Internship 2020

A.L.I.C.E.

Cornelius Yap, Tiffany Goh, Samuel Lye

A.L.I.C.E. Team

Cornelius Yap

- University: Singapore University of Technology and Design
- Major: Computer Science
- Interest: Touch Rugby, Game Design, Travelling, Food

Tiffany Goh

- University: Singapore University of Technology and Design
- Major: Computer Science
- Interests: Swimming, Interacting with animals and young children, Algorithms, Travelling

Samuel Lye

- University: Cornell University
- Major: Economics & Information Science, Minor: Computer Science
- Interests: Running, Exercising, Ultimate Frisbee, Coding, Travelling, Photography

Agenda

1. **Overview of A.L.I.C.E.**
   - Objectives
   - Background

2. **Proposed Implementation Plan**
   - Key components
   - Algorithms Used
   - Sample Use Case
   - Visualization - Input/Output

3. **Challenges / Enhancements for MP**

4. **Exploring State of the Art NLP Technologies**
   - Neural Networks
   - Attention Modelling

Objectives

a) To study and research into the most optimal NLP and visualization tools to create an informative text summarizer.

b) To conduct Proof of Concepts (POC) testing through building of various NLP models using open source tools.

c) To develop greater understanding and knowledge of fundamental machine learning concepts and NLP tools to create a functioning end product.

Background

- Automatic text summarization methods are greatly needed to address the ever-growing amount of text data available online to both better help discover relevant information and to consume relevant information faster.

- Analyzing Language Interface Created for Everyone (A.L.I.C.E.) is purposed to summarize text documents and output informative visualization displays that quickly and easily communicate the contents of the text to the user.

- The key components of A.L.I.C.E. are the Frontend (React, EngineX), Backend (Flask, UWSGI, and Machine Learning models), as well as the Container (Docker).

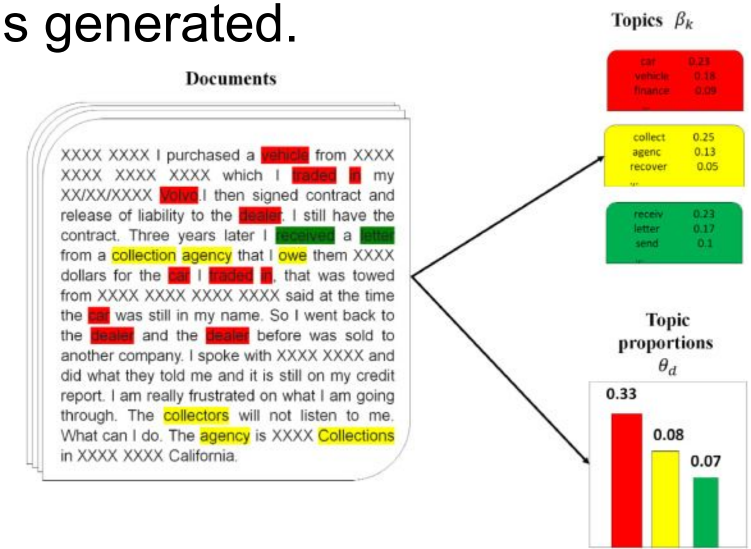
Key Components

- Visualisation
  - D3
  - Nivo
  - Word Cloud
- Front End
  - React
- Back End
  - NgineX
  - MongoDB
  - Flask
  - NLP Features
- NLP Features
  - Topic Modelling
  - NER (Named Entity Recogniser)
  - Sentiment Analysis
  - Relationship Extraction
  - Classification (LSTM)
  - Clustering
- DevOps
  - Dockers
  - Kubernetes

Algorithms Used

Topic Modelling

- Latent Dirichlet Allocation (LDA) and TF-IDF
- User has to key in the number of topics to be identified. TF-IDF runs through the text to identify the keywords in the document
- LDA then runs through all the keywords and find the probability of each particular word belonging to a topic. It then groups the words with the highest probability of belonging to the topic.
- The user can then infer what the topics are from the bags-of-words generated.



