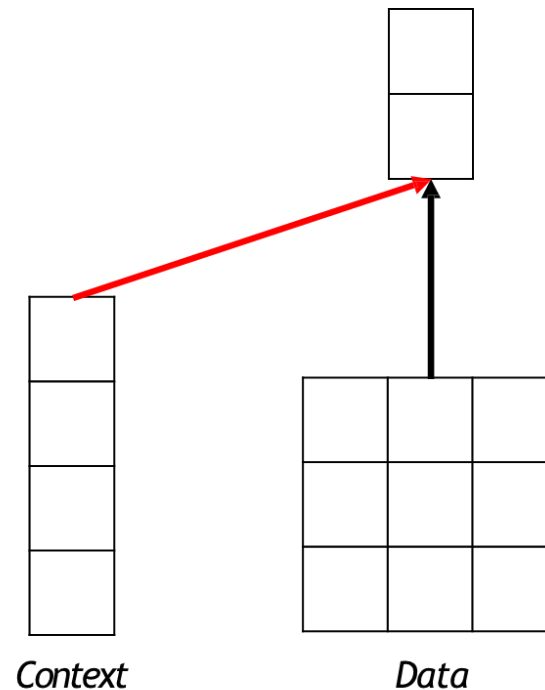
$$h_t = o_t * \tanh(C_t)$$

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

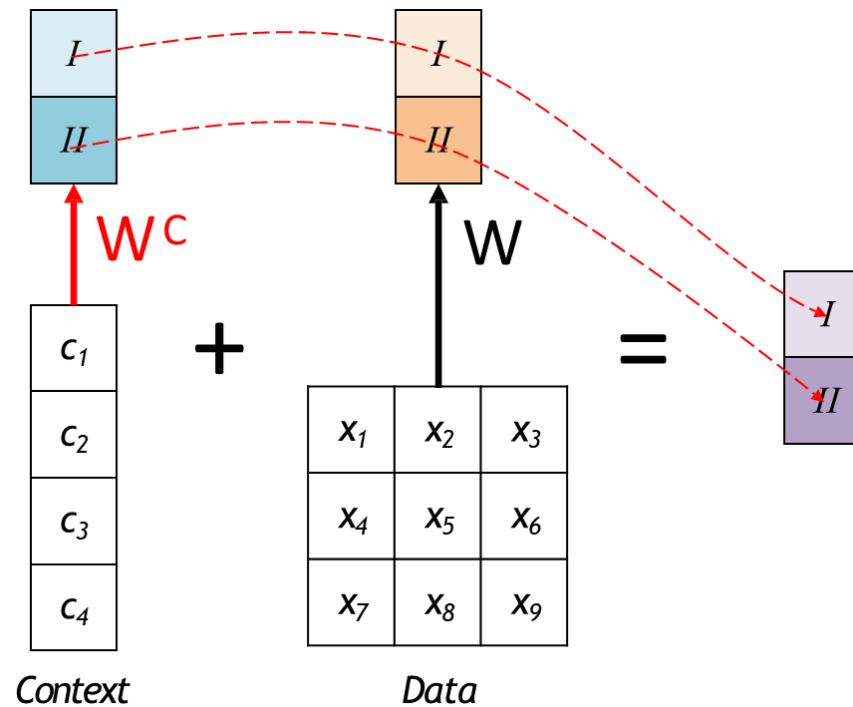
Dealing Context: Review

V to V' – Projection with Context (1)



1 Dealing Context: Review

V to V' – Projection with Context (2)



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
-------	-------	-------	-------	-------	-------	-------	-------	-------

[9x2] matrix

$w_{1,1}$	$w_{2,1}$
$w_{1,2}$	$w_{2,2}$
$w_{1,3}$	$w_{2,3}$
$w_{1,4}$	$w_{2,4}$
$w_{1,5}$	$w_{2,5}$
$w_{1,6}$	$w_{2,6}$
$w_{1,7}$	$w_{2,7}$
$w_{1,8}$	$w_{2,8}$
$w_{1,9}$	$w_{2,9}$

[1x2] matrix

$$= \left[\sum_i^9 x_i * w_{1,i}, \sum_i^9 x_i * w_{2,i} \right]$$

I II

[1 x 4] matrix

c_1	c_2	c_3	c_4
-------	-------	-------	-------

[1x2] matrix

$w_{1,1}^c$	$w_{2,1}^c$
$w_{1,2}^c$	$w_{2,2}^c$
$w_{1,3}^c$	$w_{2,3}^c$
$w_{1,4}^c$	$w_{2,4}^c$

I II

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

Diagram illustrating the transformation of a vector V to V' using a context matrix.

The input vector V is represented as a $[1 \times (9+4)]$ matrix:

$$[x_1 \ x_2 \ x_3 \ x_4 \ x_5 \ x_6 \ x_7 \ x_8 \ x_9 \ c_1 \ c_2 \ c_3 \ c_4]$$

The weight matrix is a $[(9+4) \times 2]$ matrix:

$w_{1,1}$	$w_{2,1}$
$w_{1,2}$	$w_{2,2}$
$w_{1,3}$	$w_{2,3}$
$w_{1,4}$	$w_{2,4}$
$w_{1,5}$	$w_{2,5}$
$w_{1,6}$	$w_{2,6}$
$w_{1,7}$	$w_{2,7}$
$w_{1,8}$	$w_{2,8}$
$w_{1,9}$	$w_{2,9}$
$w_{1,10}^c$	$w_{2,10}^c$
$w_{1,11}^c$	$w_{2,11}^c$
$w_{1,12}^c$	$w_{2,12}^c$
$w_{1,13}^c$	$w_{2,13}^c$

The context vector C is a $[1 \times 4]$ matrix:

$$[c_1 \ c_2 \ c_3 \ c_4]$$

The transformation is shown as:

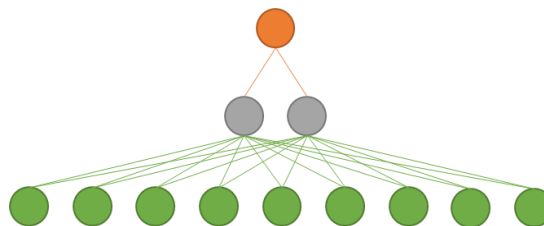
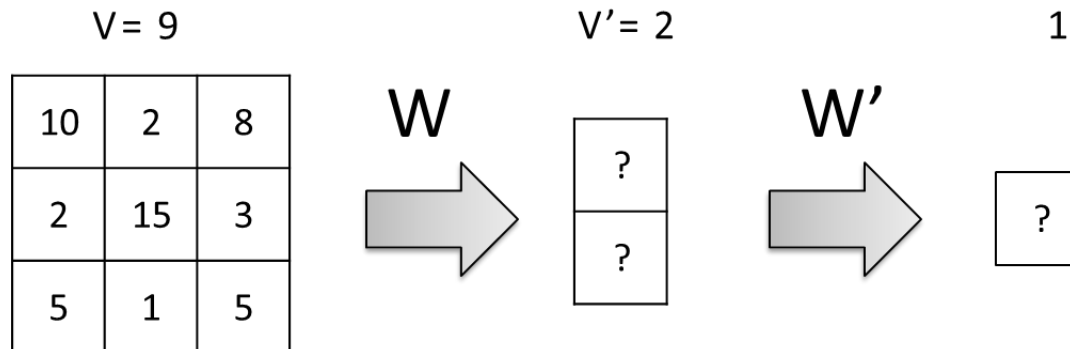
$$[1 \times (9+4)] \text{ matrix} \times \begin{bmatrix} w_{1,1} & w_{2,1} \\ w_{1,2} & w_{2,2} \\ w_{1,3} & w_{2,3} \\ w_{1,4} & w_{2,4} \\ w_{1,5} & w_{2,5} \\ w_{1,6} & w_{2,6} \\ w_{1,7} & w_{2,7} \\ w_{1,8} & w_{2,8} \\ w_{1,9} & w_{2,9} \\ w_{1,10}^c & w_{2,10}^c \\ w_{1,11}^c & w_{2,11}^c \\ w_{1,12}^c & w_{2,12}^c \\ w_{1,13}^c & w_{2,13}^c \end{bmatrix} = \begin{bmatrix} \sum_i^9 x_i * w_{1,i} & \sum_i^9 x_i * w_{2,i} \\ \sum_i^4 c_i * w_{1,i}^c & \sum_i^4 c_i * w_{2,i}^c \end{bmatrix}$$

The result is a $[1 \times 2]$ matrix:

