CSCE 590-1: From Data to Decisions with Open Data: A Practical Introduction to AI

Lecture 23: Text - Word Representation, Sentiments

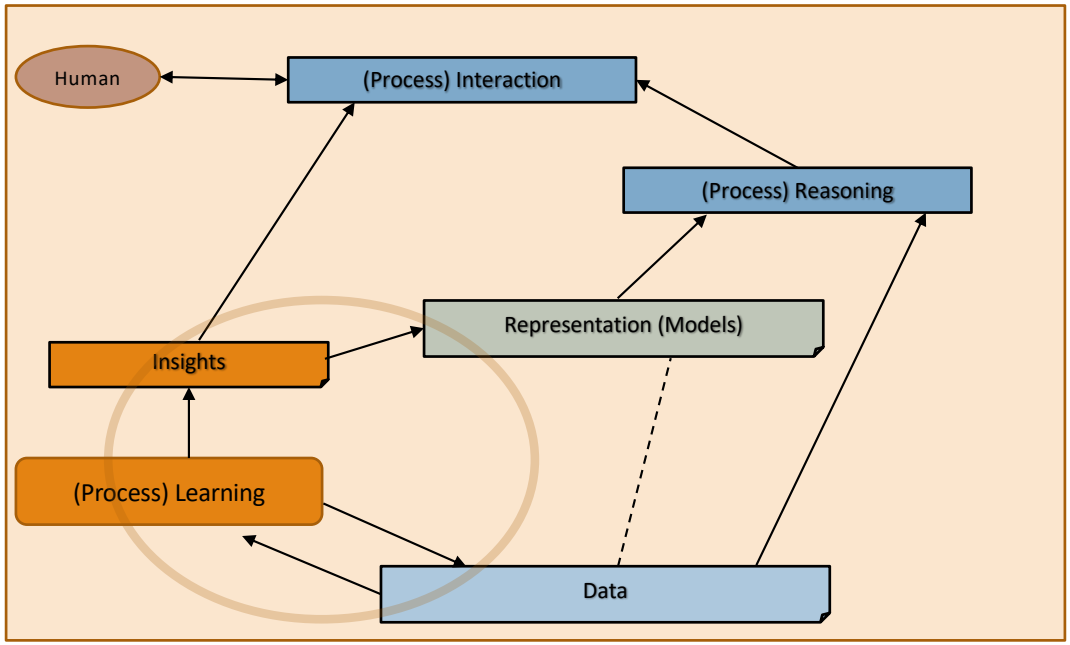
PROF. BIPLAV SRIVASTAVA, AI INSTITUTE

6TH MAR, 2021

Carolinian Creed: "I will practice personal and academic integrity."

Organization of Lecture 23

- Introduction Segment
  - Recap/ Discussion of Lecture 21
- Main Segment
  - Word representation
  - Sentiment Analysis
- Concluding Segment
  - About Next Lecture – Lecture 24
  - Ask me anything



Introduction Segment

Recap of Lecture 22

- Completed review of student projects - interim
- Text summarization

Main Segment

Why Does Word Representation Matter?

- Efficiency
  - Storage
  - Computational speed: e.g., vectors make similarity computation faster
- Effectiveness
  - Computation in the presence of typographical mistake/ noise
  - Output easy for humans to understand

Simple / Discrete Word Representations

- Illustrations
  - (Ad-hoc) integer encoding
  - Label encoding

- **Sample code**:

  https://github.com/biplav-s/course-d2d-ai/blob/main/sample-code/l23-textrepresent/word-representation.ipynb

Give Meaning to Each Dimension in Vector

- Manually assign meaning to dimensions. Examples
  - Dimension by morphological forms
  - Dimension by semantic type: e.g., days of weak

- Issues:
  - Need to decide dimensions ahead of time

Can We Do Better?

- Representations that make computation robust
  - Overcome typographical mistakes
- Learn representation that helps with down-stream tasks
  - Finding similar words
  - Finding words in similar situations
- Learn new properties from external datasources