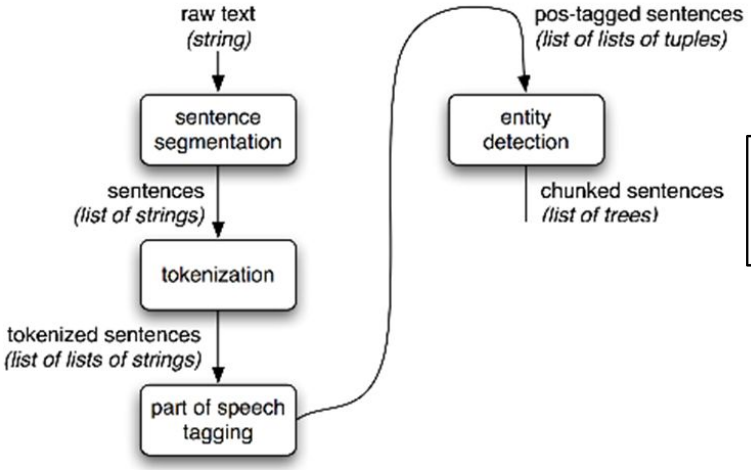
Documents

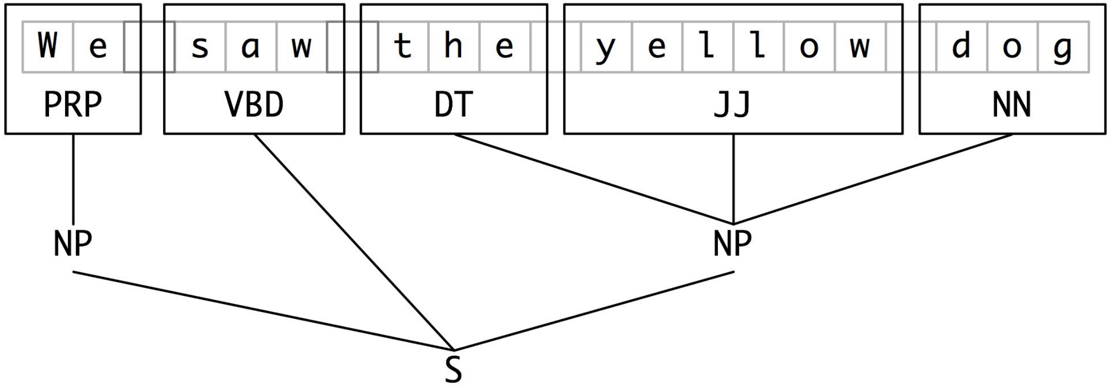
XXXX XXXX I purchased a vehicle from XXXX XXXX XXXX XXXX which I traded in my XX/XX/XXXX Volvo.I then signed contract and release of liability to the dealer. I still have the contract. Three years later I received a letter from a collection agency that I owe them XXXX dollars for the car I traded in, that was towed from XXXX XXXX XXXX XXXX said at the time the car was still in my name. So I went back to the dealer and the dealer before was sold to another company. I spoke with XXXX XXXX and did what they told me and it is still on my credit report. I am really frustrated on what I am going through. The collectors will not listen to me. What can I do. The agency is XXXX Collections in XXXX XXXX California.

Algorithms Used

Named Entity Recognizer (NER)

- NLTK Library
- We use NTLK to chunk the text to identify noun phrases with the POS Tag pattern “NP: {<DT>?<JJ>*<NN>}”.
- After the noun phrases are identified, it then uses machine learning to determine if the noun phrases is an entity based on previously trained datasets.