$$[I \quad II]$$

1 Dealing Context: Review

$$V \rightarrow V' \rightarrow 1$$

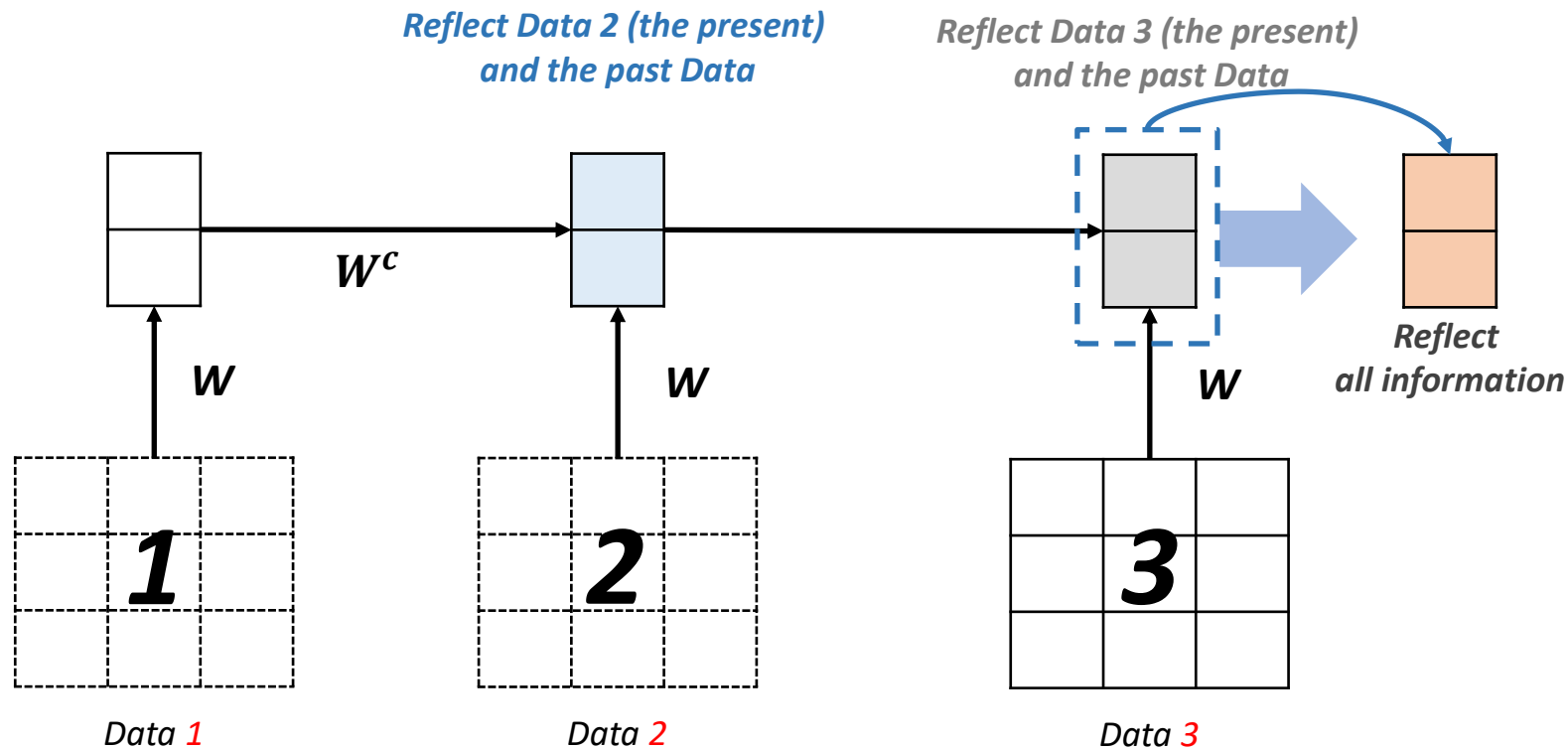


Lecture 5: Assignment1 and Language Fundamental

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

2 Assignment 1 Discussion

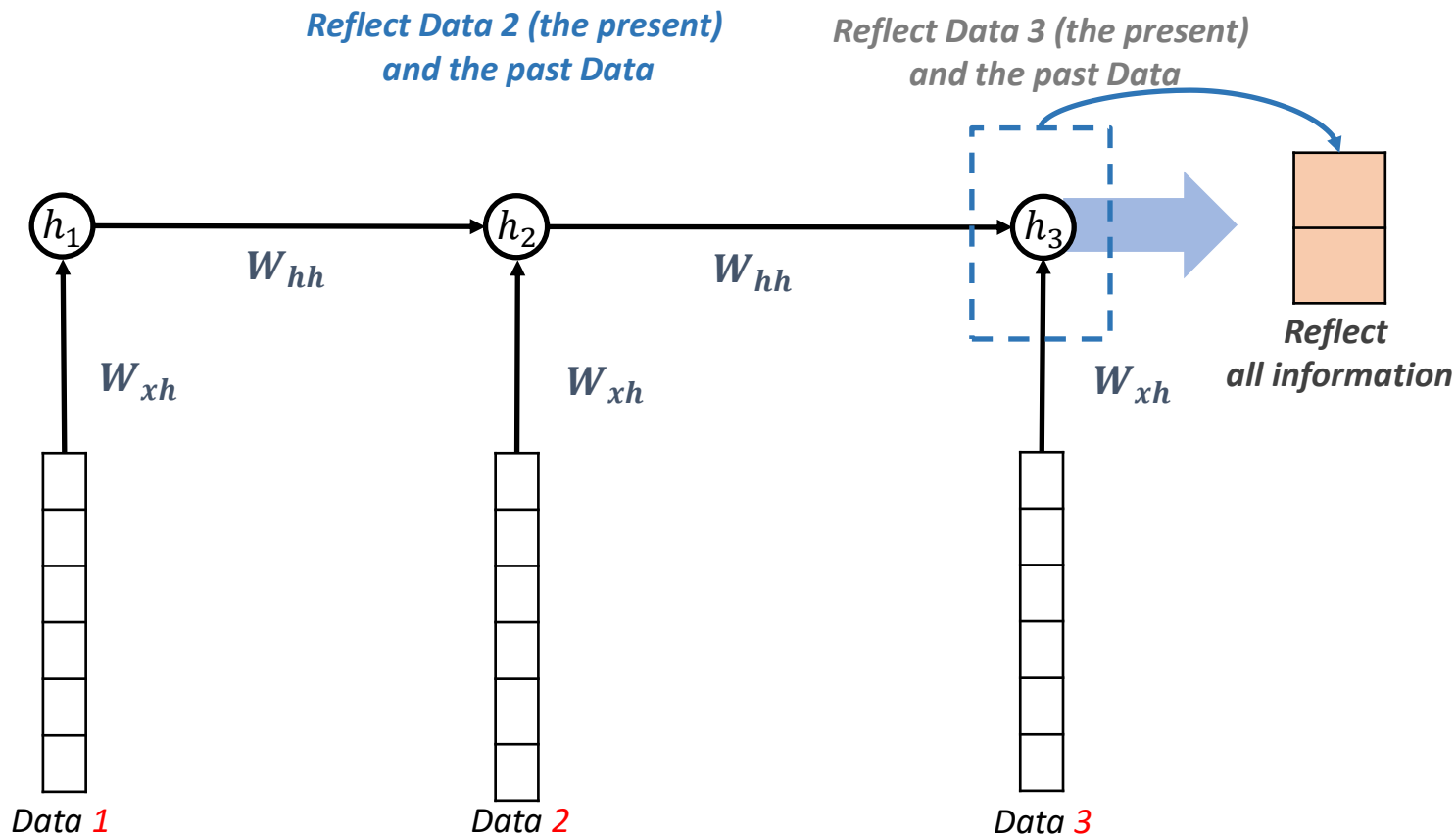
$V_s \rightarrow V's \rightarrow V'$



2

Assignment 1 Discussion

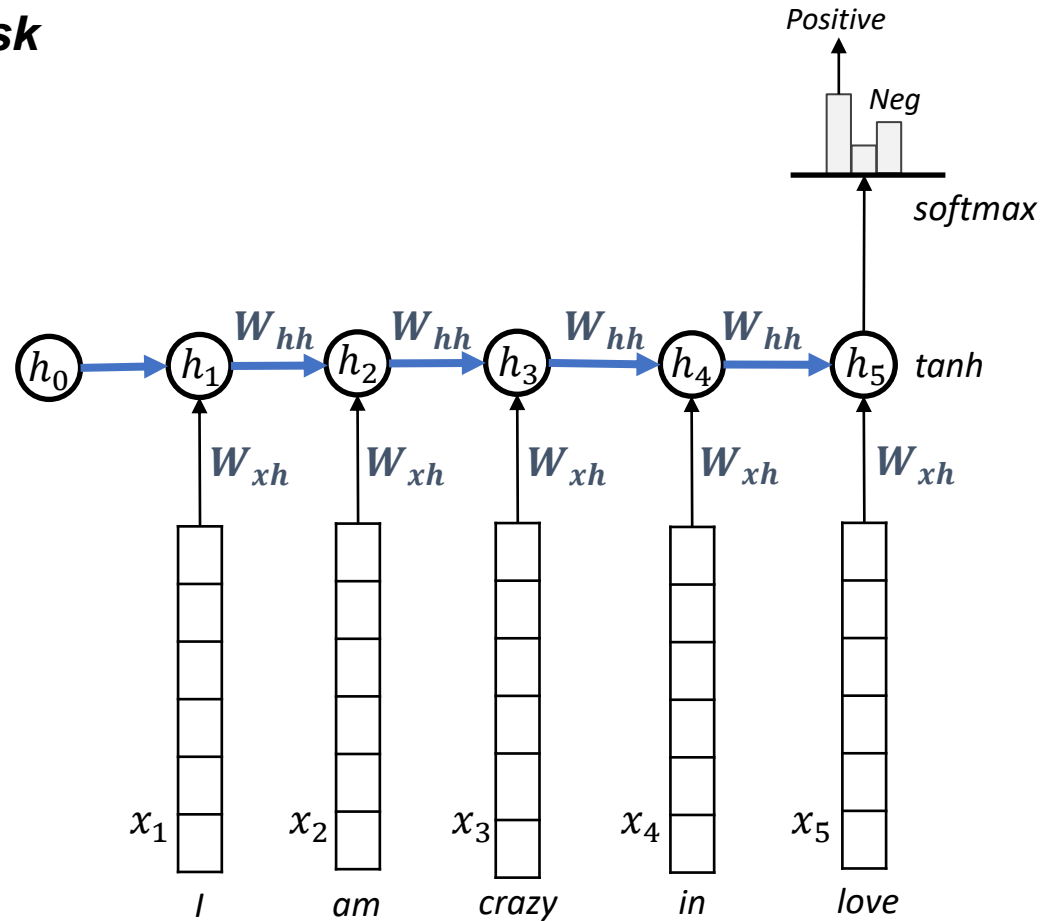
$V_s \rightarrow V's \rightarrow V'$



Assignment 1 Discussion

RNN

N to 1 Task

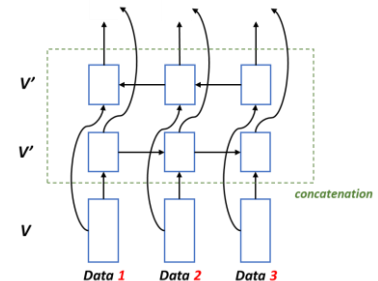
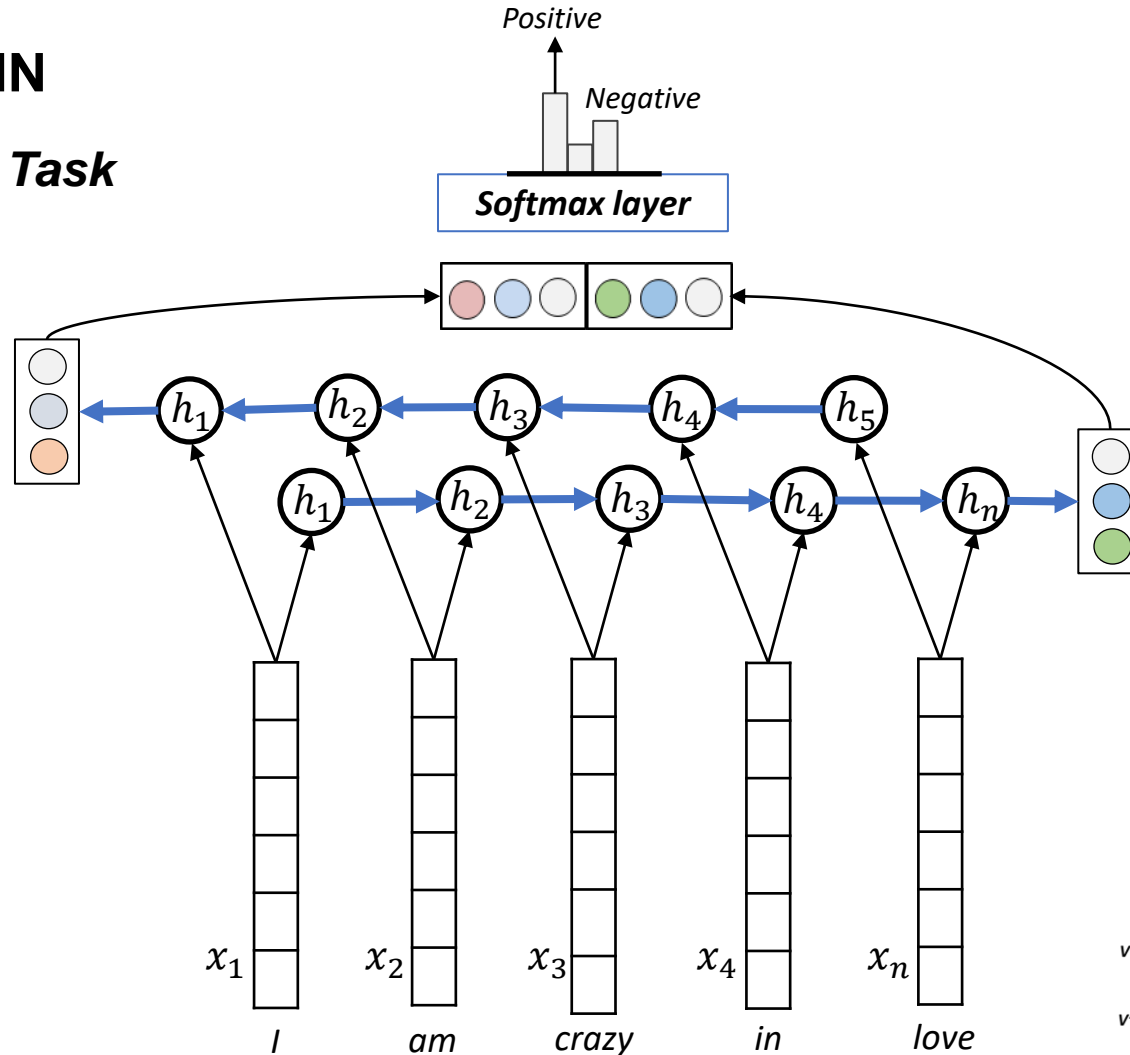


