2019 Predictive Analytics Symposium

Session 23: ALL - Natural Language Processing in the Insurance Industry

SOA Antitrust Compliance Guidelines
SOA Presentation Disclaimer

Natural Language Processing in the Insurance Industry

MengChu Tsai

Data Scientist @ Maryville University of st. louis

What is Natural Language Processing

Natural language: A language that is used for communication by humans such as English, Chinese, or German.

Natural Language Processing (NLP) is a branch of artificial intelligence, has been developed to help computers understand, interpret and generate human language.

NLP Application in Insurance Industry

- Claims Processing
- Fraud Detection
- Sentiment Analysis
- Chatbots for Automating Appointments and Choosing a Policy

Text Preprocessing

Noise Removal

Normalization

- Lowercasing
- Stemming/Lemmatization
- Remove Stop Words
- Break Up Contractions
- Convert Numbers to Textual Representation
- Remove Punctuations

Tokenization

Noise Removal

- Removing HTML, XML, etc. markup and metadata
  - ‘ Maryville University <!-- Comment --> Data Science’

    → ‘Maryville University Data Science’
- Extracting plain text from other formats, such as JSON
  - [{"\"Maryville": "University\""}, {"\"Data": "Science\""}]

    → "Maryville University" "Data Science"

Lowercasing

- Without lowercase, then we may have 2 separate data points for the same word.
  - Examples:
    - MARYVILLE, Maryville, maryville -> Maryville
    - CAT, Cat, cat -> cat
- One downside to this is some words mean different things depending on their capitalization.
  - Example: US -> us They have very different purposes.

Stemming/ Lemmatization

- Stemming/ Lemmatization are both generate the root form of the inflected words.
- Stemming follows an algorithm with steps to perform on the words which makes it faster.
  - Examples: care, cares, cared, caring-> car
- Lemmatization removes inflections by mapping a word to its root form.
  - Examples: care, cares, cared, caring-> care

Stop Word Removal

- Stop words are a set of commonly used words in a language and don't contain low information. By removing Stop words from text, we can focus on the important words instead.
  - Examples: it, her, a, an, with, over, into, about.
- Modeler can define domain specific Stop word list.
  - Examples: Analyze comments for course evaluation. Professor, Prof. appear very frequently so it was added to stop word list.

More Data Cleaning

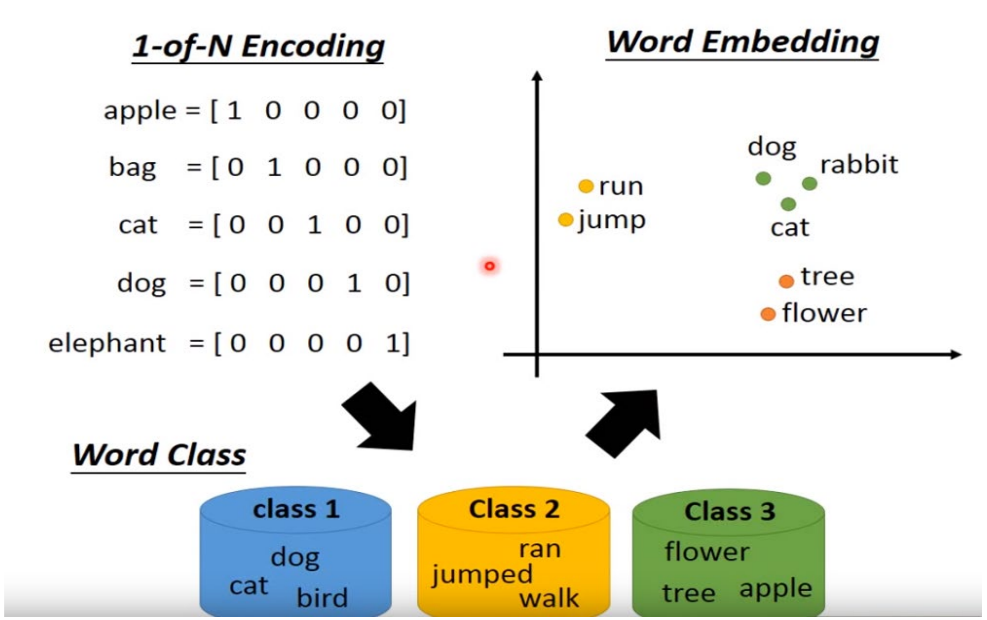
- Break Contractions
  - Examples: Don't -> Do not
- Removing Punctuations and Misspells
- Convert Numbers to Textual Representation
  - 10 -> ten

Tokenization

- Tokenization is a step which splits longer strings of text into smaller pieces, or tokens. Larger chunks of text can be tokenized into sentences, sentences can be tokenized into words.
  - Example:

    Throughout your studies at Maryville Data Science Program, you can expect a student-centered, academically rigorous and market-relevant education focused on your personal and career goals.