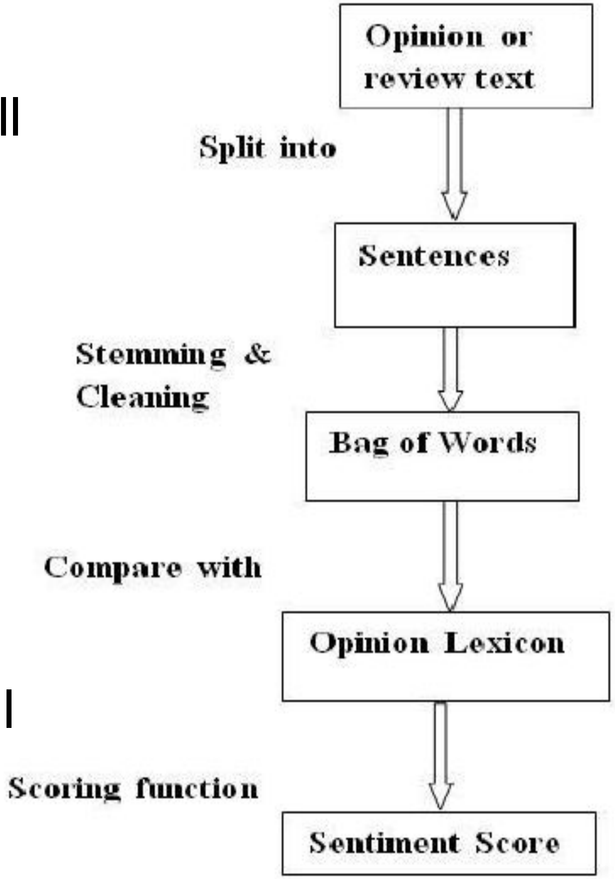




Algorithms Used

Sentiment Analysis

- NLTK Sentiment Analysis Libraries, TextBlob
- The NLTK Sentiment Analysis Libraries as well as TextBlob uses an unsupervised lexicon based method to determine the polarity and objectivity of the text.
- It maintains a lexicon which is a dictionary of words that are given a score based on how objective the word is and whether it is a positive or negative word.
- The NLTK tool then searches the scores for all the words in the text and calculate the overall polarity and objectivity of the text.



Opinion or review text

Split into

Sentences

Stemming & Cleaning

Bag of Words

Compare with

Opinion Lexicon

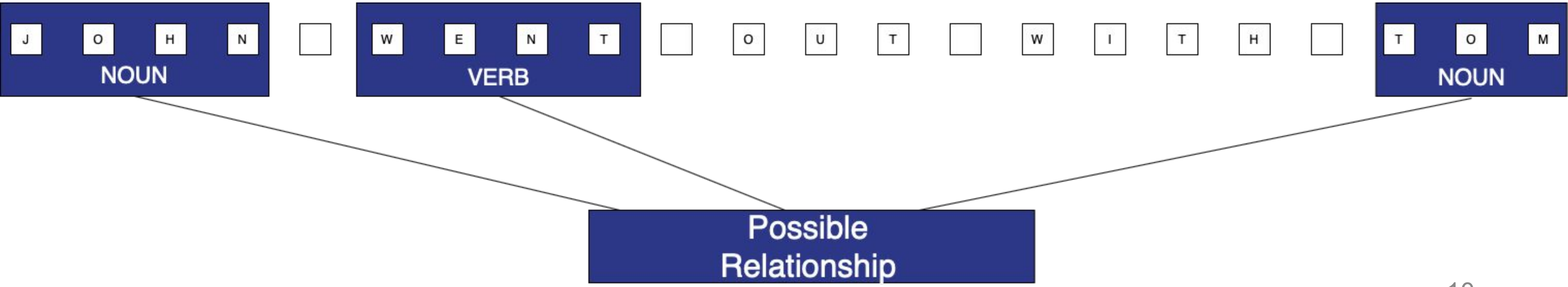
Scoring function

Sentiment Score

Algorithm Used

Relationship Extraction

- As there are no available libraries out there that can helps us perform relationship algorithms, we will have to come up with a custom algorithm ourselves
- Our algorithm aims to identify and pick up the relationship between two entities based on the contextual clues given in a single sentence. We will be using Stanford NLP Parser to POS-Tag the tokens in the given sentence and use the Tags to identify certain patterns that might indicate a relationship.



