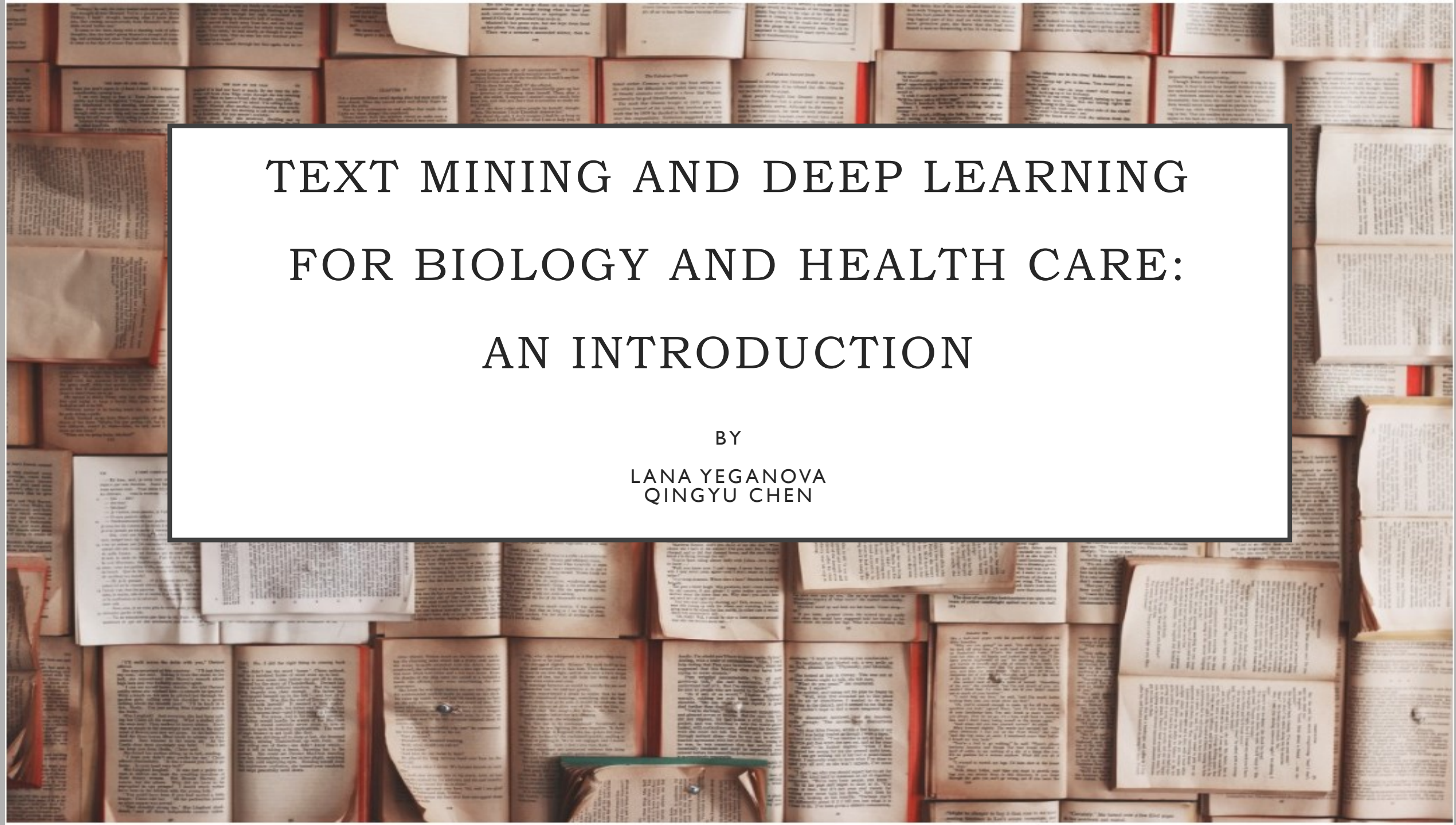
NIH.AI WORKSHOP ON BIOMEDICAL NLP

Hosted by NIH.AI and National Library of Medicine

1:00-1:40pm	Text Mining and Deep Learning for Biology and Healthcare: an Introduction
	Lana Yeganova and Qingyu Chen, NCBI/NLM
1:40-2:15pm	Automatic information extraction from free-text pathology reports using multi-task convolutional neural networks
	Hong-Jun Yoon, Oak Ridge national laboratory
2:15-2:30pm	**Break**
2:30-3:05pm	Biomedical Named Entity Recognition and Relation Extraction
	Robert Leaman and Shankai Yan, NCBI/NLM
3:05-3:40pm	Neural Approaches to Medical Question Understanding
	Asma Ben Abacha and Yassine Mrabet, LHC/NLM
3:40-3:50pm	**Break**
3:50-4:25pm	Transfer Learning in Biomedical NLP: A Case Study with BERT
	Yifan Peng, NCBI/NLM
4:25-5:00pm	Guided Discussion