    [‘studies’, ‘maryville’, ‘science’, ‘program’, ‘expect’, ‘student’, ‘center, ‘academic’, ‘rigor’ ‘market’ ‘relevant’, ‚educate’, ‘focus’, ‘person’, ‘career’, ‘goal’]



A Simplified Representation of Word Vector

Word vectors	Dimensions				
dog	-0.4	0.37	0.02	-0.34	animal
cat	-0.15	-0.02	-0.23	-0.23	domesticated
lion	0.19	-0.4	0.35	-0.48	pet
tiger	-0.08	0.31	0.56	0.07	fluffy
elephant	-0.04	-0.09	0.11	-0.06	
cheetah	0.27	-0.28	-0.2	-0.43	
monkey	-0.02	-0.67	-0.21	-0.48	
rabbit	-0.04	-0.3	-0.18	-0.47	
mouse	0.09	-0.46	-0.35	-0.24	
rat	0.21	-0.48	-0.56	-0.37	

Word Embedding

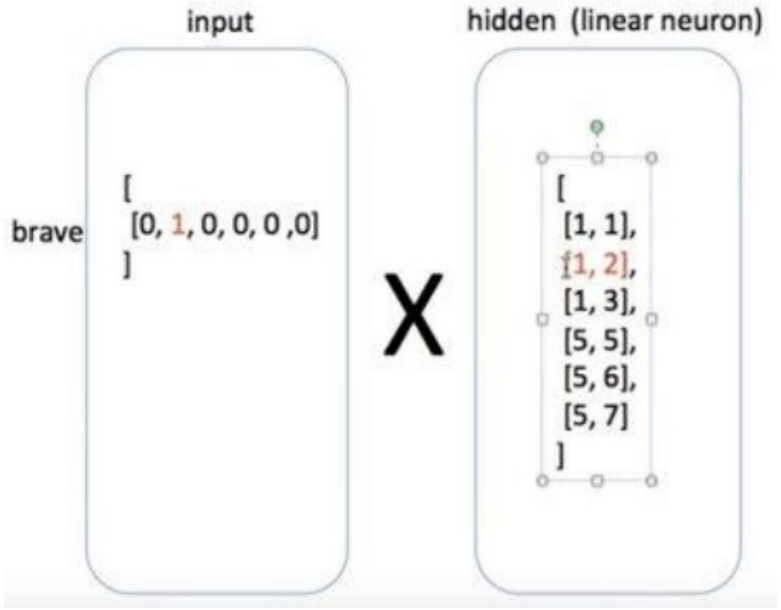
1.Frequency based Embedding: GloVe by Stanford University
2.Prediction based Embedding: Word2Vec by Google

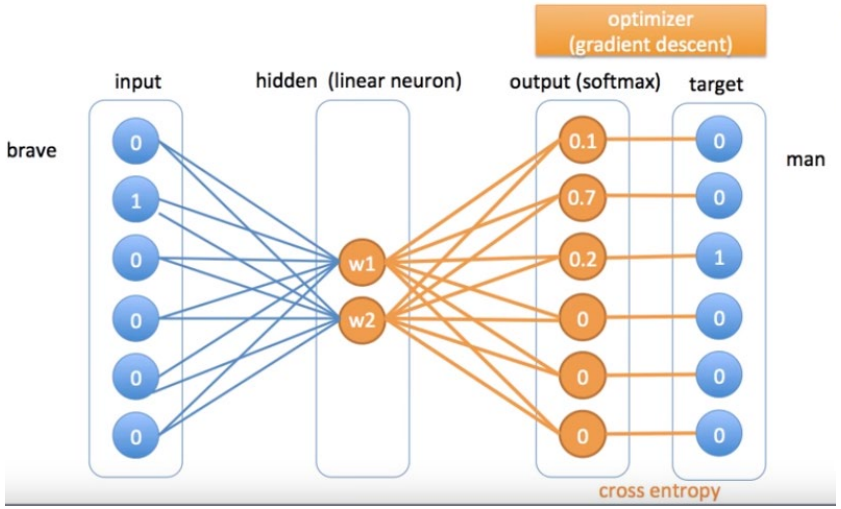
Example (Prediction based) :
“king”, “brave”, “man”, “queen”, “beautiful”, “woman”

word	neighbor
king	brave
king	man
brave	king
brave	man
man	king
man	brave
queen	beautiful
queen	woman
beautiful	queen
beautiful	woman
woman	queen
woman	beautiful