J O H N □ W E N T □ O U T □ W I T H □ T O M

NOUN

VERB

NOUN

Possible Relationship

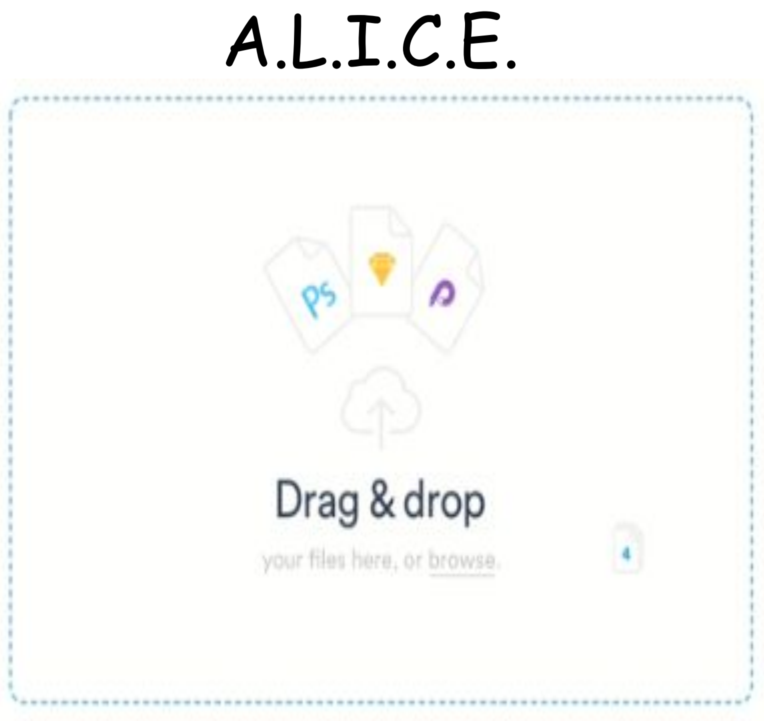
Sample Use Case

Analysis of news articles to monitor for possible Covid-19 cure

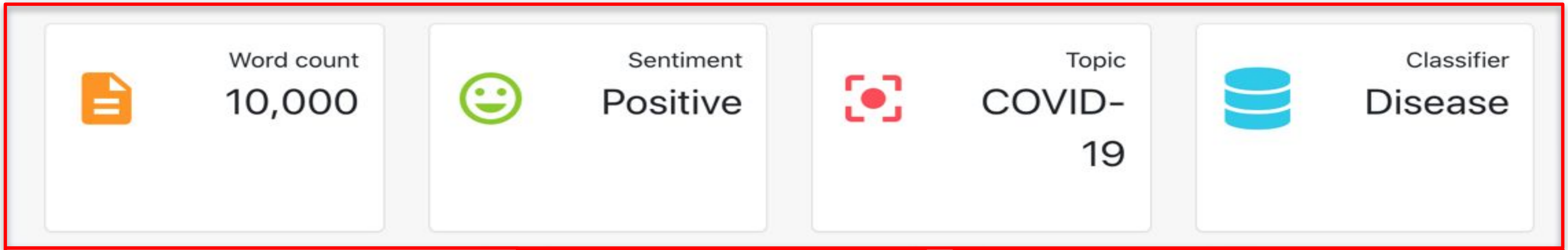
“A team of local researchers are working on a Covid-19 vaccine that can be modified within three weeks to tackle the Sars-CoV-2 virus if it mutates, and they are hoping for human trials within six months.Home-grown Esco Aster is developing the vaccine with United States biotechnology company Vivaldi Biosciences, tapping chimeric vaccines which are created by merging proteins from different viruses.The work-in-progress vaccine, currently named Esco Aster DeltaCov, was constructed by joining antigens from the Sars-CoV-2 virus - which causes Covid-19 - with a protein backbone from the flu virus.”

