Why is it hard?
Typical NLP Tasks
Operations on a text Corpus
Why Probability?
Probabilistic Language Model - Definition
VSM for Words

Document Vector Space Model
Semantically connected Word Vectors
Human/Machine Learning
Word embedding
Sequence Learning
Recurrent Neural Network
Various Approaches to MT
Machine Translation

Ramasesh an

Introduction

Ramaseshan Ramachandran

Course requirements

- Computer Science, Probability and statistics 101, Linear Algebra 101, Machine Learning 101 and lots of common sense :)

Ramaseshan

Ability to process and harness information from a large corpus of text with very little manual intervention

WHY IS IT HARD?

- Multiple ways of representation of the same scenario
- Includes common sense and contextual representation
- Complex representation information (simple to hard vocabulary)
- Mixing of visual cues
- Ambiguous in nature
- Idioms, metaphors, sarcasm (Yeah! right), double negatives, etc. make it difficult for automatic processing
- Human language interpretation depends on real world, common sense, and contextual knowledge

Ramaseshan

TYPICAL NLP TASKS

Information Retrieval	Find documents based on keywords
Information Extraction	Identify and extract personal name, date, company name, city..
Language generation	Description based on a photograph Title for a photograph
Text clustering	Automatic grouping of documents
Text classification	Assigning predefined categorization to documents. Identify Spam emails and move them to a Spam folder
Machine Translation	Translate any language Text to another
Grammar checkers	Check the grammar for any language

OPERATIONS ON A TEXT CORPUS

The basic operation on text is *tokenization*. This is the process of dividing input text into tokens/words by identifying word boundary

- Identify paragraphs, sentences
- Extract tokens
- Count the number of tokens/words in the corpus
- Find the vocabulary count
- Find patterns of words
- Term Frequency (TF)
- Type-Token Ratio
- Inverse Document Frequency (IDF)
- Zipf (term frequency $\propto \frac{1}{rank}$)and Mandelbrot hypotheses
- Find co-occurrence of words

Ramaseshan

WHY PROBABILITY?

- Provides methods to predict or make decisions to pick the next word in the sequence based on sampled data
- Make the informed decision when there a certain degree of uncertainty and some observed data
- It provides a quantitative description of the chances or likelihoods associated with various outcomes
- Probability of a sentence
- Probability of the next word in a sentence - how likely to predict "**you**" as the next word
- Likelihood of the next word is formalized through an observation by conducting experiment - counting the words in a document

Ramaseshan

Discrete Sample Space, experiment, joint and conditional probability,

PROBABILISTIC LANGUAGE MODEL

Goal: Compute the probability of a sequence of words

$$P(W) = P(w_1, w_2, w_3, ...w_n) \tag{1}$$

Task: To predict the next word using probability. Given the context, find the next word using

$$P(w_n|w_1, w_2, w_3, \ldots, w_{n-1}) \tag{2}$$

A model which computes the probability for (1) or predicting the next word (2) or complete the partial sentence is called as Probabilistic Language Model.
The goal is to learn the joint probability function of sequences of words in a language. The probability of P(The cat roars) is less likely to happen than P(The cat meows) n-grams are used to build predictive and generative language models

VECTOR SPACE MODEL FOR WORDS

Let us assume that the words in a corpus are considered as linearly independent basis vectors.
If a corpus contains $|\mathscr{V}|$ words which are linearly independent, then every word represents an axis in the continuous vector space $\mathscr{R}$.
Each word takes an independent axis which is orthogonal to other words/axes.
Then $\mathscr{R}$ will contain $|\mathscr{V}|$ axes.

Ramaseshan

Examples

1. The vocabulary size of *emma corpus* is 7079. If we plot all the words in the real space $\mathscr{R}$, we get 7079 axes
2. The vocabulary size of *Google News Corpus corpus* is 3 million. If we plot all the words in the real space $\mathscr{R}$, we get 3 million axes

DOCUMENT VECTOR SPACE MODEL

- Vector space models are used to represent words in a continuous vector space $\mathscr{R}$

Ramaseshan

Binary Incidence Matrix TF-IDF Incidence matrix Query modeling Document similarity Information Extraction Named Entity Recognition

DOCUMENT VECTOR SPACE MODEL

- Vector space models are used to represent words in a continuous vector space $\mathscr{R}$
- Combination of Terms represent a document vector in the word vector space

Ramaseshan

Binary Incidence Matrix TF-IDF Incidence matrix Query modeling Document similarity Information Extraction Named Entity Recognition