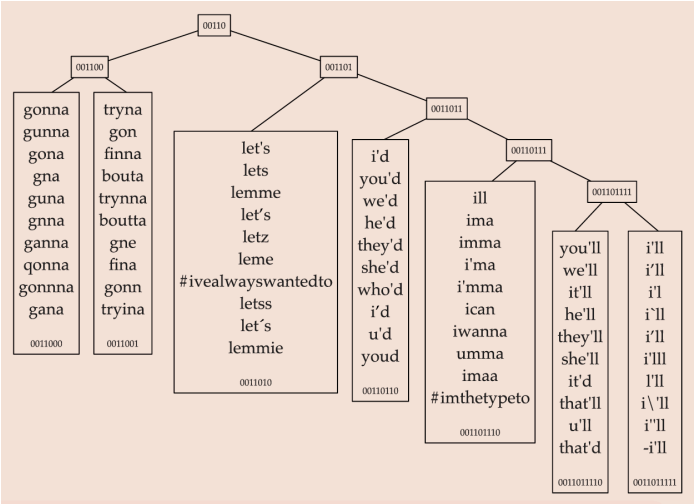
Contextual Word Representations

- Words derives meaning from its context ("**neighborhood**")

- Words with distributional assumptions:
  - Context: given a word, its nearby words or sequences of words
  - Words used in similar ways are likely to have related meanings; i.e., words used in the same (similar) context have related meanings
    - No claim about meaning except relative similarity v/s dis-similarity of words

Contextual Representation by Clustering

- Cluster words by context
- Compare with words in a manually-created taxonomy, e.g., Wordnet



The 10 most frequent words in clusters in the section of the hierarchy with prefix bit string 00110.

Owoputi, O., O'Connor, B., Dyer, C., Gimpel, K., Schneider, N., and Smith, N.A. Improved part-ofspeech tagging for online conversational text with word clusters. In Proceedings of 2013 NAACL.

Credit:
Contextual Word Representations: Putting Words into Computers", by Noah Smith, CACM June 2020

Contextual Representation by Dimensionality Reduction

- Creating word vectors in which each dimension corresponds to the frequency the word type occurred in some context.

Document: https://bit.ly/2B9uaKr
"frequency of each word occurring within two positions on either side of the word whose vector we are constructing"

context words	v(astronomers)	v(bodies)	v(objects)
't			1
,		2	1
.	1		1
1			1
And			1
Belt			1
But	1		
Given			1
Kuiper			1
So	1		
and		1	
are		2	1
between			1
beyond		1	
can			1
contains		1	
from	1		
hypothetical			1
ice		1	
including		1	
is	1		
larger		1	
now	1		
of	1		

Since these hypothetical objects are only between 1 and 10 kilometres in radius (0.6 to 6.2 miles), it's tricky to spot them from where we sit. But now astronomers think they have done it.

- Strategy 1: select contexts
  - Examples
    - Words in the neighborhood
    - Words of specific types
  - Build vectors
  - Use vector operations to derive meaning

Credit:
Contextual Word Representations: Putting Words into Computers", by Noah Smith, CACM June 2020

$$\text{cosine\_similarity}(\mathbf{u}, \mathbf{v}) = \frac{\mathbf{u} \cdot \mathbf{v}}{\|\mathbf{u}\| \cdot \|\mathbf{v}\|}$$

	astronomers	bodies	objects
astronomers	$\frac{14}{\sqrt{14} \cdot \sqrt{14}} = 1$	$\frac{0}{\sqrt{24} \cdot \sqrt{14}} = 0$	$\frac{1+1}{\sqrt{14} \cdot \sqrt{16}} \approx 0.134$
bodies		$\frac{24}{\sqrt{24} \cdot \sqrt{24}} = 1$	$\frac{2+2+2}{\sqrt{24} \cdot \sqrt{16}} \approx 0.306$
objects			$\frac{16}{\sqrt{16} \cdot \sqrt{16}} = 1$

Bodies and **objects** are **most similar** (0.306) than
- **Bodies** and **astronomers** (0)
- **Objects** and **astronomers** (0.134)

TF-IDF Method

- Given N documents
  - For document d, with terms t

- Computer two quantities
  - Term-Frequency (TF) : TF(t, d)
  - Inverse Document Frequency (IDF): **IDF**(t)

- Representation: **TF-IDF**(t, d) = TF(t, d) * IDF(t)

TF-IDF based Word Representation -1

- Given N documents

- **Term frequency (TF):** for term (word) t in document d = tf(t, d)

Variants to reduce bias due to document length

Variants of term frequency (tf) weight

weighting scheme	tf weight
binary	$0, 1$
raw count	$f_{t,d}$
term frequency	$f_{t,d} \Big/ \sum_{t' \in d} f_{t',d}$
log normalization	$\log(1 + f_{t,d})$
double normalization 0.5	$0.5 + 0.5 \cdot \frac{f_{t,d}}{\max_{\{t' \in d\}} f_{t',d}}$
double normalization K	$K + (1 - K)\frac{f_{t,d}}{\max_{\{t' \in d\}} f_{t',d}}$

Sources:
(a) sci-kit documentation
(b) Wikipedia: https://en.wikipedia.org/wiki/Tf%E2%80%93idf

TF-IDF based Word Representation -2

- Given N documents
- Term frequency (TF): for term (word) t in document d
  = tf(t, d)
- **Inverse document frequency IDF**(t)

  = log [N / **DF**(t)] + 1

DF(t) = **document frequency**, the number of documents in the document set that contain the term t.