Visualisation (Input)

- We envisage ALICE to have a simple interface to provide a transformed user experience from the conventional designs.
- Nevertheless, inputs for A.L.I.C.E. require users to select the environment of the ML models, i.e. **domain, number of topics, and number of words per topic** and this will be done in a settings page
  - Different NER models are trained on different domains and thus, have different output
  - Topic modelling requires the input to specify the number of topics and words per topic to generate an output



Visualisation (Output)

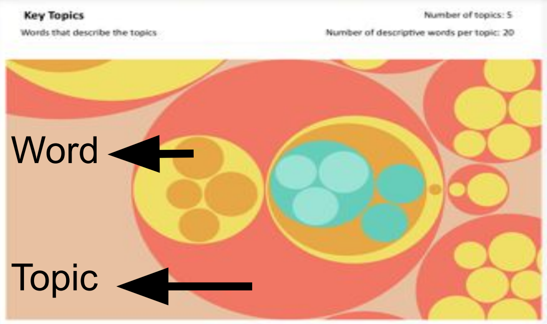


Word count
10,000

Sentiment
Positive

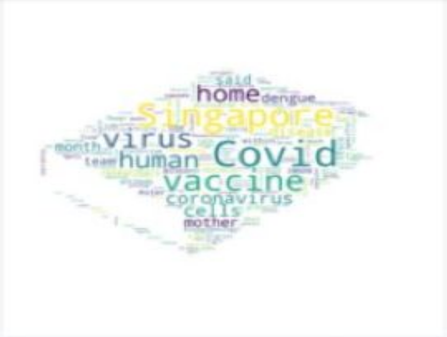
Topic
COVID-19

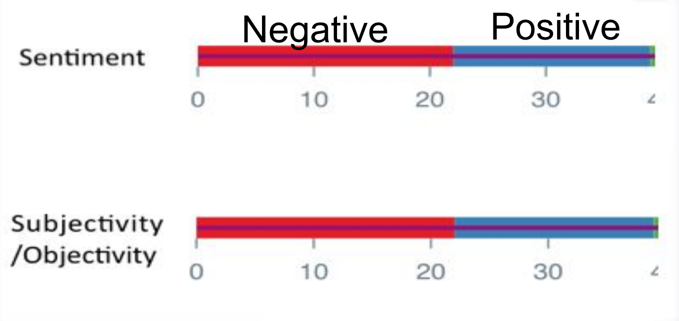
Classifier
Disease

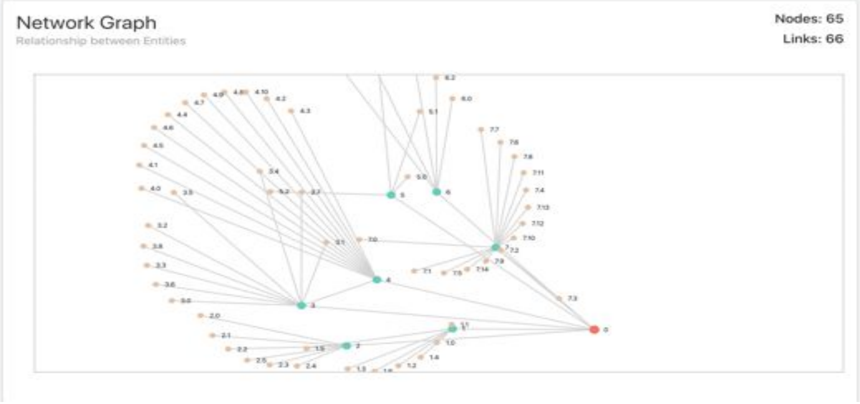


Key Topics
Words that describe the topics

Number of topics: 5
Number of descriptive words per topic: 20

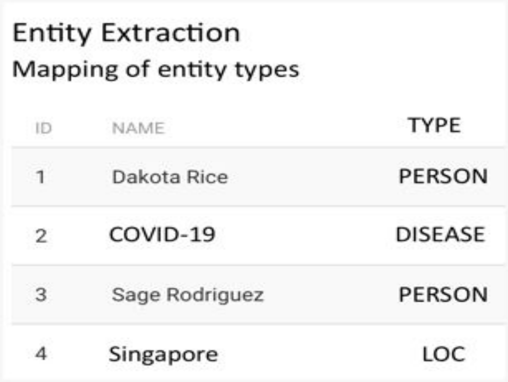






Network Graph
Relationship between Entities

Nodes: 65
Links: 66



Entity Extraction
Mapping of entity types

ID	NAME	TYPE
1	Dakota Rice	PERSON
2	COVID-19	DISEASE
3	Sage Rodriguez	PERSON
4	Singapore	LOC

Application

1. Quick summary of documents from different domains

2. Comprehensive visualisation tool that maps out the relationship between different entities in the text

3. Applicable to the security domain, where our model if trained on security-related domain, would be able to extract key information from sensitive documents and allow users to quickly understand the security concerns with regards to the documents.