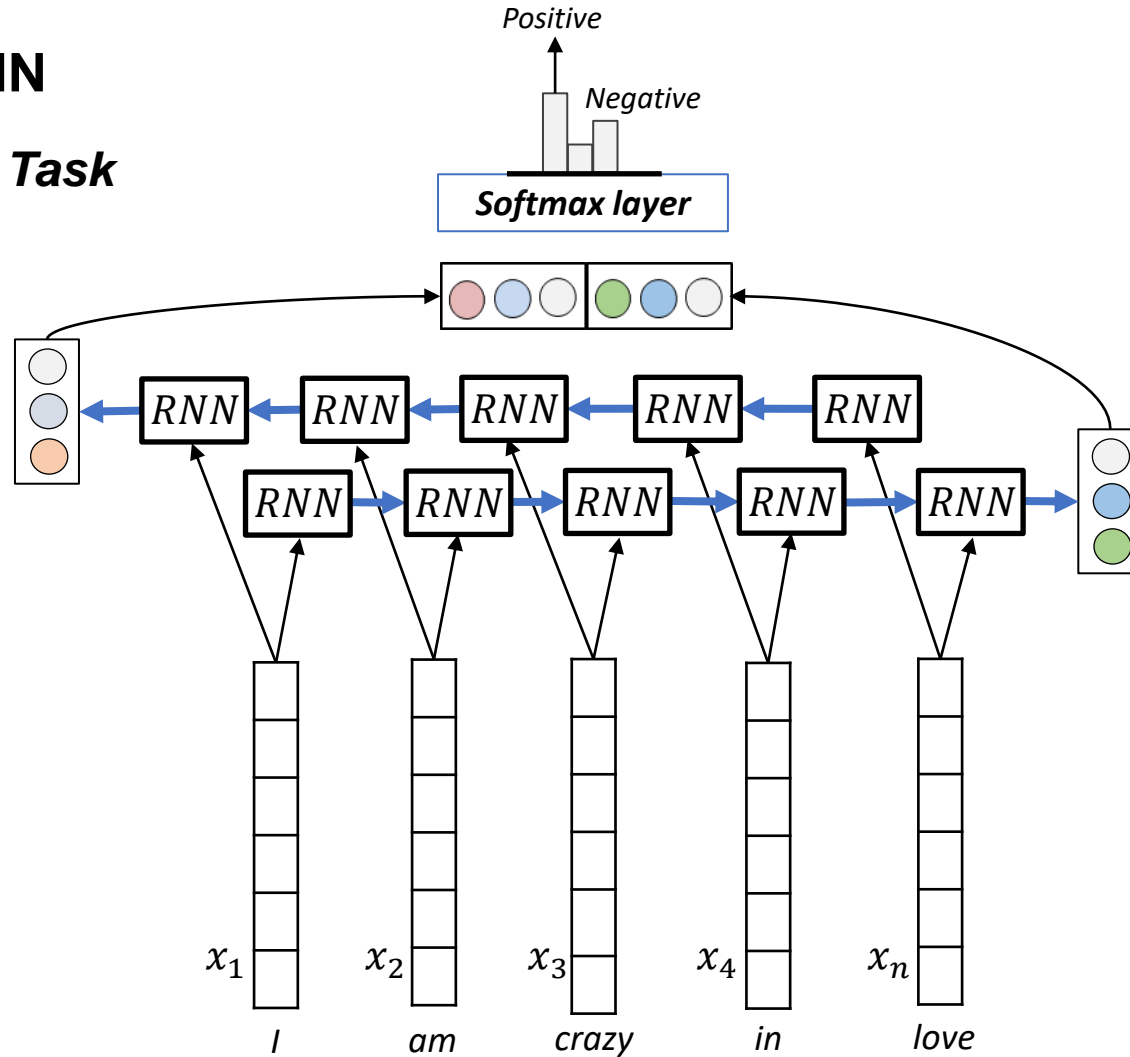
2

Assignment 1 Discussion

Bi-RNN

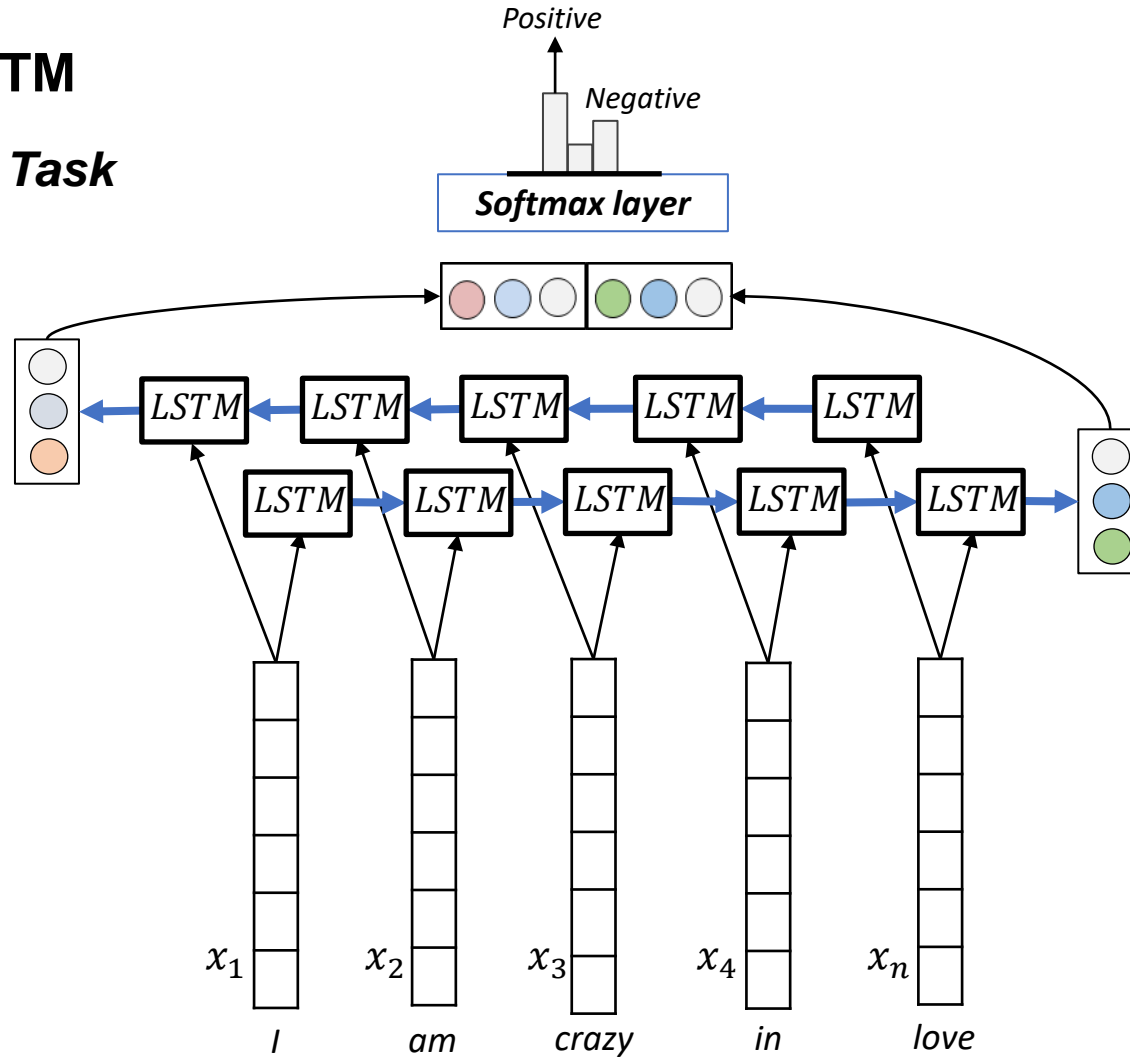
N to 1 Task



Bi-RNN***N to 1 Task***

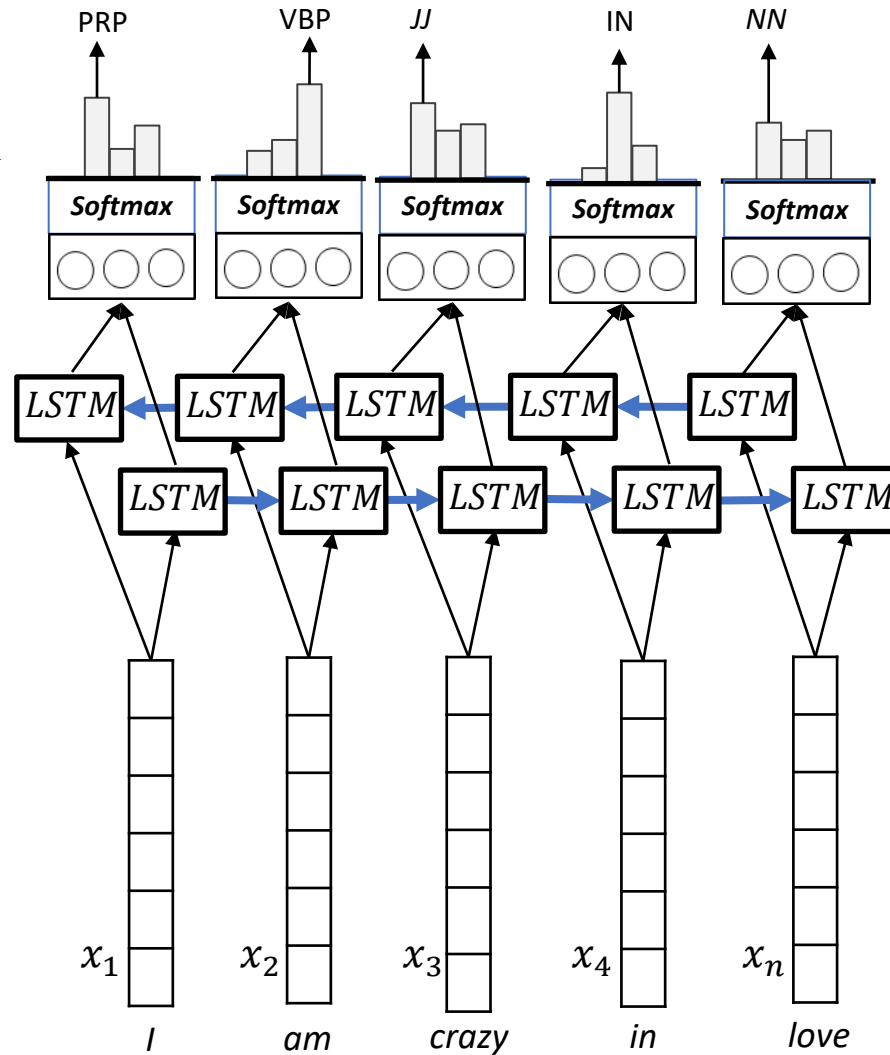
Bi-LSTM

N to 1 Task