TEXT MINING AND DEEP LEARNING FOR BIOLOGY AND HEALTH CARE: AN INTRODUCTION

BY

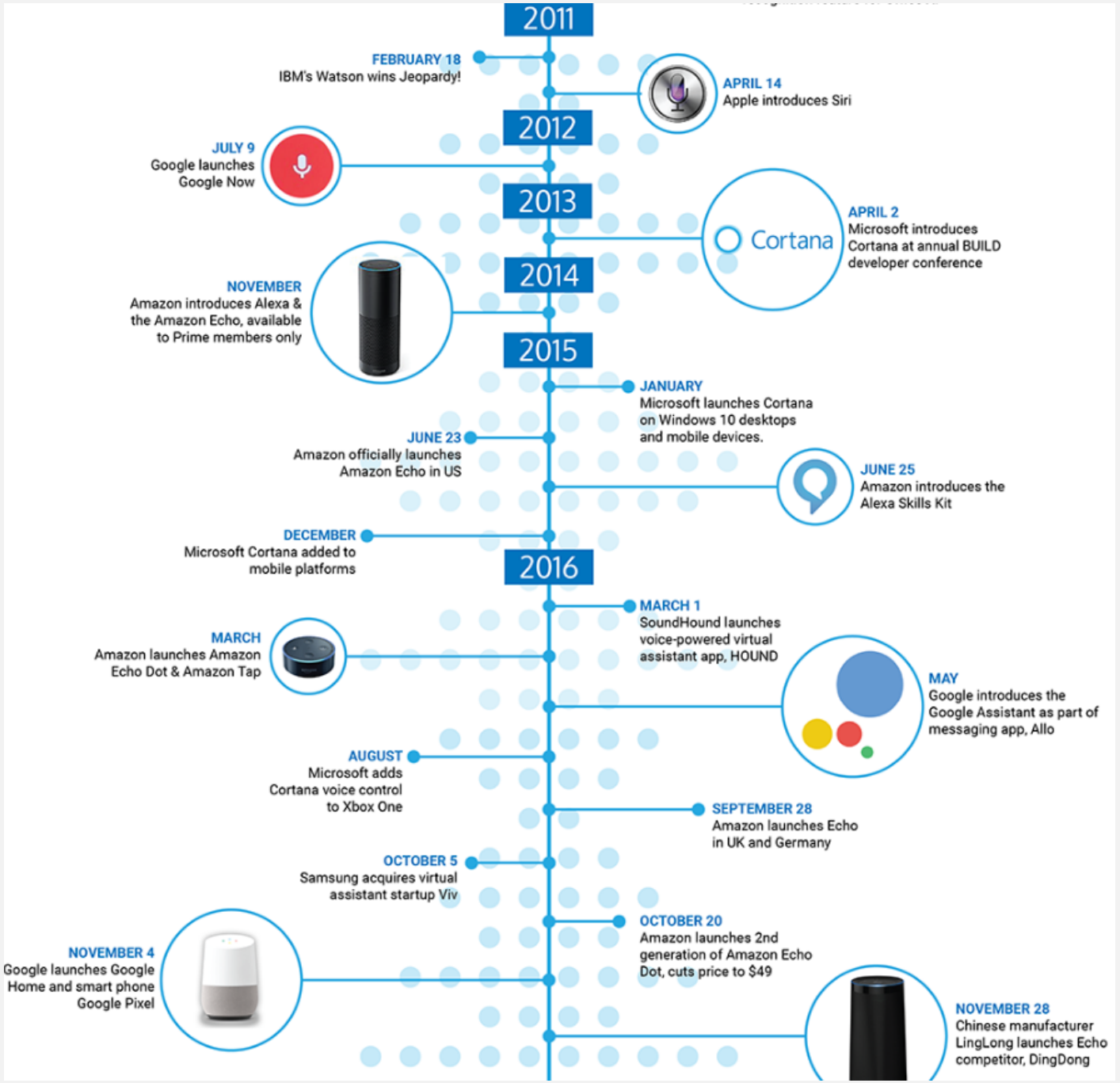
LANA YEGANOVA
QINGYU CHEN

IN THIS TALK

- Why Natural Language Processing
- How to use NLP
- What kind of problems can we solve with NLP
- What improvements do we get by using NLP
- The impact of NLP for biomedical applications
- Deep learning in NLP
- An overview of deep learning applications in bioNLP

WHY NLP?

- NLP has been a driving force fueling modern technologies
  - Information retrieval (google)
  - Machine translation (automatic translation in facebook)
  - Dialogue systems (voice bots such as cortana, alexa)
  - Summarization (news)
  - Sentiment analysis (business decisions)
  - Semantics (question answering systems)
  - Search and suggestion in search


2011
FEBRUARY 18
IBM's Watson wins Jeopardy!
APRIL 14
Apple introduces Siri
2012
JULY 9
Google launches Google Now
2013
APRIL 2
Microsoft introduces Cortana at annual BUILD developer conference
Cortana
2014
NOVEMBER
Amazon introduces Alexa & the Amazon Echo, available to Prime members only
2015
JANUARY
Microsoft launches Cortana on Windows 10 desktops and mobile devices.
JUNE 23
Amazon officially launches Amazon Echo in US
JUNE 25
Amazon introduces the Alexa Skills Kit
DECEMBER
Microsoft Cortana added to mobile platforms
2016
MARCH 1
SoundHound launches voice-powered virtual assistant app, HOUND
MARCH
Amazon launches Amazon Echo Dot & Amazon Tap
MAY
Google introduces the Google Assistant as part of messaging app, Allo
AUGUST
Microsoft adds Cortana voice control to Xbox One
SEPTEMBER 28
Amazon launches Echo in UK and Germany
OCTOBER 5
Samsung acquires virtual assistant startup Viv
OCTOBER 20
Amazon launches 2nd generation of Amazon Echo Dot, cuts price to $49
NOVEMBER 4
Google launches Google Home and smart phone Google Pixel
NOVEMBER 28
Chinese manufacturer LingLong launches Echo competitor, DingDong

BIOMEDICAL DOMAIN

- IBM Watson wins Jeopardy! But IBM Watson Health is not as successful
- Tools from general domain do not generalize to biomedical domain
- Health literature is complicated and heavy in biomedical terminology
- Targeted tools are required for Biomedical literature

WHAT KIND OF TEXT DOCUMENTS ARE AVAILABLE?

- Scientific articles: PubMed, PubMed Central
- Electronic Health Records and Clinical Notes
- FDA product labels
- Grant proposals
- Patient comments on adverse drug reactions
- Social media (subreddits, tweets), newsgroups

Imagine the new cures that could potentially emerge if only one could read, understand and synthesize all the literature?!

Swanson's discoveries

- Fish oil, Raynaud's syndrome, and undiscovered public knowledge. Swanson DR.Perspect Biol Med. 1986
- Migraine and magnesium: eleven neglected connections. Perspect Biol Med. 1988.
- Somatomedin C and arginine: implicit connections between mutually isolated literatures. Perspect Biol Med. 1990

NCBI Resources How To

PubMed.gov
US National Library of Medicine National Institutes of Health

PubMed

Advanced

Format: Abstract

Send to

Bull Med Libr Assoc. 1990 Jan;78(1):29-37.

Medical literature as a potential source of new knowledge.

Swanson DR[1].

Author information

Abstract

Specialized biomedical literatures have been found that are implicitly linked by arguments that they respectively contain, but which nonetheless do not cite or refer to each other. The combined arguments lead to new inferences and conclusions that cannot be drawn from the separate literatures. One such analysis identified one set of articles showing that dietary fish oils lead to certain blood and vascular changes, and a second set containing evidence that similar changes might benefit patients with Raynaud's syndrome. Yet these two literatures had no articles in common and had never before been cited together; neither literature mentioned the other or suggested that dietary fish oil might benefit Raynaud patients. Two years after publication of that analysis, the first clinical trial demonstrating such a beneficial effect was reported independently by others. A second example of literature synthesis, based on eleven indirect connections, led to an inference that magnesium deficiency might be a causal factor i migraine headache. A third example calls attention to implicit connections between arginine intake and blood levels of somatomedins, a potentially fruitful but neglected area of research with implications for the decline with age of thymic function and protein synthesis. A model and an online search strategy to aid in identifying other logically related noninteractive literatures is described. Such structures are probably not rare and may provide the foundation for a literature-based approach to scientific discovery.