Challenges

1. Different domains of text documents will mean that different ML models will have to be trained under those domains with a large enough dataset.

2. We need to train and create our own relation detection model as there is no extensive NLP library that provides this service. This would involve annotating a large amount of test and training data, as well as creating the resulting network diagram.

3. We need to ensure that A.L.I.C.E. is able to effectively “tell a story” and provide a flow of information to the user.

Enhancements for MP

- **Human Computer Interaction**

- Explore/implement HCI techniques to enhance user experience

- **Clustering**

- Given a group of documents, group documents with similar topics
- Using Tf-idf as a distance measurement
- The algorithm used is K means

- **Attention**

- Used to improve the accuracy of the classifier

- **Dockers**

- Implement Containers to provide a consistent environment for all the different components, libraries and dependencies

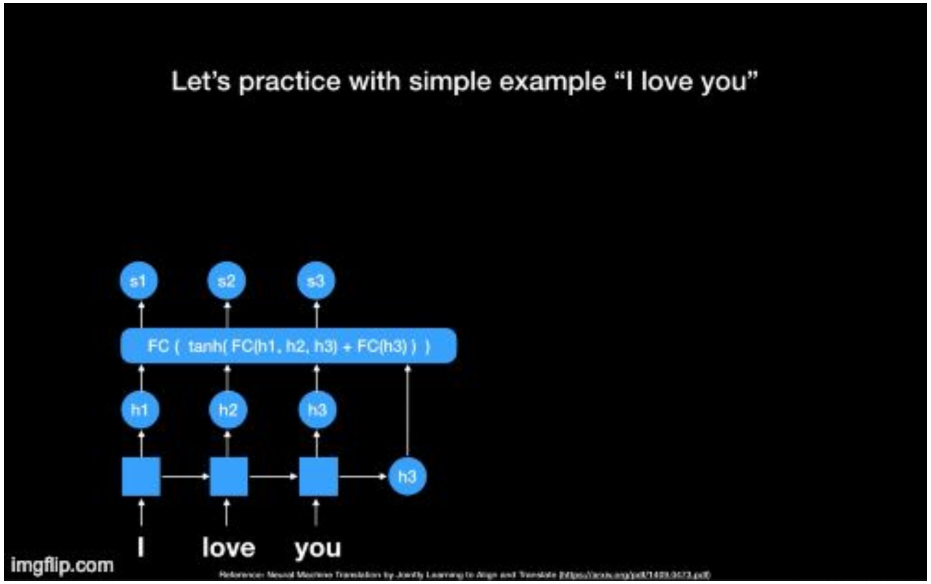
Recurrent Neural Networks

Multilayer Perceptrons (MLP)	**Convolutional Neural Networks (CNN)**	**Recurrent Neural Networks (RNN)**
Designed for classification and regression prediction problems	Designed to map image data to an output variable	Designed to work with sequence prediction problems