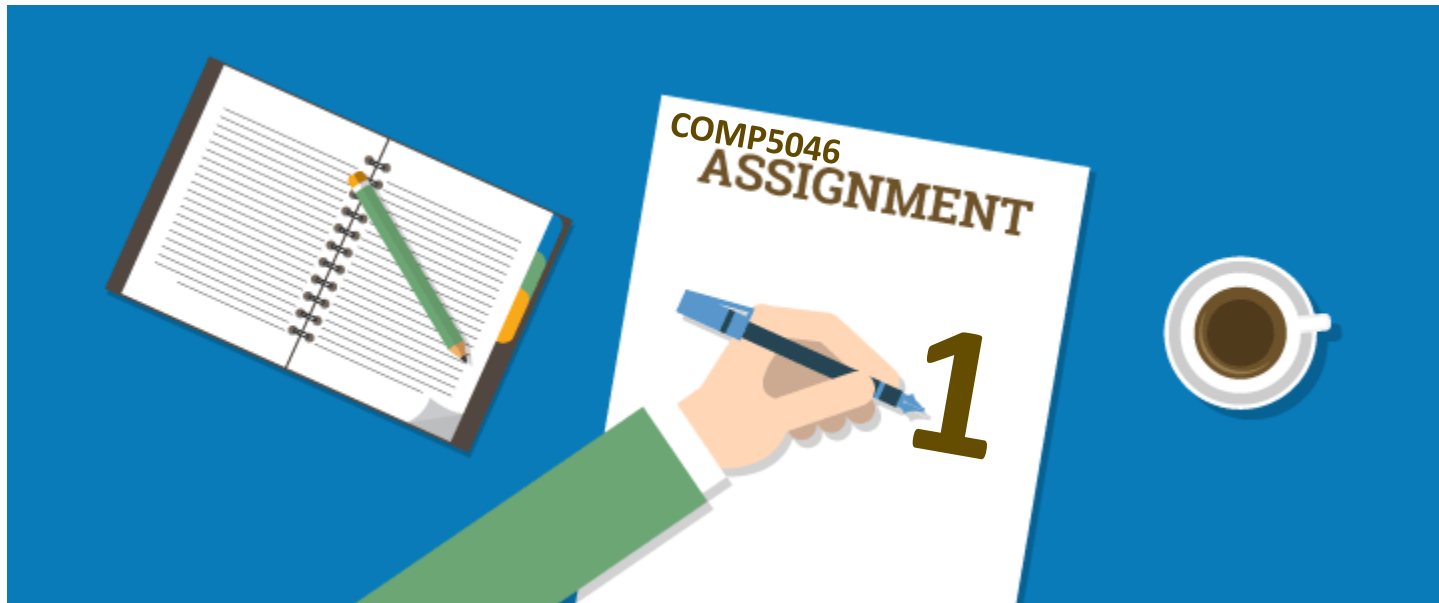
Bi-LSTM

N to N Task



Assignment 1 Discussion

Let's discuss our Assignment 1



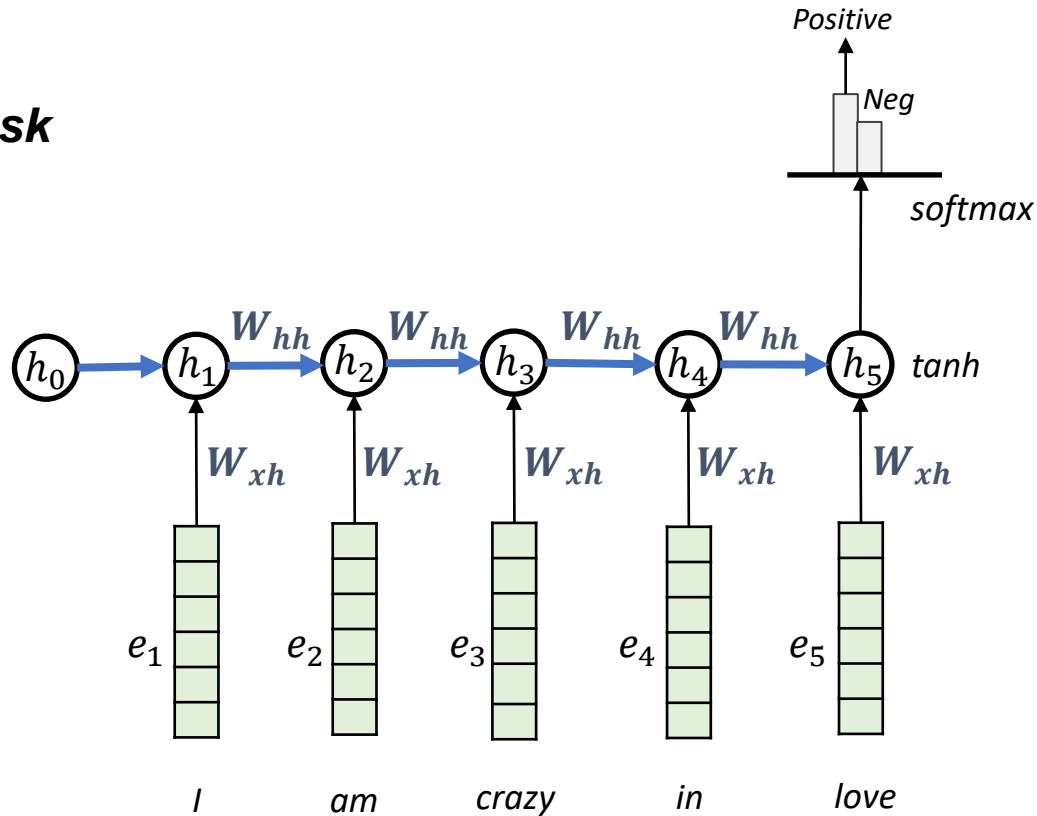
Sentiment Analysis using Recurrent Neural Networks!