DOCUMENT VECTOR SPACE MODEL

- Vector space models are used to represent words in a continuous vector space $\mathscr{R}$
- Combination of Terms represent a document vector in the word vector space
- Very high dimensional space - several million axes, representing terms and several million documents containing several terms

Balamurugan

Binary Incidence Matrix TF-IDF Incidence matrix Query modeling Document similarity Information Extraction Named Entity Recognition

CREATION OF SEMANTICALLY CONNECTED VECTORS

- Identify a model that enumerates the relationships between terms and documents
- Identify a model that tries to put similar items closer to each other in some space or structure
- A model that discovers/uncovers the semantic similarity between words and documents in the latent semantic domain
- Develop a distributed word vectors or dense vectors that captures the linear combination of word vectors in the transformed domain

Ramasesh

WHY DENSE VECTORS?

- Sparse vectors are too long and not very convenient as features machine learning
- Abstracts more than just frequency counts
- It captures neighborhood words that are connected by synonyms
  - Consider these two documents (1) Automobile association (2) car driver
  - Connects the neighbor of Automobile and the neighbor of car
  - "Automobile association" with "car driver" - driver and association could be connected using the similar words ***Automobile and car***

Ramaseshan

HUMAN/MACHINE LEARNING

- How do we solve problems when we lack sufficient knowledge?
- Finding Examples and using experience gained are useful
- Examples provide certain underlying patterns
- Patterns give the ability to predict some outcome or help in constructing an approximate model
- The model may help resolve some problems, though may not be an ideal one
- **Learning** is the key to the ambiguous world
- Linear and non-linear classification
- Perceptron, perceptron learning, cost function, feed forward neural network, back propagation algorithm

Ramaseshan

WORD EMBEDDING

- Process each word in a Vocabulary of words to obtain a respective numeric representation of each word in the Vocabulary
- Reflect semantic similarities, Syntactic similarities, or both, between words they represent
- Map each of the plurality of words to a respective vector and output a single merged vector that is a combination of the respective vectors

Ramaseshan

1. Continuous bag of words (CBOW) Model
2. Skip-gram model
3. Discuss Word2Vec model

SEQUENCE LEARNING

Sequence learning is the study of machine learning algorithms designed for applications that require sequential data or temporal data

Ramaseshan

RECURRENT NEURAL NETWORK

- Sequential data prediction is considered as a key problem in machine learning and artificial intelligence
- Unlike images where we look at the entire image, we read text documents sequentially to understand the content.
- The likelihood of any sentence can be determined from everyday use of language.
- The earlier sequence of words (int time) is important to predict the next word, sentence, paragraph or chapter
- If a word occurs twice in a sentence, but could not be accommodated in the sliding window, then the word is learned twice
- An architecture that does not impose a fixed-length limit on the prior context

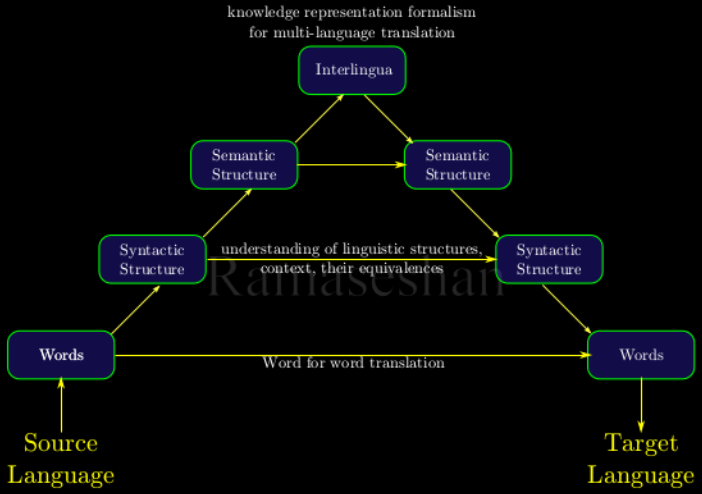
RNN—Language Model—Encoding a sentence into a fixed sized vector—Exploding and vanishing gradients—LSTM—GRU—

Ramaseshan

When I look at an article in Russian, I say "This is really written in English, but it has been coded in some strange symbols. I will now proceed to decode." (Warren Weaver, 1947)

Ramaseshan

VAUQUOIS DIAGRAM - VARIOUS APPROACHES TO MT



[1] Vauquois, B. (1968). "A survey of formal grammars and algorithms for recognition and transformation in machine translation," in Proceedings of IFIP Congress-6, pp. 254-260.