- An RNN is different as it introduces the concept of memory through the form of a different type of reverse links.
- The inclusion of links between layers in the reverse direction allows for feedback loops, which are used to help learn concepts based on context.
- Long Short-Term Memory Network (LSTM) is a type of RNN that is capable of learning long-term relationships.

Attention Model

Recurrent Neural Network (RNN)	Attention Model
- Uses fixed length vector ⊖ Rigidity	- Uses context vector ⊕ Flexibility



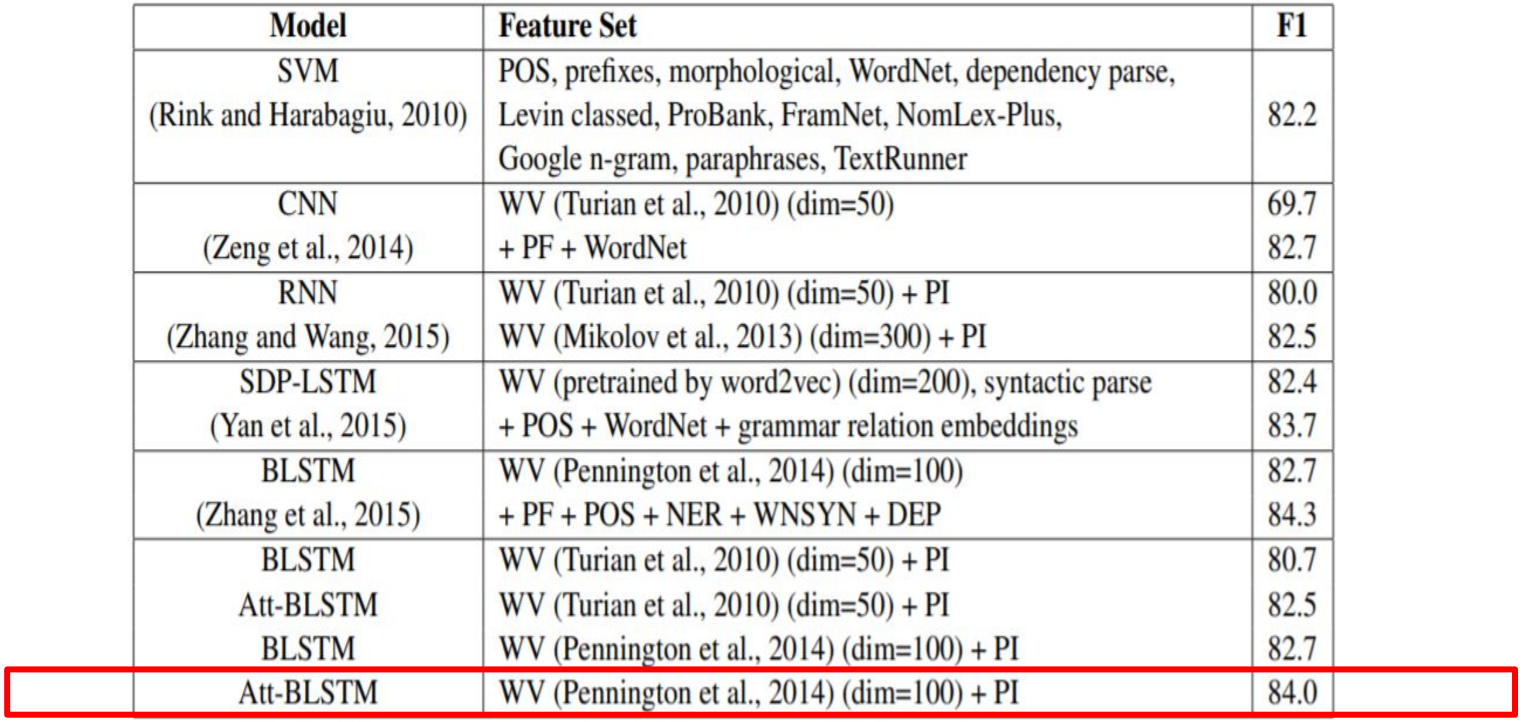
Let's practice with simple example "I love you"