Assignment 1 Discussion - Model

RNN

N to 1 Task



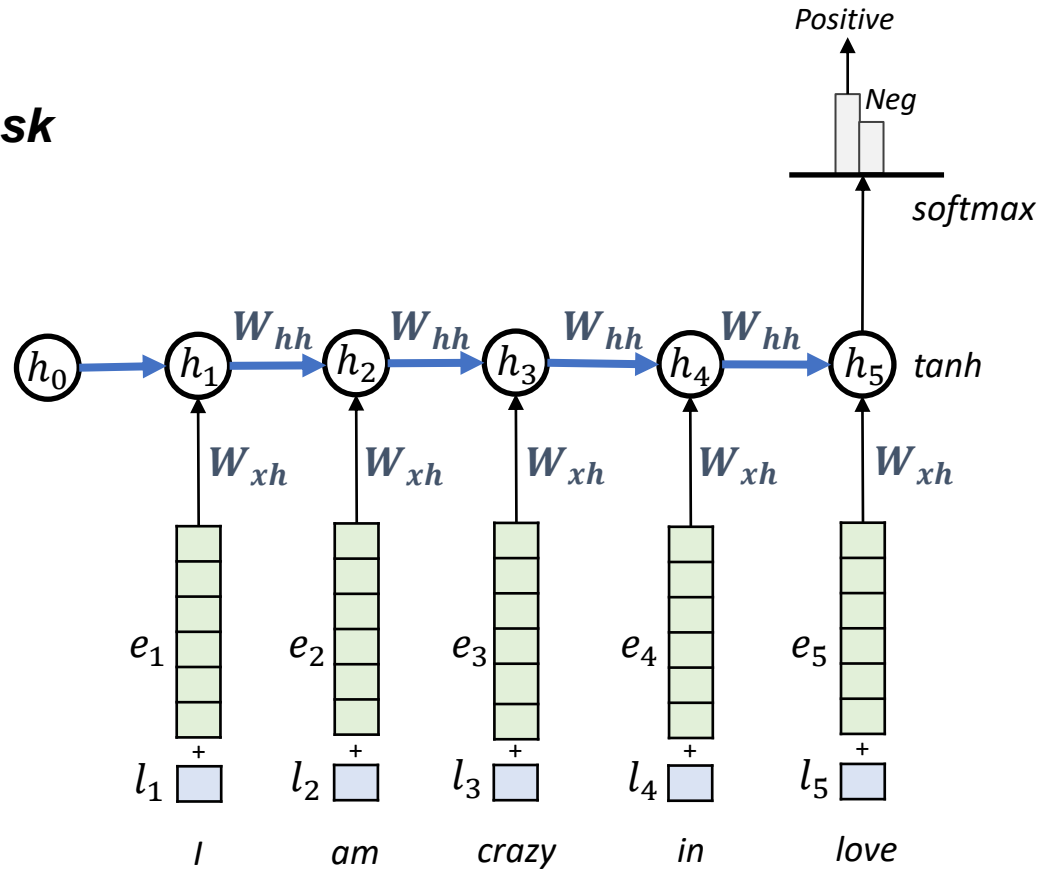
$e_n = \text{word embedding}$



2

Assignment 1 Discussion - Model

RNN

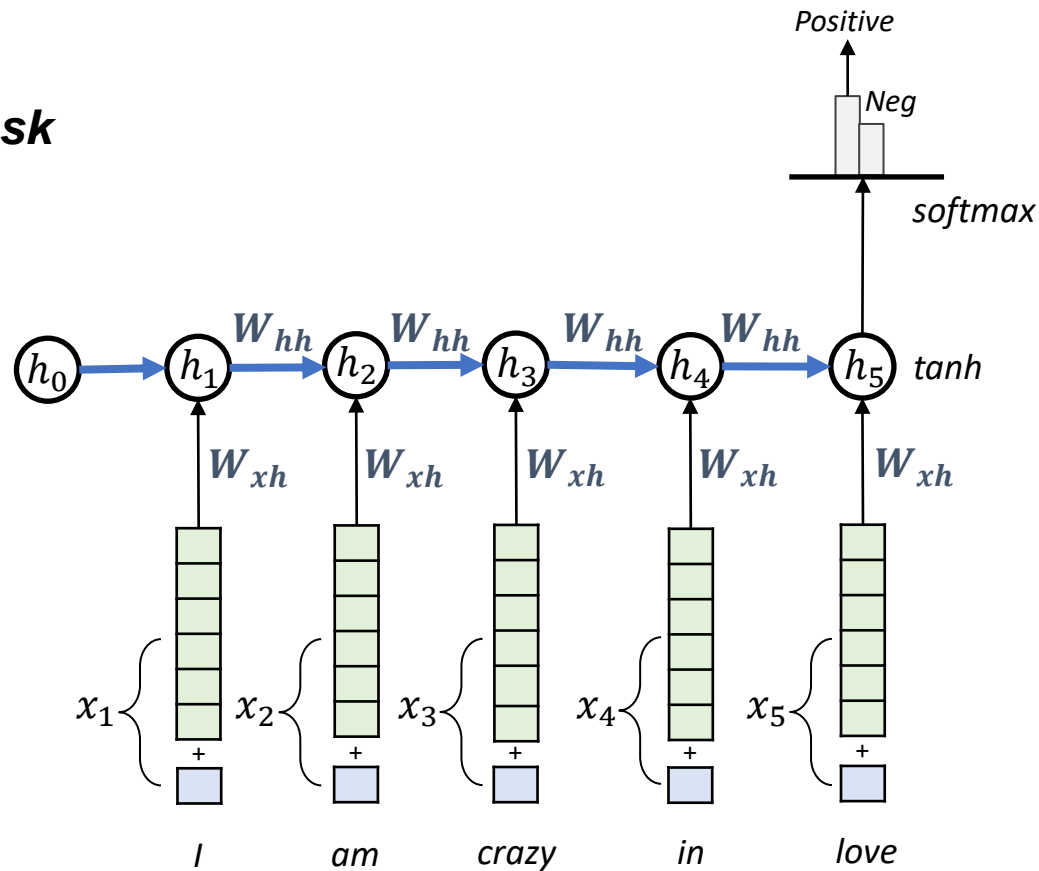
N to 1 Task



 e_n = word embedding
 l_n = lexicon embedding

RNN



N to 1 Task



e_n = word embedding
 l_n = lexicon embedding

RNN

The diagram illustrates a sequence-to-sequence model architecture for sentiment classification. At the bottom, the input sequence is the sentence "I am crazy in love", with words "I", "am", "crazy", "in", and "love" aligned under their respective input vectors x_1 through x_5 . Each input vector is a vertical stack of five green squares, with a small blue square containing a "+" sign below it. Above each input vector is an RNN block. The RNN blocks are arranged in two rows: the first row contains RNN blocks for "I", "am", and "crazy"; the second row contains RNN blocks for "in" and "love". Blue arrows indicate the sequence flow from left to right between RNN blocks. Arrows also point from each input vector to its corresponding RNN block. The output of the final RNN block (for "love") is a vector of three colored circles (white, blue, green). This vector is fed into a "Softmax layer" at the top, which produces a probability distribution over "Positive" and "Neg" classes, represented by two bars of different heights. A "Task" label with a box of three colored circles (white, blue, orange) is on the left, with an arrow pointing to the Softmax layer output.

-  $e_n = \text{word embedding}$
-  $l_n = \text{lexicon embedding}$

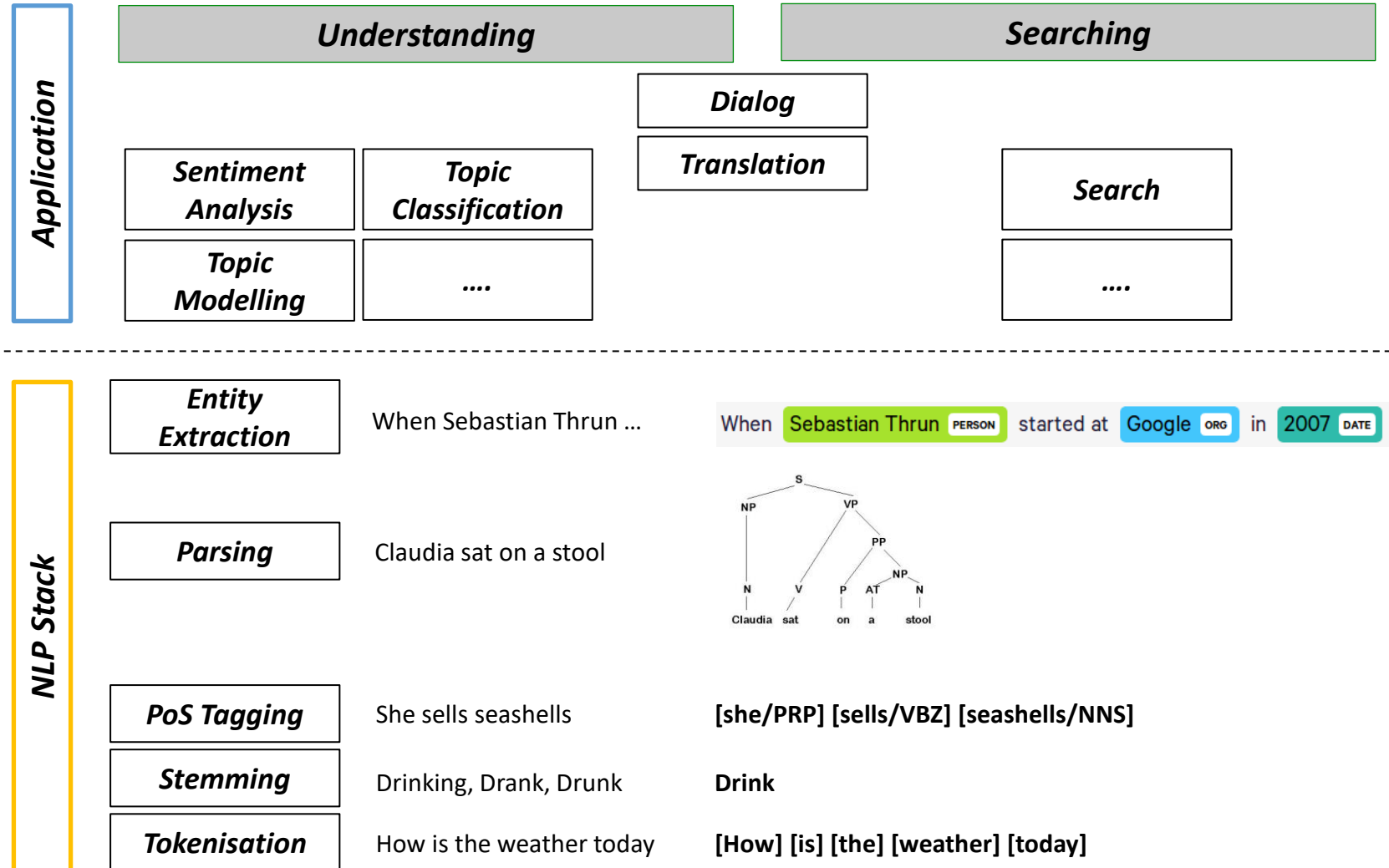
Assignment 1 Discussion - Model

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

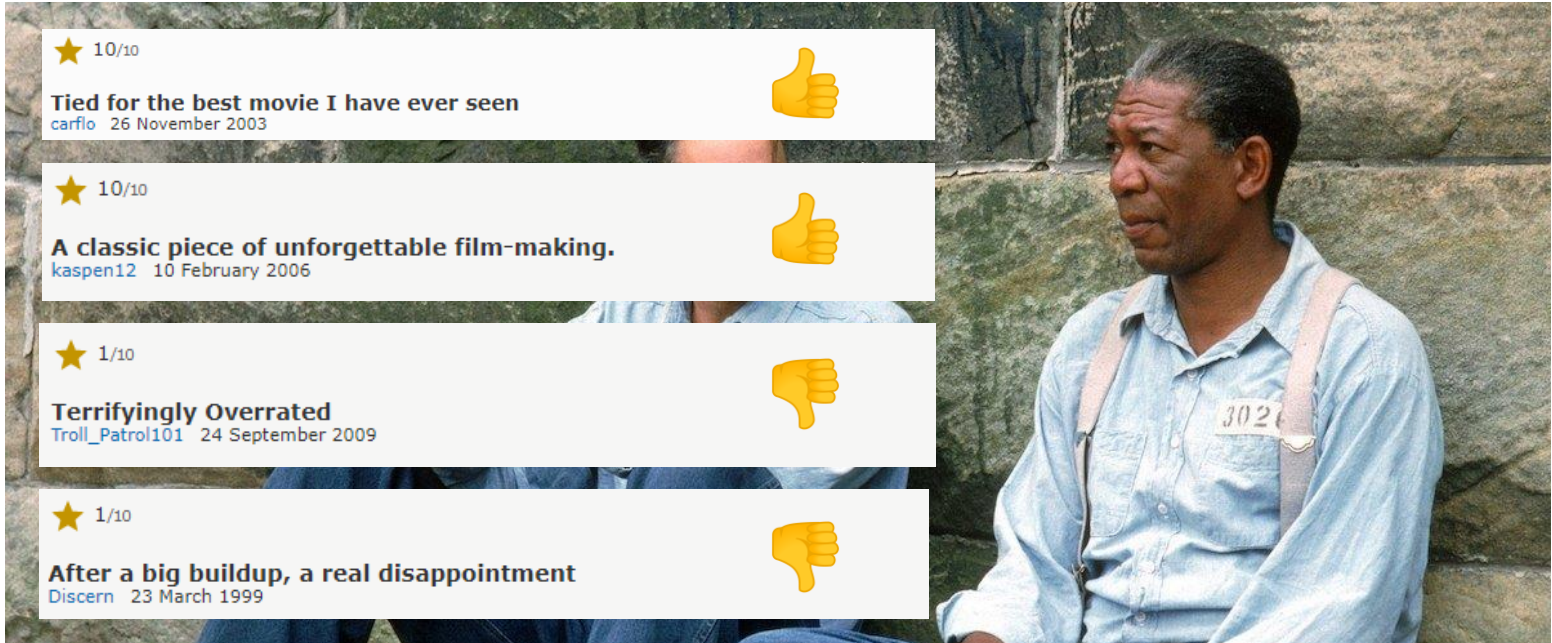
Lecture 5: Assignment1 and Language Fundamental

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

The purpose of Natural Language Processing: Overview



Movie Review – Positive or Negative



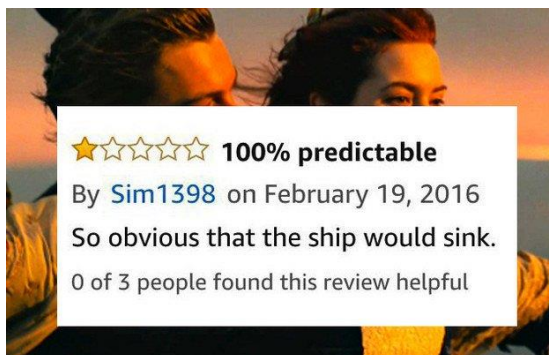
★ 10/10	👍
Tied for the best movie I have ever seen carflo 26 November 2003	
★ 10/10	👍
A classic piece of unforgettable film-making. kaspen12 10 February 2006	
★ 1/10	👎
Terrifyingly Overrated Troll_Patrol101 24 September 2009	
★ 1/10	👎
After a big buildup, a real disappointment Discern 23 March 1999	

Too easy? 😏

What is Sentiment Analysis?