- **TF-IDF**(t, d) = TF(t, d) * IDF(t),

Variants of inverse document frequency (idf) weight

weighting scheme	idf weight ($n_t = \|\{d \in D : t \in d\}\|$)
unary	1
inverse document frequency	$\log \frac{N}{n_t} = -\log \frac{n_t}{N}$
inverse document frequency smooth	$\log\left(\frac{N}{1+n_t}\right) + 1$
inverse document frequency max	$\log\left(\frac{\max_{\{t' \in d\}} n_{t'}}{1+n_t}\right)$
probabilistic inverse document frequency	$\log \frac{N-n_t}{n_t}$

Sources:
(a) sci-kit documentation
(b) Wikipedia: https://en.wikipedia.org/wiki/Tf%E2%80%93idf

TF-IDF Example Calculation

Sample code:
https://github.com/biplav-s/course-d2d-ai/blob/main/sample-code/l23-textrepresent/word-representation.ipynb

Contextual Representation by Dimensionality Reduction - 1

- Strategy 2: learn contexts from documents. Vector size is given as input

- Train a neural network to learn vector representation
  - value placed in each dimension of each word type's vector is a parameter that will be optimized
  - Selection of parameter values is done using iterative algorithms / gradient descent
  - **Hope** is that ***different senses*** in which a word is used will be captured through the learning procedure as long as the dataset is large enough to represent all senses. Paper quotes: 30 meanings of **get**

- **Optionally**: Sometime task specific inputs are given during pre-processing, processing or post-processing

Disadvantage: individual dimensions are no longer interpretable

Contextual Representation by Dimensionality Reduction -2

- Strategy 2: learn contexts from documents. Vector size is given as input

Sometime task specific inputs are given during pre-processing, processing or post-processing

- Pre-processing
  - Vector initialization by pre-training. Called **finetuning**
- Processing
  - **Knowledge-infusion** (emerging area)
- Post-processing
  - Adjust output vectors so that word types that are related in reference taxonomy (like WordNet) are closer to each other in vector space. Called **retrofitting**.

Credit:
Contextual Word Representations: Putting Words into Computers", by Noah Smith, CACM June 2020

Language Model

Problem:
Given a sentence fragment, predict what word(s) come next

Applications:
- Spelling correction
- speech recognition
- machine translation,
- …

Language Model:
estimate probability of substrings of a sentence

$$P(w_i|w_1, w_2, \ldots, w_{i-1}) = \frac{P(w_1, w_2, \ldots, w_{i-1}, w_i)}{P(w_1, w_2, \ldots, w_{i-1})}$$

Bigram approximation

$$P(w_i|w_1, w_2, \ldots, w_{i-1}) \approx \frac{P(w_{i-1}, w_i)}{P(w_{i-1})}$$

From Jurafsky & Martin

Contextual Word Embeddings

	Name	Description	URL, References
1.	Elmo (embeddings from language models)	Contextual, deep, character-based	https://allennlp.org/elmo; Deep contextualized word representations, Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, Luke Zettlemoyer. NAACL 2018.
2	Word2Vec	Word-based, prediction focus	*Mikolov, Tomas; et al. (2013). "Efficient Estimation of Word Representations in Vector Space". arXiv:1301.3781 [cs.CL]. Mikolov, Tomas (2013). "Distributed representations of words and phrases and their compositionality". Advances in Neural Information Processing Systems. arXiv:1310.4546.*
3	Glove	Word-based, count	https://nlp.stanford.edu/projects/glove/, Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation. [pdf] [bib]
4	Fasttext	Variation of word2vec, works with N-gram, words not in voacbulary	

Commentaries:
https://jalammar.github.io/illustrated-bert/ , https://cai.tools.sap/blog/glove-and-fasttext-two-popular-word-vector-models-in-nlp/