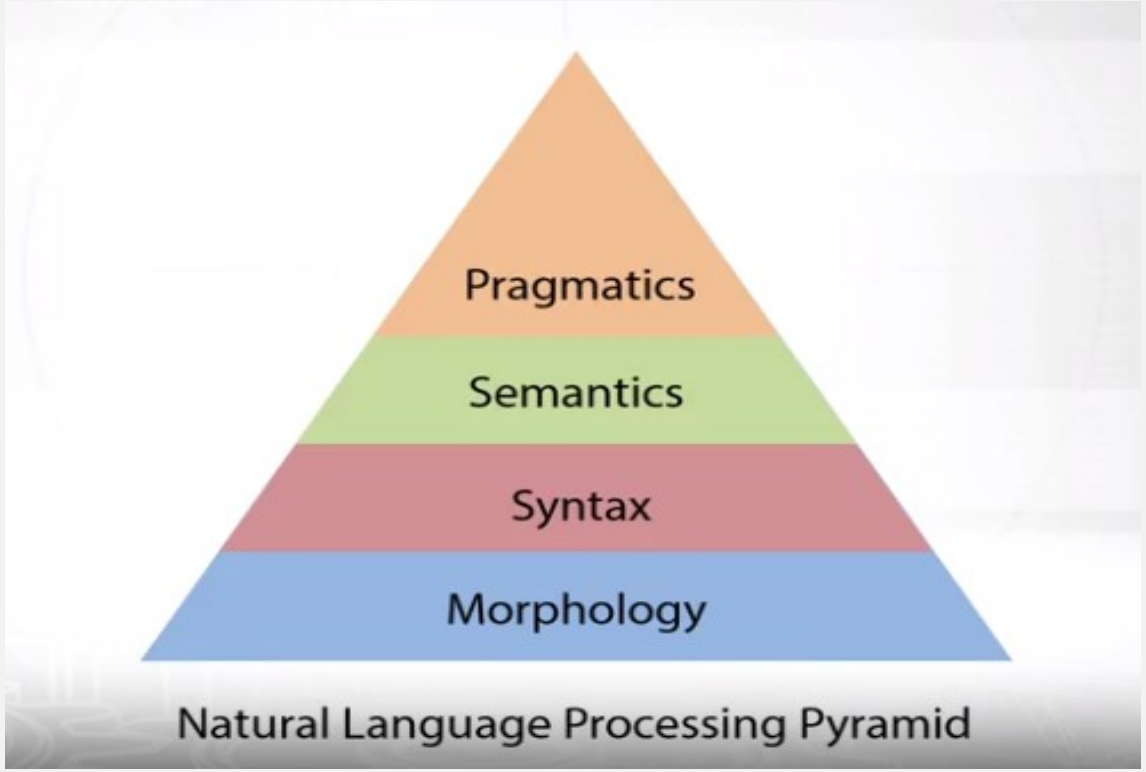
PMID: 2403828 PMCID: PMC225324

[Indexed for MEDLINE] Free PMC Article

APPLICATIONS OF NLP IN BIOMEDICAL DOMAIN

- **Text Mining & Information Extraction**
  - Biomedical Entity Recognition
    - Automatically tag genes, proteins, diseases, chemicals, drugs
  - Discover and extract unstructured knowledge hidden in the text
    - Detect drug interactions, protein interactions
- **Information Retrieval**
  - Develop smart algorithms for efficient literature search
- **Document Classification**
  - Assign text documents to predefined categories according to their content
    - Document triage, Mesh term assignment
- **Literature Clustering and Summarization**
  - Compute clusters in large volumes of literature; compute automatic summaries

THE BUILDING BLOCKS OF LANGUAGE



MORPHOLOGY – EXPLORING STRUCTURE OF WORDS

- Words have structure:
  - runs, ran, and running are inflected forms of the verb run
  - unfriendly is derived from friendly, which is derived from friend

- Words are build from minimal meaningful elements called morphemes:
  - stems and affixes (prefixes and suffixes)

WHY IS MORPHOLOGY IMPORTANT?

- The link between a query and a document may not be established because they use different forms of a term.

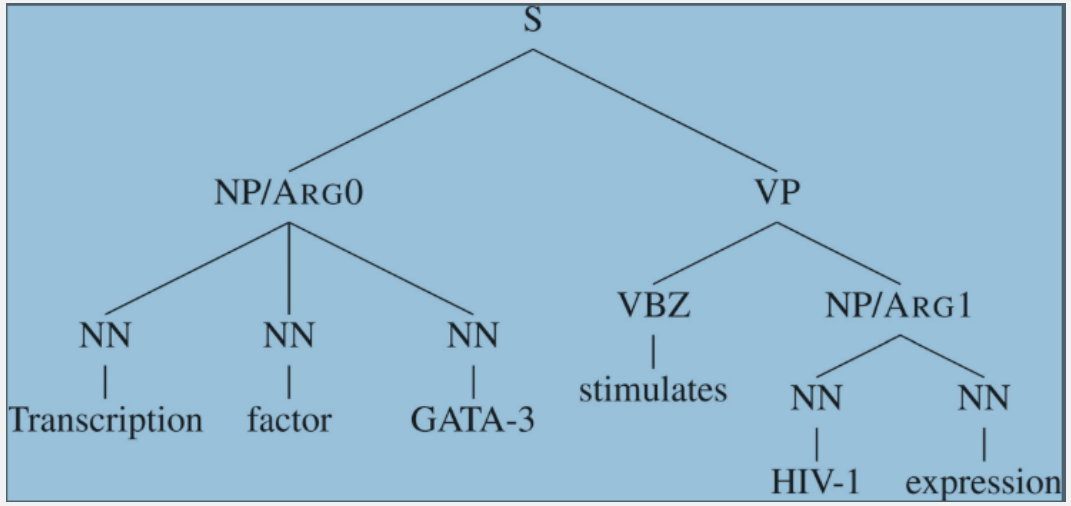
- There may be different types of differences including morphological:
  - A query for autoimmune does not retrieve documents containing autoimmunity

- Term Normalizing can be through stemming or lemmatization
  - Stemming is removing and replacing suffixes to get to the root form of the word
  - Lemmatization changes word to its base form

Yeganova, Kim, et al., PubTermVariants: Biomedical term variants and their use for Pubmed search, Proc. ACL Workshop on Biomedical Natural Language Processing 2016

SYNTAX

- Syntax refers to the grammatical structure if the text
  - Syntactic Analysis or Parsing is the process of analyzing natural language conforming to the rules of a formal grammar.



From: Domain adaptation for semantic role labeling in the biomedical domain; Bioinformatics. 2010;26(8):1098-1104. doi:10.1093/bioinformatics/btq075

SEMANTICS & PRAGMATICS

- Semantics refers to the meaning
  - Semantic Analysis is the process of understanding the meaning and interpretation of words and sentences.

    This enables computers to better understand natural language the way humans do, involving meaning and context.
- Semantics is concerned with literal meaning of a sentence
- Pragmatics is the study of intended meaning. I.e. understanding the meaning in the context.
  - Pronoun resolution
  - Multiword expressions

BASIC CONCEPTS AND NOTATIONS

- A set of all documents is called a **corpus**
  - PubMed contains over ~30 million documents
- Document is split into **tokens** and represented by a set of tokens
  - PubMed contains ~ 4.7 million unique tokens
- A set of all tokens in a corpus is called a **vocabulary**
- Usually exclude stop words (articles and prepositions)

Machines can not process strings or plain text, they require numbers as input to perform any sort of computation

VECTORIZATION OF TEXT

- Documents are represented as **vectors** in high-dimensional space
- **Dimension** of the space is equal to the vocabulary size
- Each dimension corresponds to a **word** in the corpus
- Documents can be represented as a 0-1 sparse vectors – Bag of Words

$$d_1 = (1,1,\ldots.,0,0,0,\ldots 1,0,1)$$
$$d_2 = (1,0,\ldots.,0,0,0,\ldots 1,1,0)$$
$$d_3 = (0,1,\ldots.,0,0,0,\ldots 1,0,0)$$

- Two documents can be compared using the cosine similarity between the vectors

TERM WEIGHTING SCHEMAS

$$tf_{i,j} = n_{ij} / \sum_k n_{k,j}$$

$$idf_i = \log \frac{|D|}{|\{d : t_i \in d\}|}$$

$$w_{i,j} = tf_{i,j} * idf_i$$

weighting scheme	TF weight
binary	$0, 1$
raw count	$f_{t,d}$
term frequency	$f_{t,d} / \sum_{t' \in d} f_{t',d}$
log normalization	$1 + \log(f_{t,d})$

WORD REPRESENTATIONS

- A simplest v*ector* representation of a word may be a one-hot encoded vector, i.e. [0,0,0,….,1,…0]
- A word can also be represented based on its TF-IDF value
- Word embeddings have become the representation of choice in the last decade

ONE-HOT REPRESENTATION

- Why is this representation problematic?
  - It does not give any inherent notion of relation and similarity between the words
  - Very commonly we want to know if the meaning or words are similar
    - If user searched for "Dell notebook" we would like to match documents with "Dell laptop"
    - "Rockville motel" would like to match with "Rockville hotel"

      motel [0 0 0 0 0 0 1 0 0 0]

      hotel [0 0 0 1 0 0 0 0 0 0]

CO-OCCURRENCE BASED REPRESENTATIONS

You shall know a word by the company it keep

— Firth, J.R.