★ 10/10	Tied for the best movie I have ever seen carlio 26 November 2003	👍
★ 10/10	A classic piece of unforgettable film-making. kaspen12 10 February 2006	👍
★ 1/10	Terrifyingly Overrated Troll_Patrol101 24 September 2009	👎
★ 1/10	After a big buildup, a real disappointment Discern 23 March 1999	👎



★☆☆☆☆ 100% predictable
By Sim1398 on February 19, 2016
So obvious that the ship would sink.
0 of 3 people found this review helpful



★☆☆☆☆ One Star
By Joe Watson - December 14, 2014
There were no wolves in the movie.
0 of 3 people found this review helpful



★★★★★ The snowman keeps falling apart
By Kelsey - December 1, 2014
The snowman keeps falling apart
5 of 12 people found this review helpful

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*

Sentiment Analysis



Cottonelle FreshCare Flushable Wipes for Adults, Wet Wipes, Alcohol Free, 336 Wet Wipes per Pack (Eight 42-Count Resealable Soft Packs)

by Cottonelle

★★★★☆ 11,351 ratings

Available from these sellers.

Style Name: **8 Packs of Flushable Wipes**

- Superior Clean CleaningRipples texture provides softness while removing more cleans better versus using dry bath tissue alone
- 100 percent flushable & the No. 1 Flushable Wipe Brand among national flushable wipes brands
- Immediately Starts to Break Down After Flushing – Cottonelle bathroom wipes break down 6X's faster than Dude Wipes (based on strength loss testing) and are sewer safe & septic safe with SafeFlush Technology
- Moist wipes made from fibers that are 100 percent biodegradable
- Adult wipes that are infused with the gentle cleansing power of water and are perfect for man wipes, feminine wipes and more

Customer reviews

★★★★☆ 4.6 out of 5

11,351 customer ratings



▼ How does Amazon calculate star ratings?

Review this product

Share your thoughts with other customers

Write a customer review

Top international reviews



MustLoveDogs

★★★★☆ **Just because you CAN flush it, doesn't mean you should!**

Reviewed in the United States on 14 July 2018

Style: 8 Packs of Flushable Wipes | **Verified Purchase**

Flushable? Not according to the plumber I just paid \$200 to. Be careful folks. Other than the misleading "flushable" advertising, I liked product, but can't afford plumbing bills.

354 people found this helpful

Helpful

Report abuse



Zack Fischmann

★★★★☆ **These are NOT unscented -- one of the ingredients is "fragrance/parfum"**

Reviewed in the United States on 15 January 2019

Style: 8 Packs of Flushable Wipes | **Verified Purchase**

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*

Sentiment Analysis: Examples

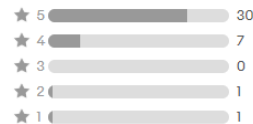
Apple iPhone 7 - 128GB - Rose Gold (Unlocked)

★★★★★ 39 product ratings | [About this product](#)

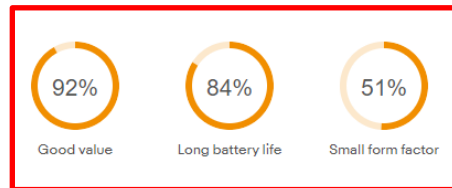


Ratings and reviews

4.6
★★★★★
39 product ratings



Aspects



[Write a review](#)

Most relevant reviews

[See all 24 reviews](#)

★★★★★

by [judeel2](#)

18 Jul, 2019

Excellent phone

Works excellently well, the screen is very very clear. Photos are better than my iPhone 5se, even though they are both 12mp. Front facing camera is 7mp, 5se is less. The only downside is the battery life. It doesn't last all day for me. I have small hands but the larger size isn't too big. Can highly recommend, good value.

Verified purchase: Yes | Condition: Pre-Owned

★★★★★

by [noadaughert_31](#)

26 Apr, 2018

Really good for price

Had virtually no scratches and battery life is optimal despite being refurbished. Good value for your money. Only complaint was that there wasnt any accessories such as the bluetooth ear buds required for listening to music or the lightning to AUX adapter. But no accessories were listed in the description.

Verified purchase: Yes | Condition: Pre-Owned

★★★★★

by [diannpedlo_0](#)

03 Jan, 2019

Good practical iPhone.

It's just so much better than my previous iPhone 6 as it was damaged & difficult to use. The iPhone 7 feels good to use. I'm not really sure it was the best price as I didn't shop around but am happy regardless,

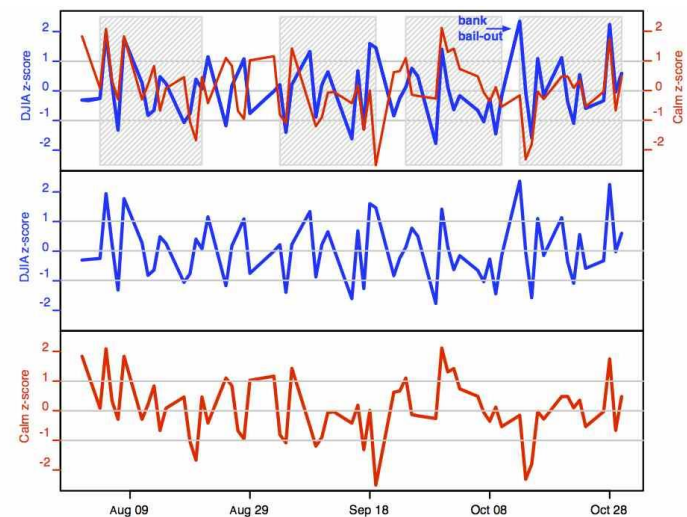
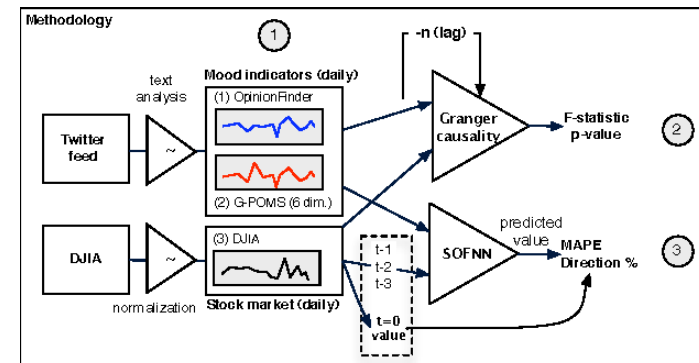
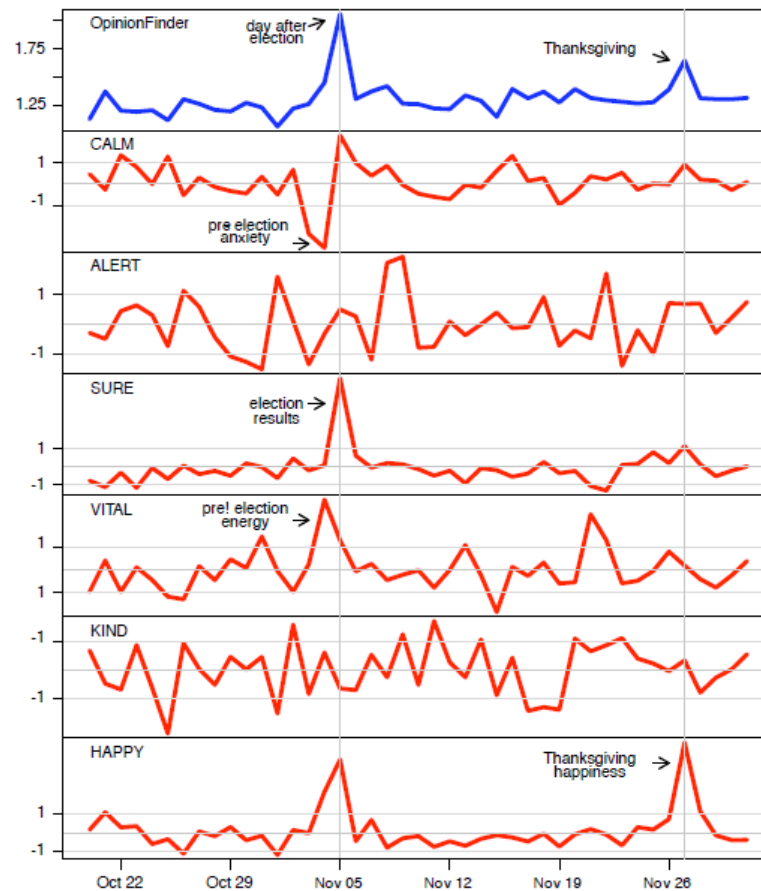
Verified purchase: Yes | Condition: New

Sentiment Analysis: Sentiment viz



Sentiment Analysis: Examples

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



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*

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*

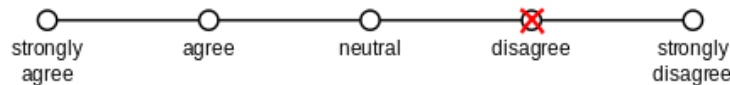
What is Sentiment Analysis?

- *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***

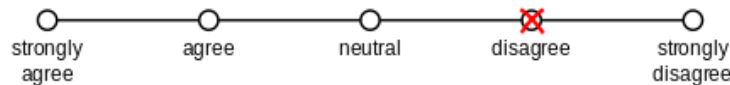
What is Sentiment Analysis?