MACHINE TRANSLATION

- The idea of *the ability to make anyone speak to anyone without the boundary of languages* is the most appealing idea
- The goal of the automatic translation is to produce error-free translation
  - Preserve the meaning of the source language
- AMT is a hard problem
- Parallel corpora aids in the development of AMT

Ramasesan

- Translation by analogy: Example based machine translation (EBMT) (lazy learning)

  This is my house - Hii ni nyumba yangu
  My dog loves to run - Mbwa wangu anapenda kukimbia
  I run with my dog - Mimi kukimbia na mbwa wangu
  My house is blue in color - Nyumba yangu ni rangi ya bluu
  This is my dog -

Ramaseshan

- Translation by analogy: Example based machine translation (EBMT) (lazy learning)

  This is my house - Hii ni nyumba yangu
  My dog loves to run - Mbwa wangu anapenda kukimbia
  I run with my dog - Mimi kukimbia na mbwa wangu
  My house is blue in color - Nyumba yangu ni rangi ya bluu
  This is my dog - Hii ni mbwa wangu

Ramaseshan

- Learn MT models from data: Statistical Machine Learning
  - Translation models with language-specific parameters
  - Train model parameters & apply to unseen data

Attention—word and Phrase-based Translations

NEURAL MACHINE TRANSLATION

Neural Machine Translation (NMT) is the mechanism of modeling the Machine translation process using artificial neural network

We could consider translations as a sequence with the source and the destination sentences $((E_{t1}, F_{t2}))$ appearing in a time series. The words within E, F appear in different time $(t_{11}, t_{12}, t_{13}, ..., t_{1n})$ and $(t_{21}, t_{22}, t_{23}, ..., t_{2m})$, respectively

Unlike the phrase-based SMT models, NMT attempts to build and train a single, large neural network that reads a sentence and outputs a correct translation

Ramaseshan

HYPOTHESIS GENERATION

- Dr.Swanson, an information scientist, proved that two distinct knowledge sources together contain implications that cannot be seen within either of the two sets by using an independent lens
- He had shown that how seemingly unconnected resources could be combined to form a new hypothesis, though he was not an expert in all the fields that he chose to converge
- Stress is associated with migraines
- Stress can lead to loss of magnesium
- Calcium channel blockers prevent some migraines
- Magnesium is a natural calcium channel blocker
- Spreading cortical depression (SCD) is implicated in some migraines
- High levels of magnesium inhibit SCD
- Migraine patients have high platelet aggregation
- Magnesium can suppress platelet aggregation

Ramaseshan

SUMMARY

1. Gather statistical information about corpus, words
2. Understanding words from context or context words from a word
3. Learn to encode the contextual information about a word
4. Predict the next word based on the context
5. Learn to encode a sentence - understanding the context of a sentence
6. Predict how likely a new sentence could be a valid sentence
7. Learn to automatically translate from one language to another

Ramaseshan

Ramaseshan

PREPROCESSING

- Corpus creation
- Modifications required to make the text suitable for processing
- Understanding the problem is very crucial in choosing a preprocessing steps
- Preprocessing steps are unique to a problem
- Improper preprocessing schemes may lead to loss of lexical context

Ramaseshan

COMMONLY USED PREPROCESSING STEPS FOR ENGLISH

Preprocessing consists of (a) tokenization, (b) normalization and (c) substitution

- Case folding - Convert all text into lower case
- Stemming - running $\rightarrow$ run
- Lemmatization - best $\rightarrow$ good
- Remove misspellings
- Punctuations

- white space, newline, tabs...
- Removing contractions - isn't $\rightarrow$ is not, I'd $\rightarrow$ i would
- Remove scripts, form variables, for HTML and XML
- Tokenization

Ramaseshan

PREPROCESSING - SPECIFIC TO THE PROBLEM

Can a machine solve problems all by itself by using the advancements in ML, AI and NLP and later explain it to the student in his/her native language?
AInstein

Ramaseshan

A kangaroo is capable of jumping to a height of 2.62 m. Determine the takeoff speed of the kangaroo.

Ramaseshan

Classifier

The most important step - to find out which problem class this text belongs to

Meta-knowledge
Class = Kinematics + Against Gravity

Question Resolver

Independently identifies possible questions from the meta knowledge. It matches with the problem question and determines the question type - Direct or indirect

Meta-knowledge
Class = Kinematics + Against Gravity
Acceleration = -9.8 ms^{-2}
Final velocity= 0.0 ms^{-1}
Height = 2.62 m
Direction = up
Direct question
Takeoff Speed = ? ms^{-1}

Context Clues

Close reading engine identifies the context and suggests missing information that could be used to find the solution

Meta-knowledge
Class = Kinematics + Against Gravity
Acceleration = -9.8 ms^{-2}
Final velocity= 0.0 ms^{-1}

Formula Identifier

$v=u+at$
$v=s/t$
$s=ut+0.5at^2$
$v^2=u^2+2as$

FI finds the right formula from the known quantities

The formula $v^2=u^2+2as$ fits the need. All quantities except initial velocity, u, are known.