Sentiment Detection

Types of Sentiment Tasks

- Sentiment-oriented Word Embedding
- Sentence-level Models
  - Input: Set of sentences, each made up of a set of words
  - Output: A set of labels (positive, negative, neutral)
- Document-level Models
  - Input: Set of documents, each made up of a set of sentences, each made up of a set of words
  - Output: A set of labels (positive, negative, neutral)
- Fine-grained sentiment labels
  - (e.g., sentiment strength)

A Simple Rule-Based Sentiment Engine

- Process input to get tokens
  - Perform: Stemming, tokenization, part-of-speech tagging and semantic parsing.
- Use lexicons to find polarity of words
- Use a method to aggregate over polarity of words

SentiWordNet

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010 SENTIWORDNET 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. LREC-2010

- Home page: https://github.com/aesuli/SentiWordNet
- All WordNet synsets automatically annotated for degrees of positivity, negativity, and objectivity
- # ObjScore = 1 - (PosScore + NegScore)

Examples (from):
https://raw.githubusercontent.com/aesuli/SentiWordNet/master/data/SentiWordNet_3.0.0.txt

- a 00006032 0.25 0.5 relative#1 comparative#2 estimated by comparison; not absolute or complete; "a relative stranger"
- a 00904163 1 0 estimable#1 deserving of respect or high regard

Source: Jurafsky & Martin

Scherer's Typology of Affective States

Emotion: relatively brief episode of synchronized response of all or most organismic subsystems in response to the evaluation of an event as being of major significance

angry, sad, joyful, fearful, ashamed, proud, desperate

Mood: diffuse affect state …change in subjective feeling, of low intensity but relatively long duration, often without apparent cause

cheerful, gloomy, irritable, listless, depressed, buoyant

Interpersonal stance: affective stance taken toward another person in a specific interaction, coloring the interpersonal exchange

distant, cold, warm, supportive, contemptuous

Attitudes: relatively enduring, affectively colored beliefs, preferences predispositions towards objects or persons

liking, loving, hating, valuing, desiring

Personality traits: emotionally laden, stable personality dispositions and behavior tendencies, typical for a person

nervous, anxious, reckless, morose, hostile, envious, jealous

Source: Jurafsky & Martin

The General Inquirer

Philip J. Stone, Dexter C Dunphy, Marshall S. Smith, Daniel M. Ogilvie. 1966. The General Inquirer: A Computer Approach to Content Analysis. MIT Press

- Home page: http://www.wjh.harvard.edu/~inquirer
- List of Categories: http://www.wjh.harvard.edu/~inquirer/homecat.htm
- Spreadsheet: http://www.wjh.harvard.edu/~inquirer/inquirerbasic.xls

Categories:

- 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 for Research Use

Source: Jurafsky & Martin

Basic Sentiment Analysis

Sample code for TextBlob, Vader:
https://github.com/biplav-s/course-d2d-ai/blob/main/sample-code/l23-textrepresent/Basic%20Sentiment.ipynb