Reference: Neural Machine Translation by Jointly Learning to Align and Translate (https://arxiv.org/pdf/1409.0473.pdf)

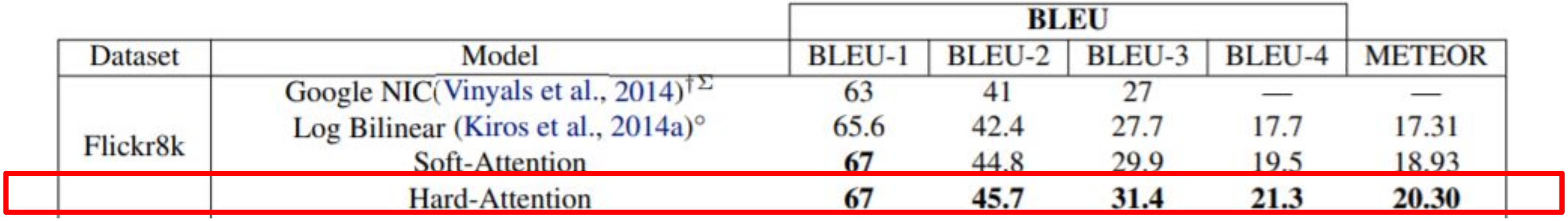
Applications of Attention Models

- Neural Image Caption Generation with Visual Attention
- Hierarchical Attention Network for Document Classification
- Attention-based Bidirectional Long Short-Term Memory Networks for Relation Classification
- Effective Approaches to Attention-based Neural Network Translation

Comparison between models



Model	Feature Set	F1
SVM (Rink and Harabagiu, 2010)	POS, prefixes, morphological, WordNet, dependency parse, Levin classed, ProBank, FramNet, NomLex-Plus, Google n-gram, paraphrases, TextRunner	82.2
CNN (Zeng et al., 2014)	WV (Turian et al., 2010) (dim=50) + PF + WordNet	69.7 82.7
RNN (Zhang and Wang, 2015)	WV (Turian et al., 2010) (dim=50) + PI WV (Mikolov et al., 2013) (dim=300) + PI	80.0 82.5
SDP-LSTM (Yan et al., 2015)	WV (pretrained by word2vec) (dim=200), syntactic parse + POS + WordNet + grammar relation embeddings	82.4 83.7
BLSTM (Zhang et al., 2015)	WV (Pennington et al., 2014) (dim=100) + PF + POS + NER + WNSYN + DEP	82.7 84.3
BLSTM	WV (Turian et al., 2010) (dim=50) + PI	80.7
Att-BLSTM	WV (Turian et al., 2010) (dim=50) + PI	82.5
BLSTM	WV (Pennington et al., 2014) (dim=100) + PI	82.7
Att-BLSTM	WV (Pennington et al., 2014) (dim=100) + PI	84.0



		BLEU				
Dataset	Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR
Flickr8k	Google NIC(Vinyals et al., 2014)$^{\dagger\Sigma}$	63	41	27	—	—
	Log Bilinear (Kiros et al., 2014a)°	65.6	42.4	27.7	17.7	17.31
	Soft-Attention	**67**	44.8	29.9	19.5	18.93
	Hard-Attention	**67**	**45.7**	**31.4**	**21.3**	**20.30**

MINISTRY OF HOME AFFAIRS

Internship 2020

A.L.I.C.E.

Thank you!