- Idea is to quantify co-occurrence of terms in the corpus
- Co-occurrence matrix with a fixed context window
  - Compute co-occurrence matrix based on co-occurrence of words within a small window
  - Perform PCA on the co-occurrence matrix
  - Requires huge memory to store the co-occurrence matrix

PREDICTION BASED EMBEDDINGS

***Don't count, predict!* A systematic comparison of context-counting vs. context-predicting semantic vectors**

Marco Baroni and **Georgiana Dinu** and **Germán Kruszewski**

GloVe: Global Vectors for Word Representation

Jeffrey Pennington, Richard Socher, Christopher D. Manning
Computer Science Department, Stanford University, Stanford, CA 94305
jpennin@stanford.edu, richard@socher.org, manning@stanford.ed

Enriching Word Vectors with Subword Information

Piotr Bojanowski* and **Edouard Grave*** and **Armand Joulin** and **Tomas Mikolov**
Facebook AI Research
{bojanowski,egrave,ajoulin,tmikolov}@fb.com

Distributed Representations of Words and Phrases and their Compositionality

Tomas Mikolov
Google Inc.
Mountain View
mikolov@google.com

Ilya Sutskever
Google Inc.
Mountain View
ilyasu@google.com

Kai Chen
Google Inc.
Mountain View
kai@google.com

Greg Corrado
Google Inc.
Mountain View
gcorrado@google.com

Jeffrey Dean
Google Inc.
Mountain View
jeff@google.com

Deep contextualized word representations

Matthew E. Peters†, Mark Neumann†, Mohit Iyyer†, Matt Gardner†,
{matthewp,markn,mohiti,mattg}@allenai.org

Christopher Clark*, Kenton Lee*, Luke Zettlemoyer†*
{csquared,kentonl,lsz}@cs.washington.edu

WORD EMBEDDINGS IN DEEP LEARNING

- Traditional one-hot vectors or bag-of-word models rely on exact word matching.
- It fails to capture distinct words with similar meanings, e.g., cancer vs tumor.
- For example, given the following pair of sentences, it will give a similarity of 0 (assume stop words are removed):
  - Obama speaks to the media in Illinois
  - The president greets the press in Chicago

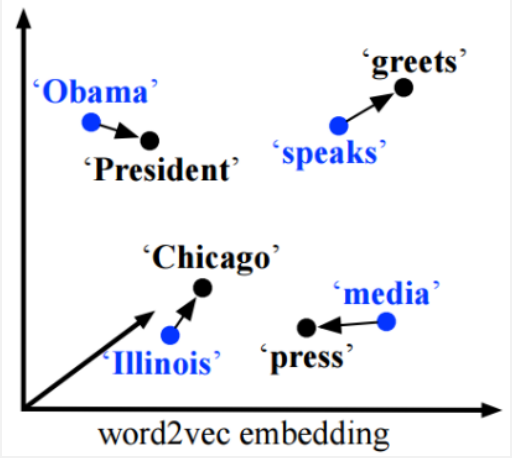
Example from http://proceedings.mlr.press/v37/kusnerb15.pdf

WORD EMBEDDINGS IN DEEP LEARNING

- With word embeddings every word used in a language can be represented by a set of real numbers (a vector).
- Word embeddings are N-dimensional vectors that try to capture word-meaning and context in their values.
- Characteristics:
  - Every word has a unique word embedding vector, which is just a list of numbers for each word.
  - The word embeddings are multidimensional; typically embeddings are between 50 and 500 in length.
  - For each word, the embedding captures the "meaning" of the word.
  - Similar words end up with similar embedding values.

WORD EMBEDDINGS IN DEEP LEARNING

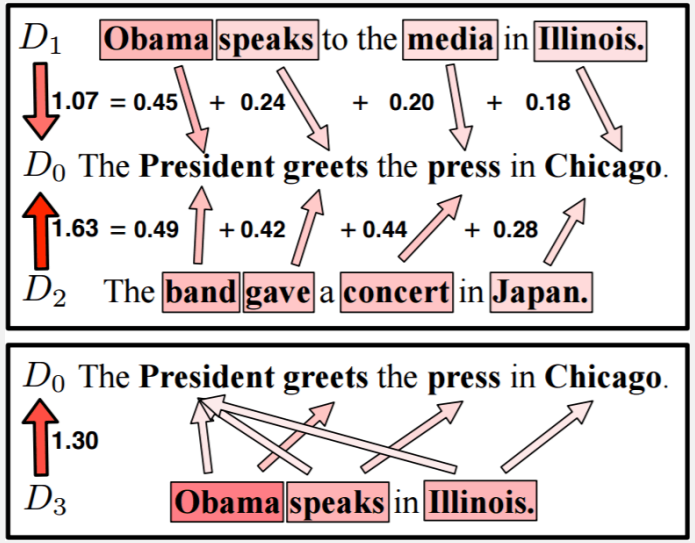
- Embeddings learn the semantic space based on the context of a word, which does not reply on exact word matching.
- Words with similar context are close to each other in embeddings.



Example from http://proceedings.mlr.press/v37/kusnerb15.pdf

WORD EMBEDDINGS IN DEEP LEARNING

- We can use the similarity between word vectors to find similar or related words, which is beyond keyword matching.



Example from http://proceedings.mlr.press/v37/kusnerb15.pdf

HOW TO TRAIN WORD EMBEDDINGS?

- Word2Vec
  - Continuous bag of words - predicts the probability of a word given a context.
  - Skip-gram model - predict the context given a word
- Glove
  - Instead of implicitly modelling word co-occurrences like word2vec, it focuses on aggregated global co-occurrence statistics from a training corpus.
- FastText
  - An extension of word2vec approach using character n-grams to represent a word.
  - If a word does not exist in the vocabulary, it can still produce an estimated vector based on its character n-grams.

AVAILABLE BIOMEDICAL EMBEDDINGS

- Our group has made a few biomedical embeddings publicly available:
- BioWordVec: https://github.com/ncbi-nlp/BioWordVec
  - fastText embeddings trained using PubMed and MeSH terms
- BioSentVec: https://github.com/ncbi-nlp/BioSentVec
  - Sentence embeddings trained using PubMed and MIMIC-III
- BioConceptVec: https://github.com/ncbi-nlp/BioConceptVec
  - Four versions of concept embeddings which aim to find similar biological concepts, e.g., find related genes.

EMBEDDINGS IN BIOMEDICAL APPLICATIONS

Sentence-level search

- Compare new findings with previous knowledge
- Perform evidence attribution
- Assist biomedical question answering

LitSense

About

Given a query, LitSense finds the best-matching **sentences** from **over half a billion statements** from PubMed and PubMed Central.

https://www.ncbi.nlm.nih.gov/research/litsense/

COMBINING TRADITIONAL IR WITH EMBEDDINGS

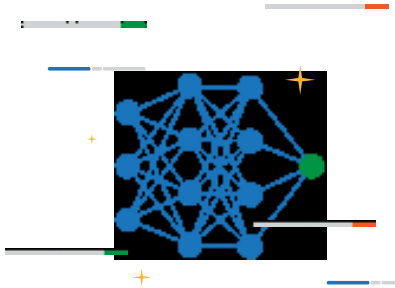
Traditional IR (TF-IDF, BM25,...)