word	word one hot encoding	neighbor	neighbor one hot encoding
king	[1, 0, 0, 0, 0, 0]	brave	[0, 1, 0, 0, 0, 0]
king	[1, 0, 0, 0, 0, 0]	man	[0, 0, 1, 0, 0, 0]
brave	[0, 1, 0, 0, 0, 0]	king	[1, 0, 0, 0, 0, 0]
brave	[0, 1, 0, 0, 0, 0]	man	[0, 0, 1, 0, 0, 0]
man	[0, 0, 1, 0, 0, 0]	king	[1, 0, 0, 0, 0, 0]
man	[0, 0, 1, 0, 0, 0]	brave	[0, 1, 0, 0, 0, 0]
queen	[0, 0, 0, 1, 0, 0]	beautiful	[0, 0, 0, 0, 1, 0]
queen	[0, 0, 0, 1, 0, 0]	woman	[0, 0, 0, 0, 0, 1]
beautiful	[0, 0, 0, 0, 1, 0]	queen	[0, 0, 0, 1, 0, 0]
beautiful	[0, 0, 0, 0, 1, 0]	woman	[0, 0, 0, 0, 0, 1]
woman	[0, 0, 0, 0, 0, 1]	queen	[0, 0, 0, 1, 0, 0]
woman	[0, 0, 0, 0, 0, 1]	beautiful	[0, 0, 0, 0, 1, 0]

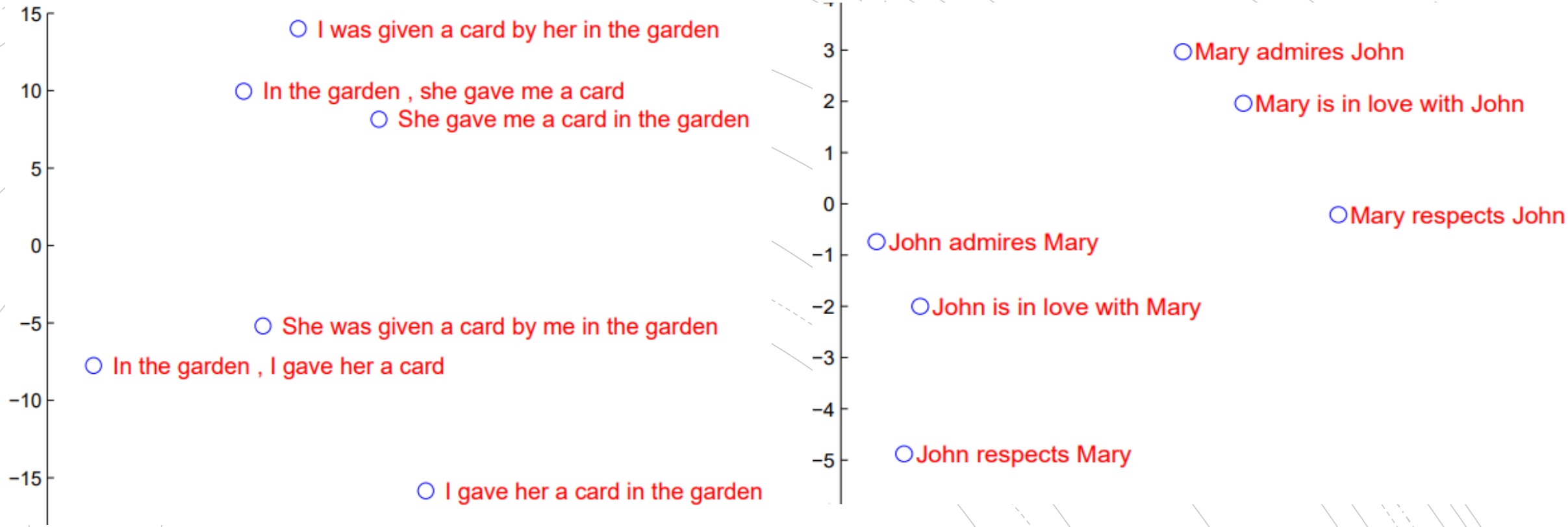
Word Embedding





unique word	encoding	word2vec embedding
king	[1, 0, 0, 0, 0, 0]	[1, 1]
man	[0, 0, 0, 1, 0, 0]	[1, 3]
queen	[0, 0, 0, 1, 0, 0]	[5, 5]
woman	[0, 0, 0, 0, 0, 1]	[5, 7]

Word Embedding



Sentence Embedding

Pre-trained Word Embedding Model

Glove : It uses a *co-occurence* counts matrix to make the embeddings. The matrix values represent the frequency a word appears in a given context. Then, dimensionality reduction is applied to this matrix to create the resulting embedding matrix.

Word2Vec : It is a predictive model, it trains by trying to predict a target word given a context or the context words from the target.

Fasttext : It is an extension and supposedly improvement of the vanilla Word2Vec model. Unlike Word2Vec that considers a word as a single entity, Fasttext considers each word as a Bag of Character n-grams. When an OOV(out of vocabulary) word is encountered it will try and build a vector by summing up subword vectors that would make up the word.

Model Comparison

Dataset contains 500,000 reviews text and rating (1-5 Stars).

Model: Long-Short Term Memory Network (Recurrent Neural Networks)

NLP Tools: Gensim, NLTK

Model Comparison:

- Model_1 with pre-trained model (GloVe)
- Model_2 without pre-trained model
- Use the same neural network architecture
- The result shows the accuracy of Model_1 is 80.00% and the accuracy of Model_2 is 70.06%

Thank you!