- *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*

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:

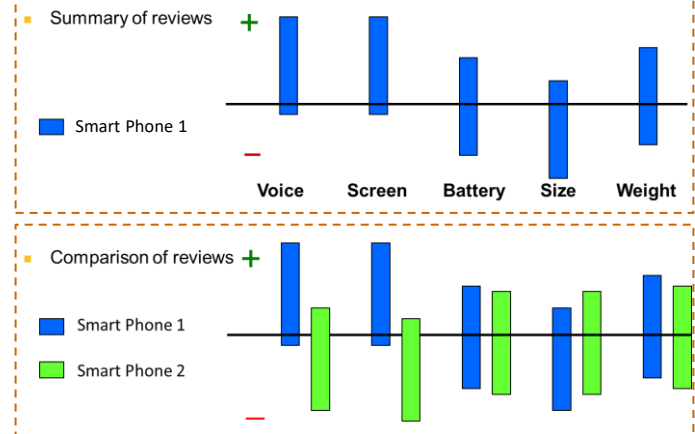
1) Retina display is awesome. Everything looks more defined and sharper. There is much color and clarity out there... or should I say, in those digital images and videos... needless to say, the camera as well captures great images.

....."

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



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]*

Useful Sentiment Lexicons

Name	Details
The General Inquirer http://www.wjh.harvard.edu/~inquirer http://www.wjh.harvard.edu/~inquirer/homecat.htm http://www.wjh.harvard.edu/~inquirer/inquirerbasic.xls	Categories <ul style="list-style-type: none">• Positiv (1915 words) and Negativ (2291 words)• Strong vs Weak, Active vs Passive, Overstated versus Understated• Pleasure, Pain, Virtue, Vice, Motivation, Cognitive Orientation, etc Free to use
LIWC Linguistic Inquiry and Word Count http://www.liwc.net/	2300 words and less than 70 classes Affective Processes <ul style="list-style-type: none">• negative emotion (bad, weird, hate, problem, tough)• positive emotion (love, nice, sweet) Cognitive Processes <ul style="list-style-type: none">• Tentative (maybe, perhaps, guess), Inhibition (block, constraint)• Pronouns, Negation (no, never), Quantifiers (few, many) \$30 or \$90 fee
MPQA Subjectivity Cues Lexicon http://www.cs.pitt.edu/mpqa/subj_lexicon.html	Each word annotated for intensity (strong, weak) 6885 words from 8221 lemmas <ul style="list-style-type: none">• 2718 positive• 4912 negative GNU GPL (widely-used free software license)
Opinion Lexicon http://www.cs.uic.edu/~liub/FBS/opinion--lexicon--English.rar	6786 words <ul style="list-style-type: none">• 2006 positive/ 4783 negative Free to use
SentiWordNet http://swn.isti.cnr.it/	All WordNet synsets automatically annotated for degrees of positivity, negativity, and neutrality/objectiveness <ul style="list-style-type: none">• [estimable(J,3)] “may be computed or estimated” Pos 0 Neg 0 Obj 1• [estimable(J,1)] “deserving of respect or high regard” Pos .75 Neg 0 Obj .25 Free to use

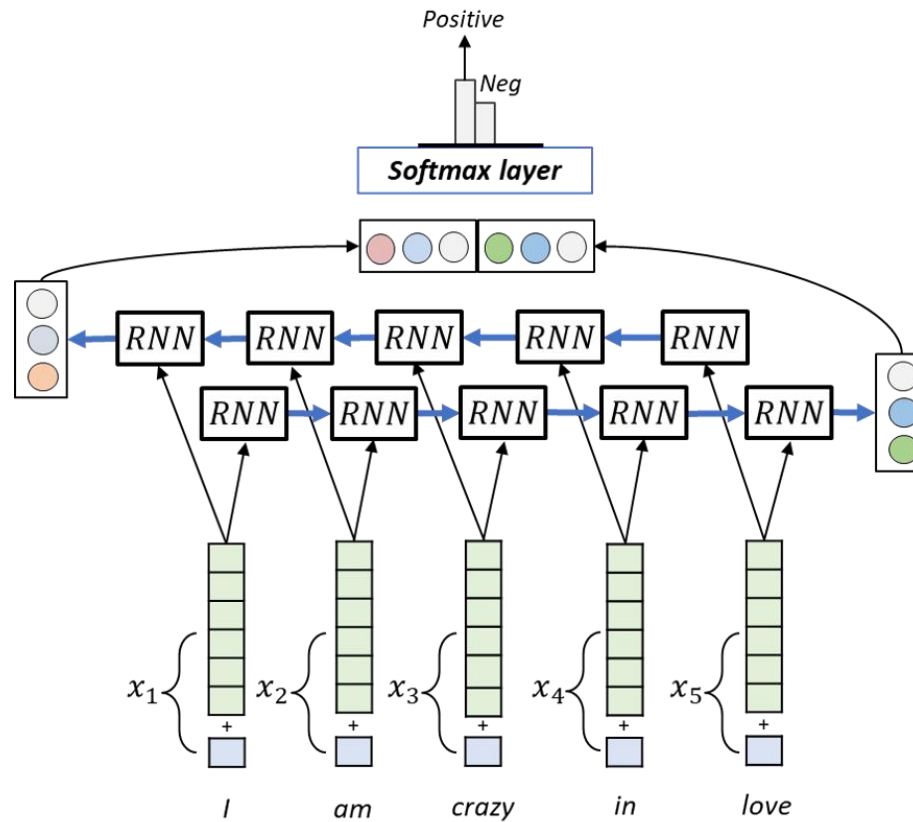
Can you build the sentiment lexicon by yourself?

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*



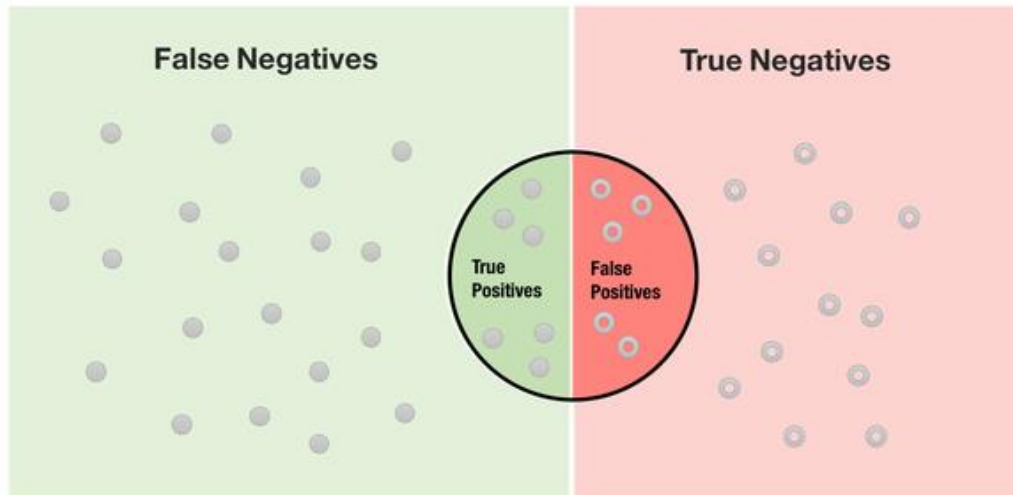
Assignment 1: Sentiment Analysis



e_n = word embedding

l_n = lexicon embedding

Assignment 1: Sentiment Analysis



$$precision = \frac{TP}{TP + FP}$$

$$recall = \frac{TP}{TP + FN}$$

$$F1 = \frac{2 \times precision \times recall}{precision + recall}$$

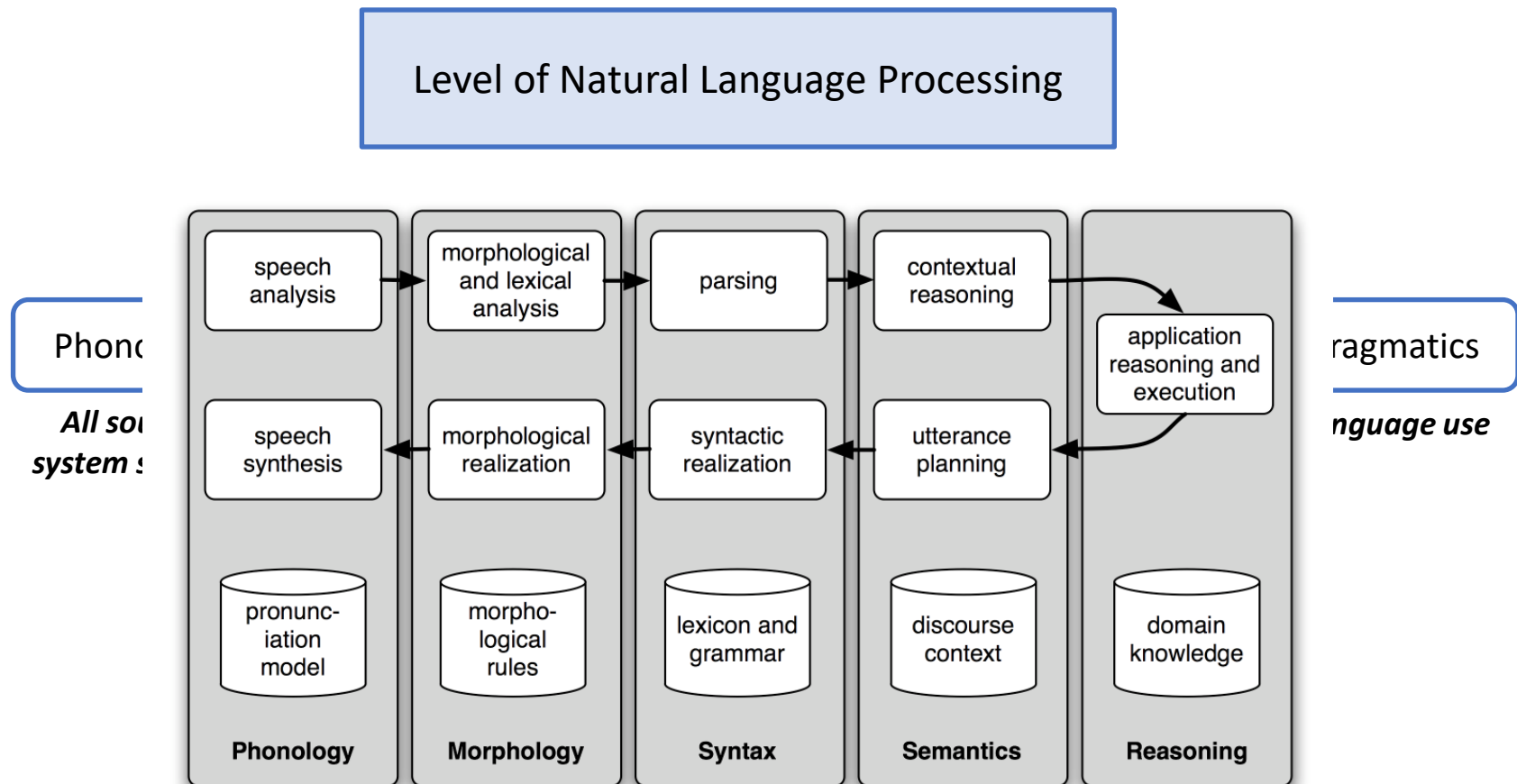
$$accuracy = \frac{TP + TN}{TP + FN + TN + FP}$$

$$specificity = \frac{TN}{TN + FP}$$

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

Level of Natural Language Processing



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**

We know how sounds can combine

Often shown when a word from one language is borrowed into another:



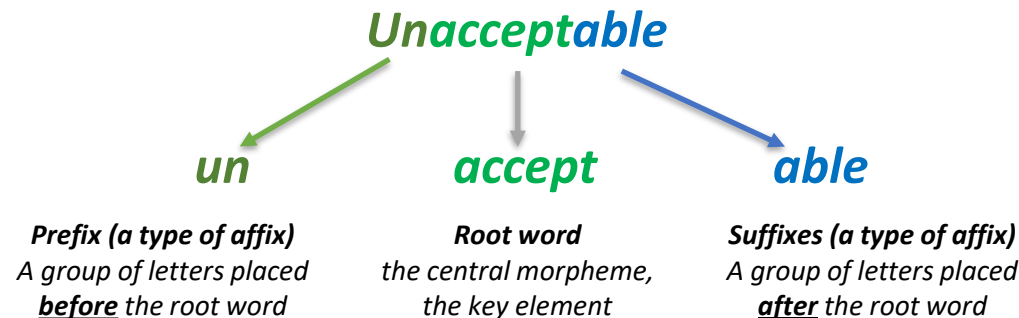
- McDonalds — in English consonant clusters allowed ([mk] and [ldz]) becomes...