Fast

Sentence = Bag Of Words

Semi-exact matching
(stemming, manual synonyms)

Neural Embeddings (sent2vec)



Slow

Sequence information

'Discovered' synonyms

Combine and re-rank

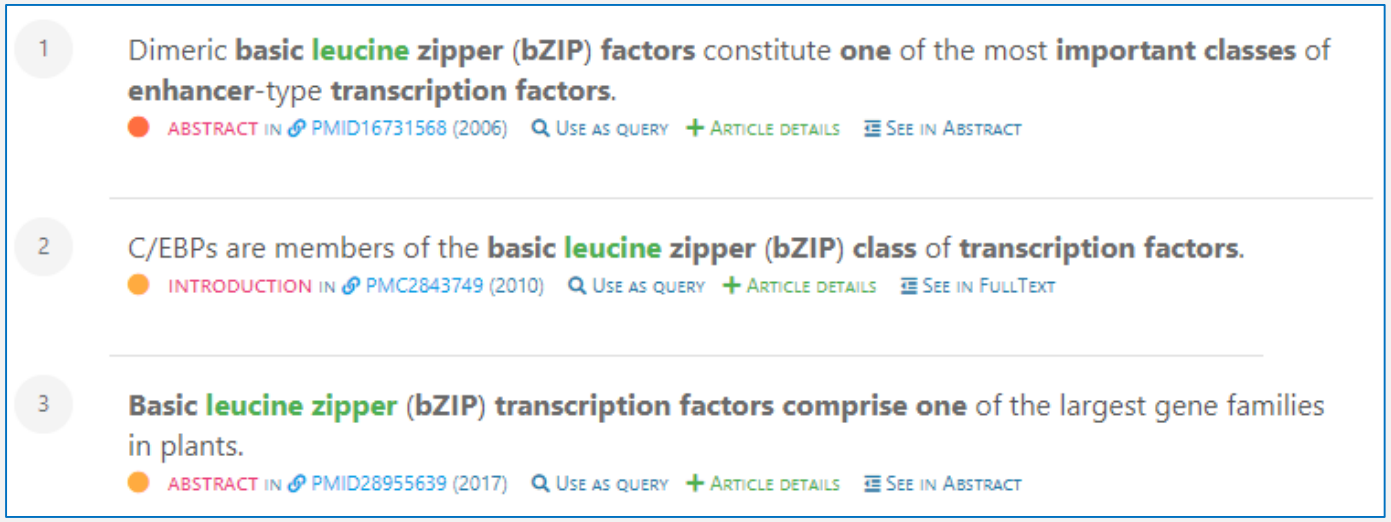
LitSense

EXAMPLE: EVIDENCE ATTRIBUTION

Curators of databases (ex: CDD, UniProt) need to link **manually annotate proteins** to **evidence in literature**.

How to find evidence for the summary of the bZIP superfamily :
"Basic leucine zipper (bZIP) factors comprise one of the most important classes of enhancer-type transcription factors" ?

Search it on LitSense !



RECENT DEVELOPMENT ON EMBEDDINGS

- Contextualized language embeddings
  - Same words may have different meanings in different context.
  - However, current word embeddings treat them equally.
  - Recent development focus on contextualized language embeddings:
    - ELMO: learns multiple context representations of a word.
    - BERT: we will detail BERT in the later session.

BEYOND EMBEDDINGS

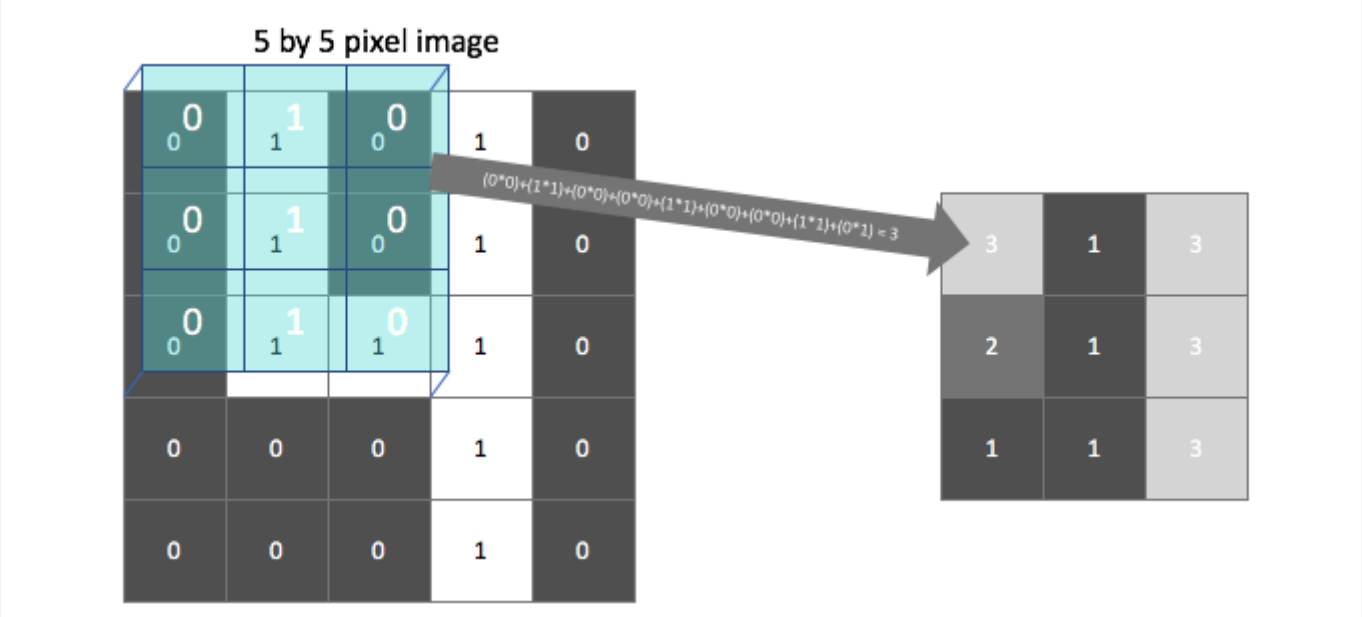
- Embeddings are often used as an intermediate component in downstream NLP tasks.
- Deep learning models take embeddings as an input.
- Major deep learning models include:
  - CNN
  - RNN (GRU, LSTM, and with attention)
  - Transformer (stacked encoder-decoder)

CONVOLUTIONAL NEURAL NETWORK(CNN)

- A model was designed originally for computer vision.
- Image classification was a bottleneck task in computer vision, where traditional machine learning models achieve low accuracy.
- In 2012, a CNN model achieved a top-5 error of ~15% in a large-scale image classification competition, significantly outperforming other models.
- Essentially, a CNN model has two primary operations: convolution and pooling.