Neural Network Based Methods for Sentiment Classification

Stanford Sentiment Resources

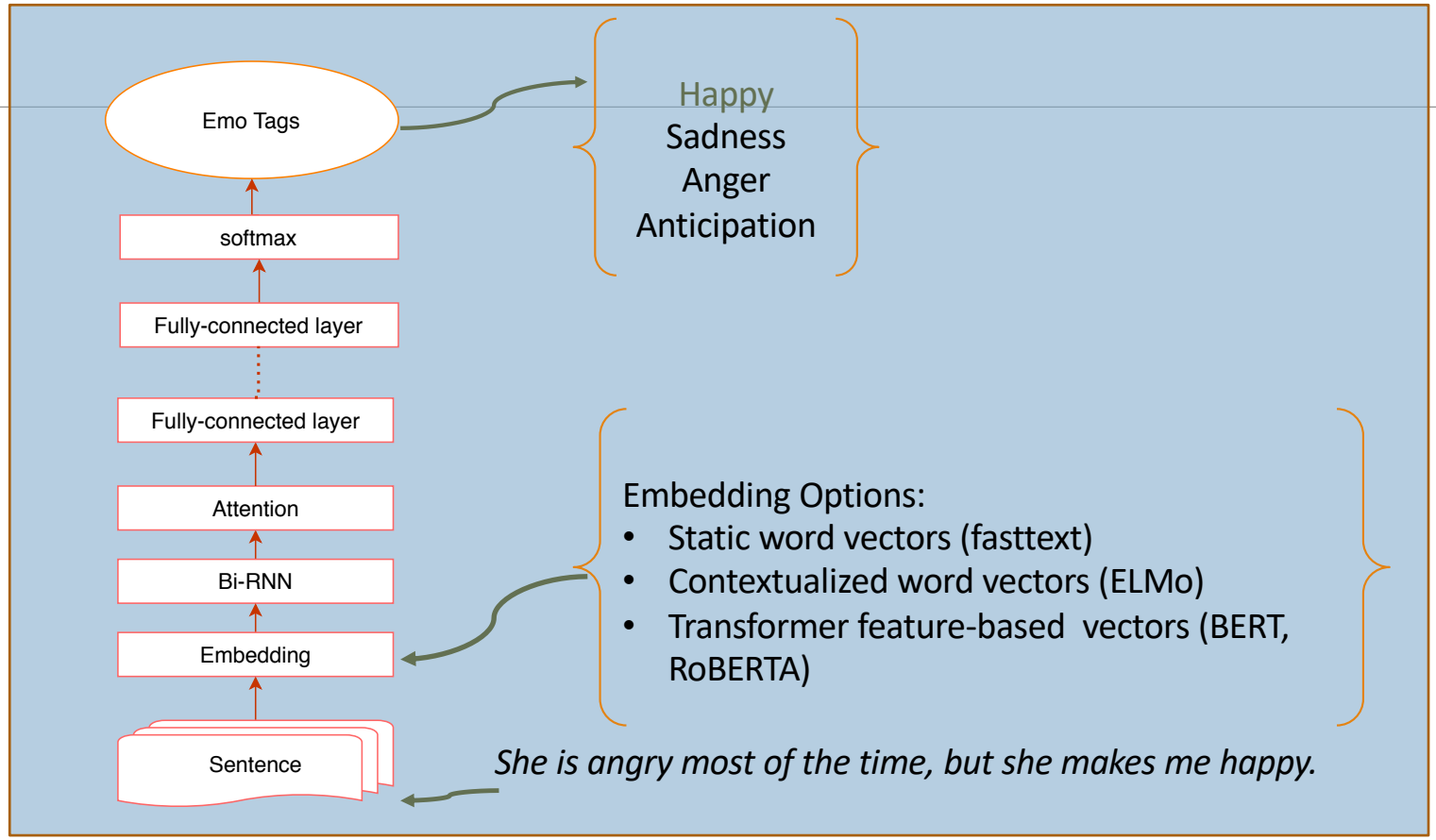
- IMDB:
  - Dataset and paper (ACL 2011)
  - https://ai.stanford.edu/~amaas/data/sentiment/

- Comments
  - The dataset has 50,000 reviews from IMDB, allowing no more than 30 reviews per movie.
  - The dataset has even number number of positive and negative movie reviews.
  - The dataset contains highly polar movie reviews data. A negative review has a score ≤ 4 out of 10, and a positive review has a score ≥ 7 out of 10.
  - The dataset is evenly divided into training and test sets.
  - This dataset is widely used to benchmark new work.

Acknowledgement:

Fawad Kirmani's CSCE 771 project, Fall 202

Automation of Emotion Classification



Slide: Shabnam Tafreshi

Concluding Segment

Lecture 23: Concluding Comments

- We discussed word representation
  - Discrete words
  - Contextual word representation
- Looked at sentiment models
  - Lexicon-based methods
  - NN-based methods

Upcoming Classes

Upcoming Classes

21	**Mar 25 (Th)**	**Review: project presentations, Discussion**	
	Mar 30 (Tu)	Wellness Holiday	
22	Apr 1 (Th)	Text: Text Summarization	
23	Apr 6 (Tu)	Text: Representation, Sentiment	
24	Apr 8 (Th)	Text: Sentiment, Visualization	
25	Apr 13 (Tu)	Advanced: Bias and Trust Issues	Quiz 4
26	Apr 15 (Th)	Paper presentations – Graduate students	Final assignment for Graduate students
27	Apr 20 (Tu)	**Invited Guest** – Javid Huseynov – Case Study: Finance	
28	Apr 22 (Th)	Project presentations	
	Apr 27 (Tu)	Reading day	Reading day
29	Apr 29 (Th)	Project presentations	Final assignment given (undergrad)
30	May 4 (Tu)	Course Recap	Final assignment due (undergrad), Paper summary due (grad)

About Next Lecture – Lecture 24

Lecture 24: More on Text

- Sentiment
  - Problem of bias
  - Using sentiment in business setting
- Text visualization