マクドナルド	麦当劳	맥도날드
Makudonarudo	Màidāngláo	Maegdonaldeu

in other language — consonant clusters are not allowed

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).



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

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 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

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

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

Text Preprocessing

How many words?

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

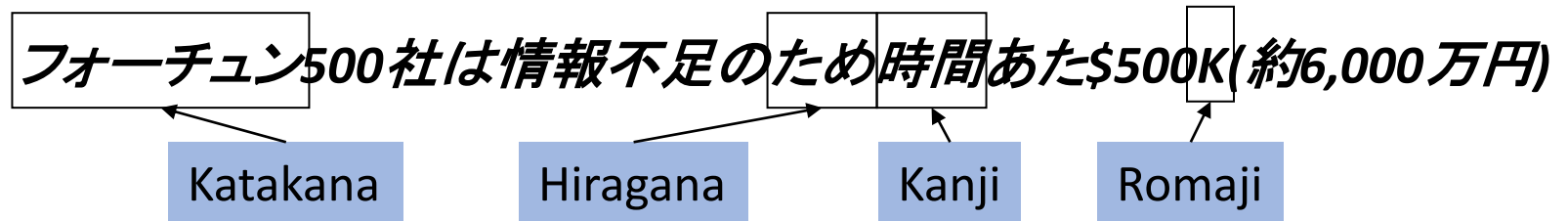
	Tokens = N	Types = $ V $
Switchboard phone conversations	2.4 million	20 thousand
Shakespeare	884,000	31 thousand
Google N-grams	1 trillion	13 million

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*

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



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.

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

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)

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'

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

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*.

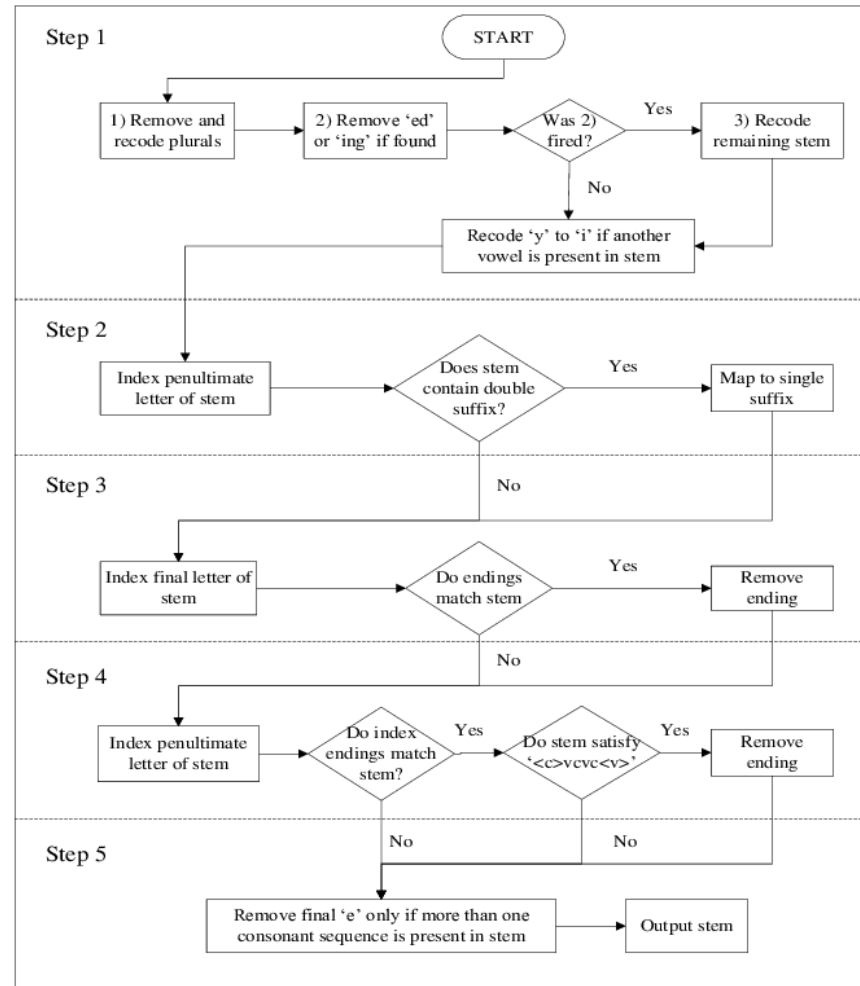
*for example compressed
and compression are both
accepted as equivalent to
compress.*



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

Porter's algorithm: The most common English stemmer

Porter Stemming Algorithm



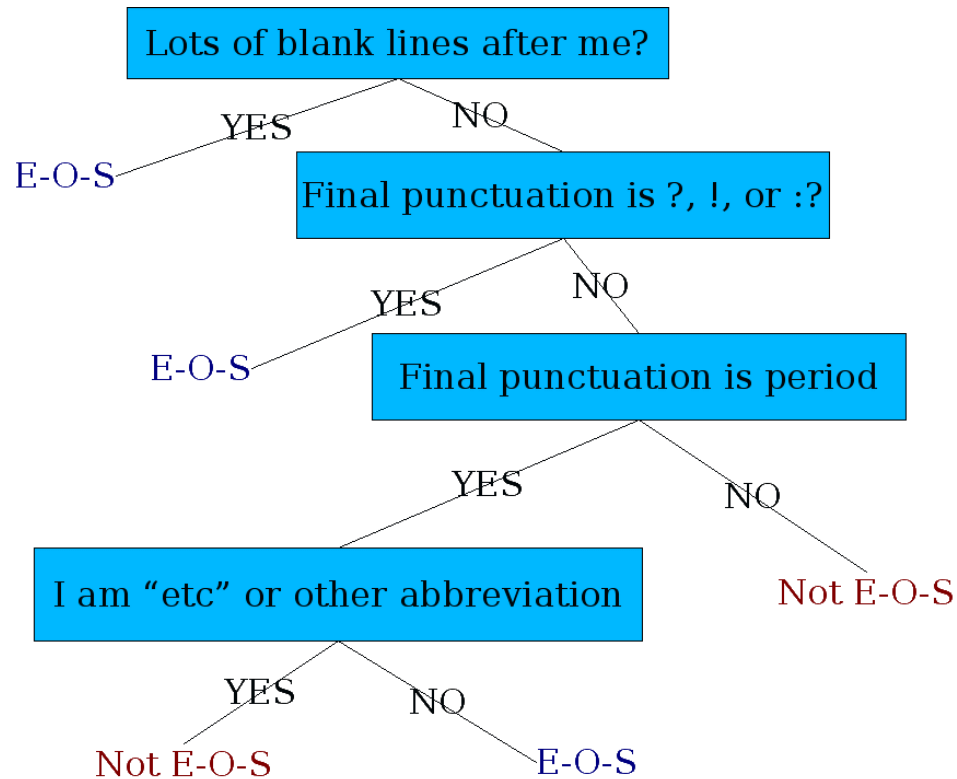
Dealing with complex morphology is sometimes necessary

- Some languages requires complex morpheme segmentation
 - Turkish
 - Uygarlastiramadiklarimizdanmissinizcasina
 - `(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' + casina `as if'

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

Sentence Segmentation using a Decision Tree



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.

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



Regular Expressions: Disjunctions

- Letters inside square brackets []

Pattern	Matches
<code>[wW]oodchuck</code>	Woodchuck, woodchuck
<code>[1234567890]</code>	Any digit

- Ranges [A-Z]

Pattern	Matches	
<code>[A-Z]</code>	An upper case letter	<u>D</u> renched Blossoms
<code>[a-z]</code>	A lower case letter	<u>m</u> y beans were impatient
<code>[0-9]</code>	A single digit	Chapter <u>1</u> : Down the Rabbit Hole

Regular Expressions: Negation in Disjunction

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

Pattern	Matches	
[^A-Z]	Not an upper case letter	O <u>y</u> fn pripetchik
[^Ss]	Neither 'S' nor 's'	<u>I</u> have no exquisite reason"
[^e^]	Neither e nor ^	Look h <u>e</u> re
a^b	The pattern 'a carat b'	Look up <u>a^b</u> now

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

Regular Expressions: More Disjunction

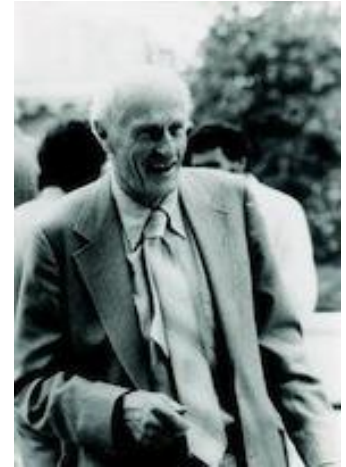
- Woodchucks is another name for groundhog!
- The pipe | for disjunction

Pattern	Matches
<code>groundhog woodchuck</code>	
<code>yours mine</code>	yours mine
<code>a b c</code>	= <code>[abc]</code>
<code>[gG]roundhog [Ww]oodchuck</code>	



Regular Expressions: ? * + .

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



Stephen C Kleene

Kleene *, Kleene +

Text Preprocessing

Regular Expressions: Anchors [^] ^{\$}

Pattern	Matches
[^] [A-Z]	<u>P</u> alo Alto
[^] [^A-Za-z]	<u>1</u> <u>"Hello"</u>
\. ^{\$}	The end <u>.</u>
. ^{\$}	The end <u>?</u> The end <u>!</u>

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

- 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."