CONVOLUTIONAL NEURAL NETWORK(CNN)

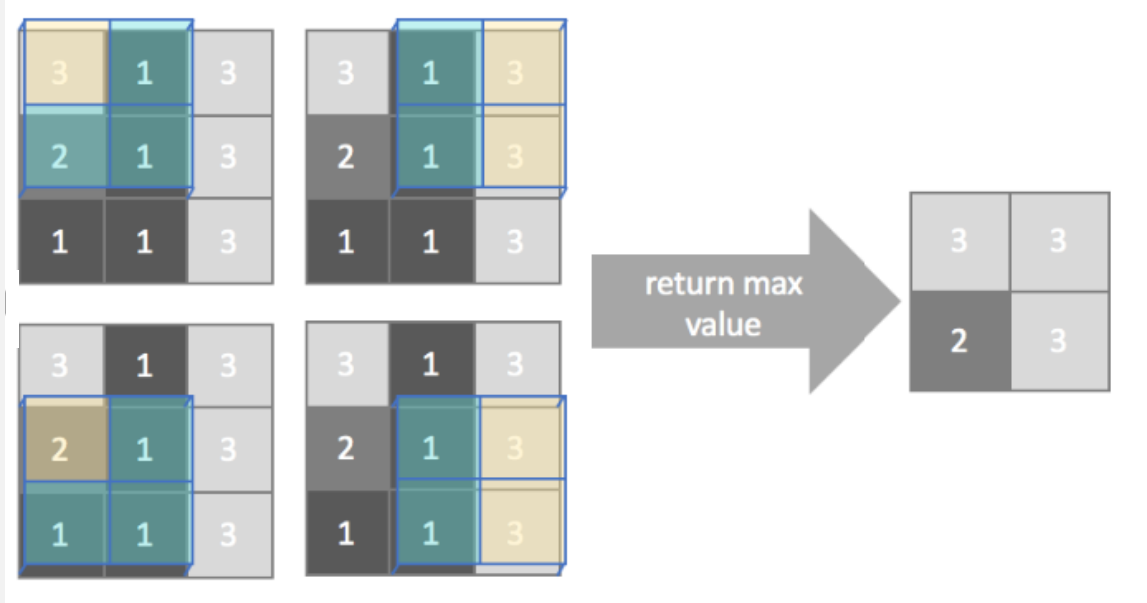
- Convolution operations aim to extract important characteristics or features from an image.
- The following example uses a vertical edge filter for convolution operation.



Example from https://www.jeremyjordan.me/convolutional-neural-networks/

CONVOLUTIONAL NEURAL NETWORK(CNN)

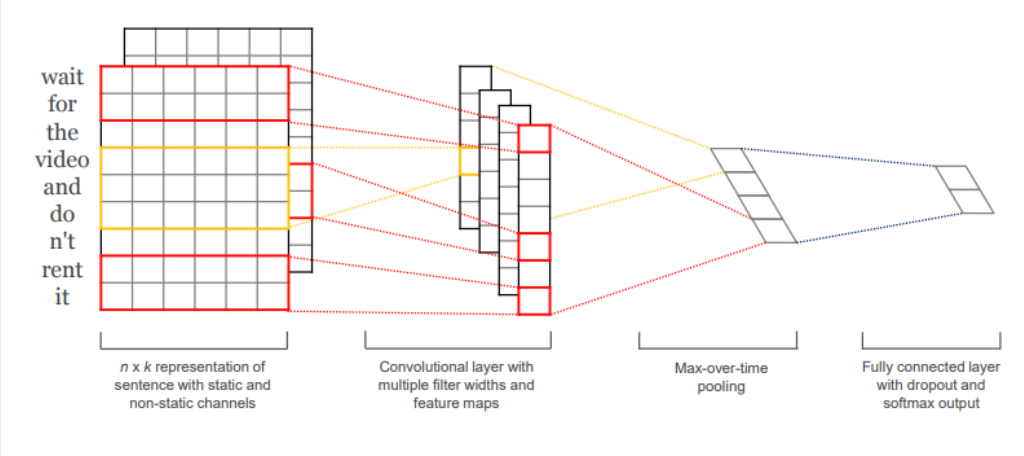
- Pooling operations are often performed after convolution operations.
- It significantly reduces the spatial dimension while keeps the most import features



Example from https://www.jeremyjordan.me/convolutional-neural-networks/

CNN IN NLP APPLICATIONS

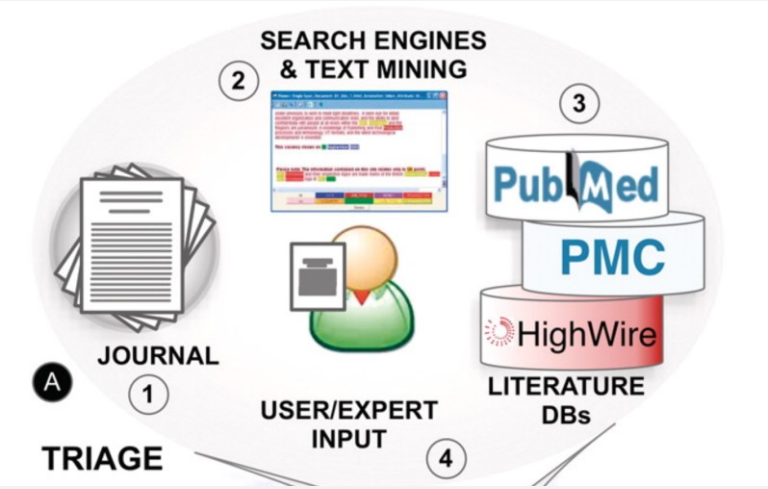
- CNN models have been used widely in NLP applications, including:
  - Sentence classification
  - Document triage
  - Sentiment analysis



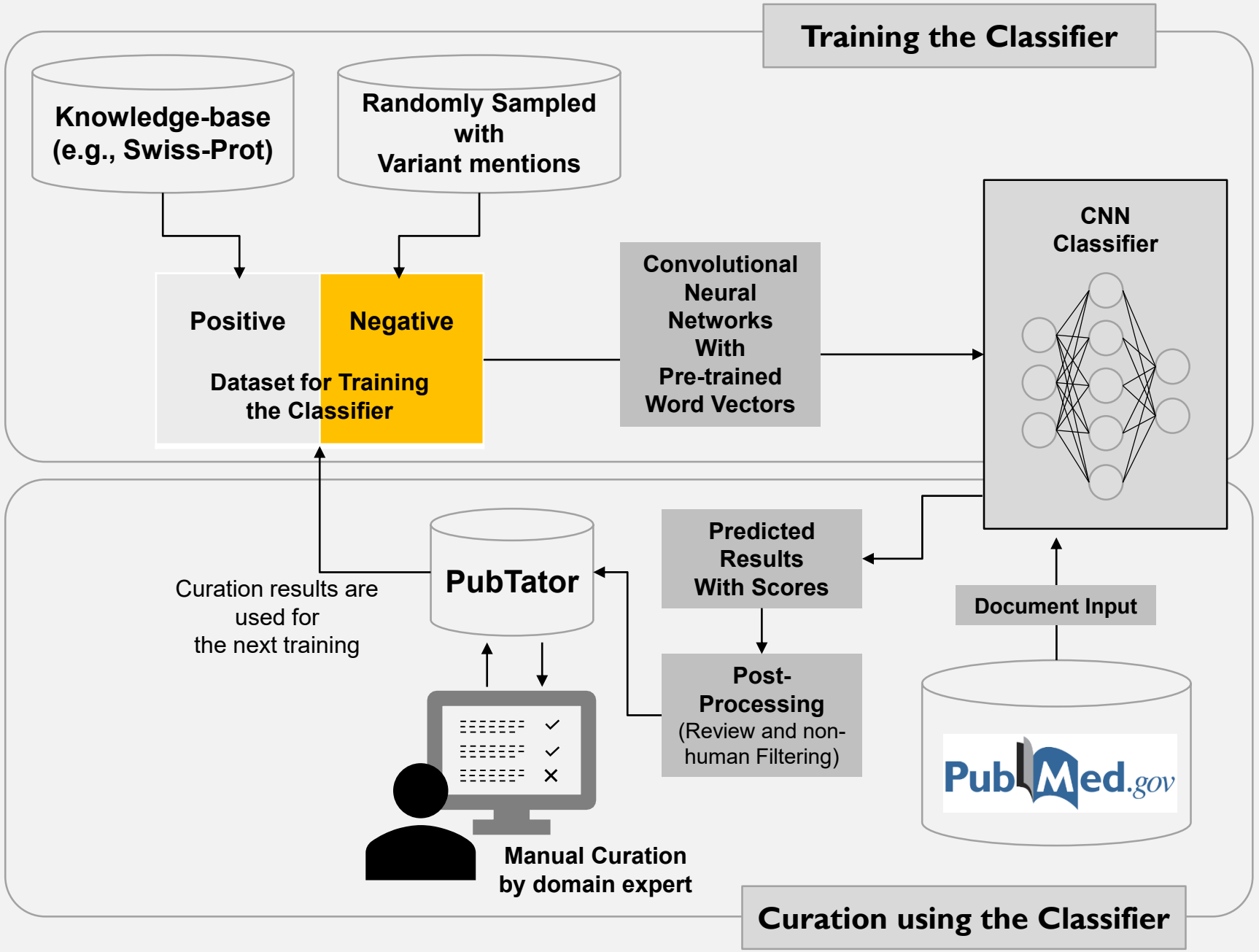
Example from https://arxiv.org/pdf/1408.5882v2.pdf