$u=\sqrt{v^2-2as}$ ms^{-1}

Meta-knowledge
Class = Kinematics + Against Gravity
Acceleration = -9.8 ms^{-2}
Final velocity = 0.0 ms^{-1}
Height = 2.62 m
Direction = up
Direct question
Takeoff Speed = ? ms^{-1}

{Q} Known Quantities

Valid number and SI unit are combined. Meter refers to distance and height "is-a" distance

Meta-knowledge
Class = Kinematics + Against Gravity
Acceleration = -9.8 ms^{-2}
Final velocity= 0.0 ms^{-1}
Height = 2.62 m
Direction = up

Solution

Takeoff speed = 7.17 ms^{-1}

PREPROCESSING SAMPLE - FSHARP CODE

```
(true,PPSError.NoRegexError,(   doc |>
                                removeNewLine |>
                                removeHints |>
                                removeMathMinus |>
                                separateNumberEqualToSignBySpace |>
                                removeDashBetweenNumberAlpha |>
                                separateNumberAlphaBySpace |>
                                replaceTenPowerNotationToENotation |>
                                fixMSPattern |>
                                regexReplace directionReplacementTuples |>
                                regexReplace unitReplacementTuples |>
                                identifyAndMarkDegree |>
                                singlespace).ToLower())
```

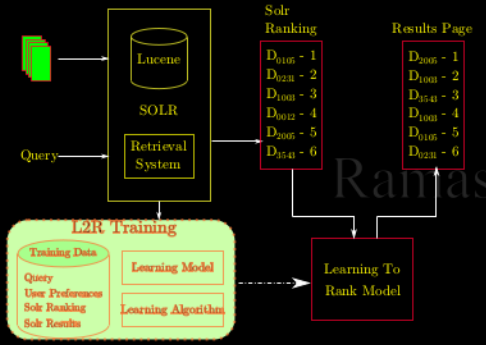
CORPUS

- An airplane accelerates down a runway at **3.20m/s2** for **32.8 s** until is finally lifts off the ground. Determine the distance traveled before takeoff
- How far will a car travel in **25 min** at **12 kmh**?
- A jalopy with an initial speed of 23.7 km/h accelerates at a uniform rate of **0.92 m/s2**Â for **3.6 s**. Find the final speed and the displacement of the jalopy during this time
- A college student wants to toss a textbook to his roommate who is leaning out of a window directly above him. He throws the book upwards with an initial velocity of **8.0 m/s**. The roommate catches it while it is traveling at **3.0 m/s** [up]. a) How long was the book in the air? b) How far vertically did the book travel? A car accelerates in a straight line from rest at the rate of **2.3 m/s 2** . What is its final velocity after **55 m**? What is its time?
- 4. What height will a dart achieve **7 seconds** after being blown straight up at **50 m/s**?

PROCESSED CORPUS

- An airplane accelerates down a runway at **3.20** $< m/s**2 >$ for **32.8** $< s >$ until is finally lifts off the ground. Determine the distance traveled before takeoff
- How far will a car travel in **25** $< minute >$ at **12** $< km/h >$?
- A jalopy with an initial speed of **23.7** $< km/h >$ accelerates at a uniform rate of **0.92** $< m/s**2 >$Â for **3.6** $< s >$. Find the final speed and the displacement of the jalopy during this time
- A college student wants to toss a textbook to his roommate who is leaning out of a window directly above him. He throws the book upwards with an initial velocity of **8.0** $< m/s >$. The roommate catches it while it is traveling at **3.0** $< m/s >$ [up]. a) How long was the book in the air? b) How far vertically did the book travel? A car accelerates in a straight line from rest at the rate of **2.3** $< m/s**2 >$. What is its final velocity after **55 m**? What is its time?
- 4. What height will a dart achieve **7** $< s >$ after being blown straight up at **50** $< m/s >$?

HTML PREPROCESSING



Ramaseshan

1. Convert HTML to text
2. Case folding - convert content to lower case
3. Remove scripts
4. Tokenize
5. Term Frequency
6. Extract Vocabulary
7. Remove stop words - not for Language modeling
8. Stemming/Lemmatization

Ramaseshan