CNN IN BIOMEDICAL NLP APPLICATIONS

- Biomedical document triage is an essential step in biocuration pipeline.
- The relevant documents are automatically collected instead of tedious manual check.
- Our group developed a CNN model for biomedical document classification.
- It has been used in UniProtKB/Swiss-Prot and GWAS catalog curation pipeline for document triage.



Example from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3328793/

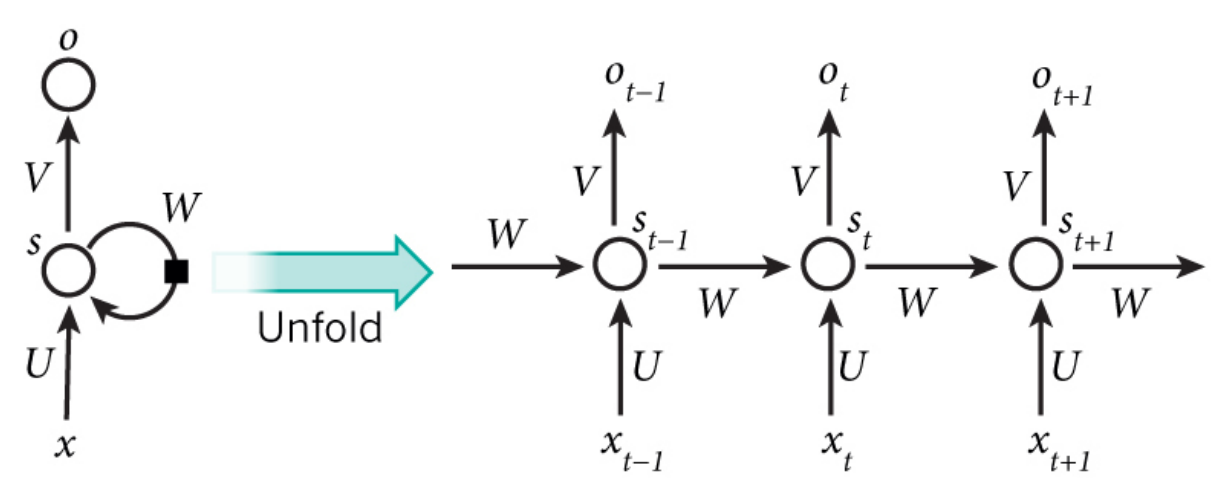

Training the Classifier
Knowledge-base (e.g., Swiss-Prot)
Randomly Sampled with Variant mentions
Positive
Negative
Dataset for Training the Classifier
Convolutional Neural Networks With Pre-trained Word Vectors
CNN Classifier
Predicted Results With Scores
Post-Processing (Review and non-human Filtering)
Document Input
PubTator
Curation results are used for the next training
Manual Curation by domain expert
PubMed.gov
Curation using the Classifier

RECURRENT NEURAL NETWORK(RNN)

- CNN models are effective, but only focuses on the current input, without considering previous inputs.
- However, sequential information is critical in many NLP tasks.
- RNN is designed to model sequential information, which previous inputs are also considered.

RECURRENT NEURAL NETWORK(RNN)

- For an input sequence, an RNN takes both current input and previous input, and processes in its hidden states. This process is repeated for all the inputs in the sequence.



Example from https://www.nature.com/articles/nature14539

RECURRENT NEURAL NETWORK(RNN)

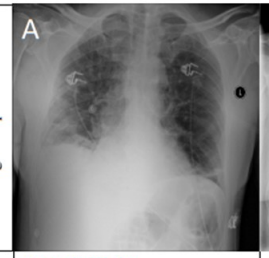
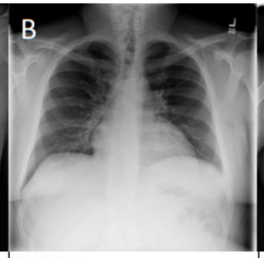
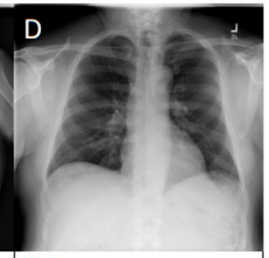
- There are a few variations of RNN:
  - Long short term memory (LSTM)
    - RNN has a limitation of losing important information for long sequential input
    - LSTM uses different gates aiming to preserve important information and ‘forget’ unnecessary information
  - Gated Recurrent Units (GRU)
    - A simplified version of LSTM which only has two gates (update and reset)
  - LSTM vs GRU
    - There is no universal agreement on which is better.
    - GRU is more computationally efficient.

RECURRENT NEURAL NETWORK(RNN)

- There are a few enhancements on RNN-like models.
  - Bidirectional RNN:
    - Keep track of sequential information from two directions.
  - Attention schema:
    - Based on the architecture of RNN models, the final unit needs to capture all the information of the input sequence.
    - Attention schema allows RNN models to focus on a particular information at each step so that important information can be captured.

RNN IN BIOMEDICAL APPLICATIONS

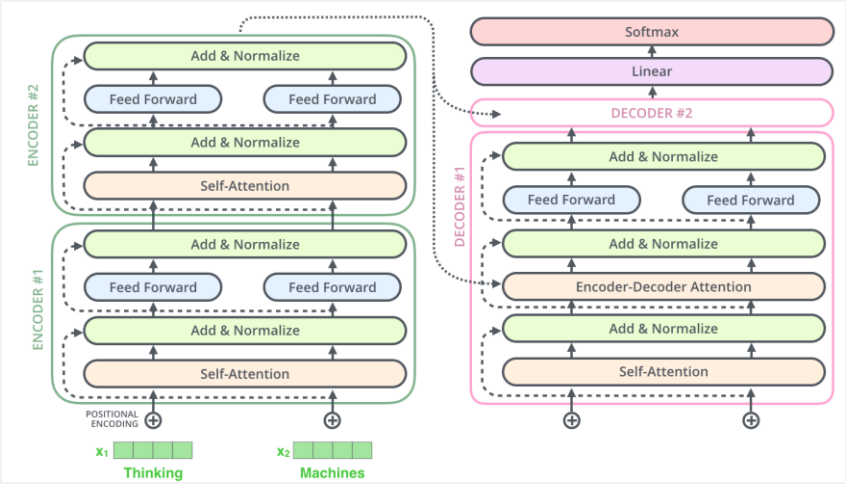
- Our group developed an LSTM model as a part of clinical report generation framework.

Image Sample cases				
P	Atelectasis Effusion	No finding	Nodule Pneumothorax Mass Consolidation	Mass
Original report	findings : a single ap view of the chest demonstrates increasing bibasilar interstitial opacities with decreased overall aeration . increasing blunting of right costophrenic angle. ... impression : increasing bibasilar atelectasis with possible development of right pleural effusion .	Normal no evidence of lung infiltrate .	findings : heart and mediastinum unchanged . multiple lung nodules . evidence of recent left chest surgery with left chest tube in place . very small left apical pneumothorax . lungs unchanged , no evidence of acute infiltrates . impression : stable chest .	findings : large left suprahilar and infrahilar masses as well as the well circumscribed nodule the level of the aortic knob . the right infrahilar mass as well . no effusion . impression : metastatic lung disease .
Generated Report	findings : a single ap view of the chest demonstrates unchanged bilateral reticular opacities , consider atelectasis . continued left basilar atelectasis , no evidence of developing infiltrate . the cardiac and mediastinal contours are stable . impression : no evidence of developing infiltrate .	findings : pa and lateral views of the chest demonstrate lungs that are clear without focal mass , infiltrate or effusion . cardiomediastinal silhouette is normal size and contour . pulmonary vascularity is normal in caliber and distribution . impression : no evidence of acute pulmonary pathology	findings : pa and lateral views of the chest demonstrate unchanged bilateral chest tubes . again pulmonary nodules are seen on the right and cardiac silhouette unchanged . the cardiac and mediastinal contours are stable . impression : 1. bilateral masses and left lower lung field consolidation . 2.new bilateral lung masses .	comparison is to previous upright study of no significant interval change is seen in the appearance of the chest . the mediastinal soft tissue and pulmonary vascularity are stable . there are blastic bone lesions in the chest . bones , soft tissues are normal . the lung fields are clear . there are calcified lymph nodes in the left lower lung . impression : . sclerotic lesions in the left humeral , consistent with metastasis.

Example from https://arxiv.org/abs/1801.04334

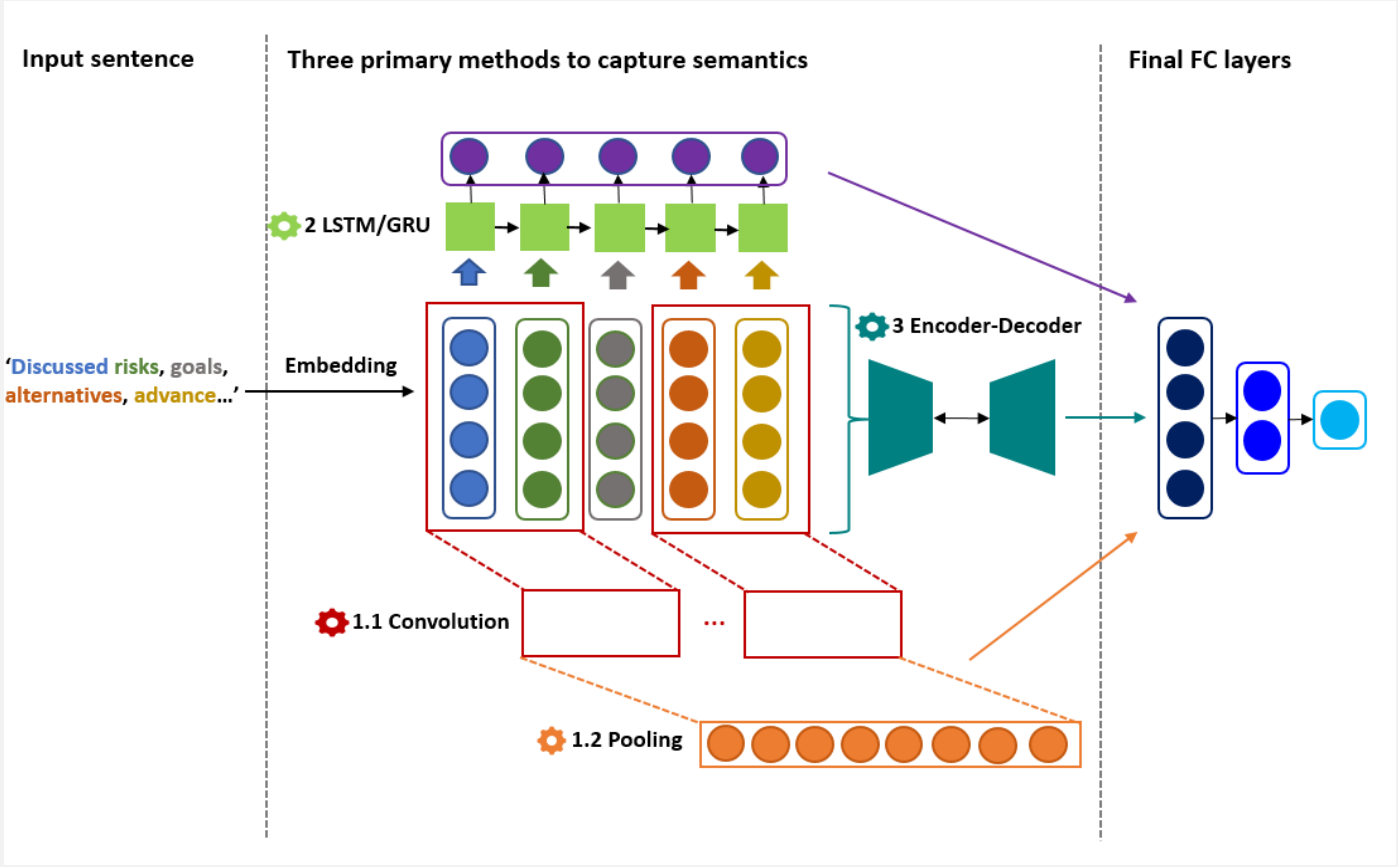
RECENT DEVELOPMENT ON DEEP LEARNING MODELS IN NLP

- Transformer
  - Stacked encoder-decoder architecture.
  - Skip-connection and multi-head attention.
- BERT
  - Bidirectional encoder representations based on transformers.
  - We will detail in the later session.



Example from http://jalammar.github.io/illustrated-transformer/

A SUMMARY OF DEEP LEARNING MODELS IN NLP



ACKNOWLEDGEMENTS

- **Dr. Zhiyong Lu's Information Retrieval Group**

Won Kim	Natalie Xie	Wei Chih-Hsuan
Rezarta Islamaj	Wanli Liu	Yan Shankai
Donald Comeau	Alexis Allot	Qingyu Chen
Sun Kim	Lee Kubim	Lana Yeganova
Yifan Peng	Robert Leaman	

- Workshop organizers: **NIH.AI** and **NLM**

- This research was supported by the NIH Intramural Research Program of the National Library of Medicine.

Thank You!