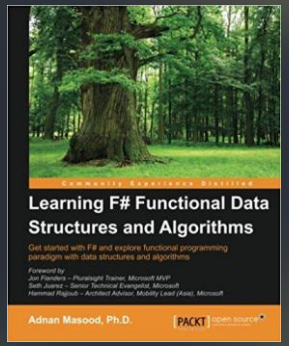
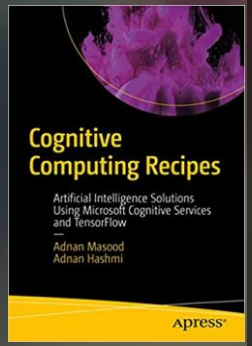
Drive Intelligence from Text Comprehension in Enterprise Apps

Adnan Masood, PhD.

Microsoft Azure
+AI Conference
CO-PRODUCED BY
Microsoft & DEVintersection

About the Speaker





Adnan Masood, Ph.D. is an Artificial Intelligence and Machine Learning researcher, software architect, and Microsoft MVP (Most Valuable Professional) for Artificial Intelligence. As Chief Architect of AI and Machine Learning, at Global, he collaborates with Stanford Artificial Intelligence Lab, and MIT AI Lab for building enterprise solutions

Author of Amazon bestseller in programming languages, **"Functional Programming with F#",** Dr. Masood teaches Data Science at Park University, and has taught Windows Communication Foundation (WCF) courses at the University of California, San Diego. He is a regular speaker to various academic and technology conferences (WICT, DevIntersection, IEEE-HST, IASA, and DevConnections), local code camps, and user groups. He also volunteers as STEM (Science Technology, Engineering and Math) robotics coach for elementary and middle school students

A strong believer in giving back to the community, Dr. Masood is a co-founder and president of the Pasadena .NET Developers group, co-organizer of Tampa Bay Data Science Group, and Irvine Programmer meetup. His recent talk at Women in Technology Conference (WICT) Denver highlighted the importance of diversity in STEM and technology areas, and was featured by variety of news outlets.

Drive Intelligence from Text Comprehension in Enterprise Apps

Natural language understanding and processing has revolutionized the way we deal with unstructured data within an enterprise. This talk provides a comprehensive overview of variety of text mining techniques including Q&A (Dr. Q&A, BERT, Stanford BiDAF), classification, summarization, topic modeling, annotation, and contract analysis. Along with Microsoft Cognitive Services, the session covers how to use machine learning libraries to drive insights and intelligence in your applications and covers how to work with unstructured text and turn unstructured text into meaningful insights into mobile, web and line of business applications.

In this code-focused session, we will demonstrate how to use a few lines of code to easily analyze sentiment, extract key phrases, detect topics, and detect language for any kind of text. The techniques include TF-IDF, LDA, Word2Vec, Doc2Vec, word embedding, BERT, ELMO, and BiDAF etc. to showcase their use in enterprise applications with real world use cases.

On Feb. 14, OpenAI announced yet another feat of machine learning ingenuity in a blog post detailing how its researchers had trained a language model using text from 8 million webpages to predict the next word in a piece of writing. The resulting algorithm, according to the nonprofit, was stunning: It could "[adapt] to the style and content of the conditioning text" and allow users to "generate realistic and coherent continuations about a topic of their choosing." To demonstrate the feat, OpenAI provided samples of text that GPT-2 had produced given a particular human-written prompt.

For example, researchers fed the generator the following scenario:

> In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.

The GPT-2 algorithm produced a news article in response:

> The scientist named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown to science. Now, after almost two centuries, the mystery of what sparked this odd phenomenon is finally solved. Dr. Jorge Pérez, an evolutionary biologist from the University of La Paz, and several companions, were exploring the Andes Mountains when they found a small valley, with no other animals or humans. Pérez noticed that the valley had what appeared to be a natural fountain, surrounded by two peaks of rock and silver snow. Pérez and the others then ventured further into the valley. "By the time we reached the top of one peak, the water looked blue, with some crystals on top," said Pérez.

What is Dark Data

The multi-modal the mass of text, tables, and images that are widely collected and stored but which cannot be exploited by standard relational tools.

Natural Language Text

… The Namurian Tsingyuan Formation from Ningxia, China, is divided into three members …

Formation-Time (Location)

Formation	Time
Tsingyuan Fm.	Namurian

Formation	Location
Tsingyuan Fm.	Ningxia

Table

TABLE 2—Ranged abundance of gastropod genera from Tsingyuan Formation.

Genus	No. of specimens
Euphemites	6
Retispira	128
Sinutina	5

Taxon-Formation

Taxon	Formation
Retispira	Tsingyuan Fm.

Document Layout

Genus STROBEUS Meek and Worthen, 1866

STROBEUS RECTILINEA (Phillips, 1836)

Figure 5.16, 5.17

Buccinum rectineum PHILLIPS, 1836

Macrochilina tumida DE KONINCK, 1881

Macrochilina obesa DE KONINCK, 1881

Macrochilina intermadia DE KONINCK, 1881

Taxon-Taxon

Taxon	Taxon
Strobeus Rectilinea	Buccinum Rectineum

Image

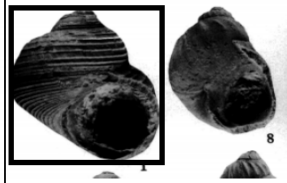


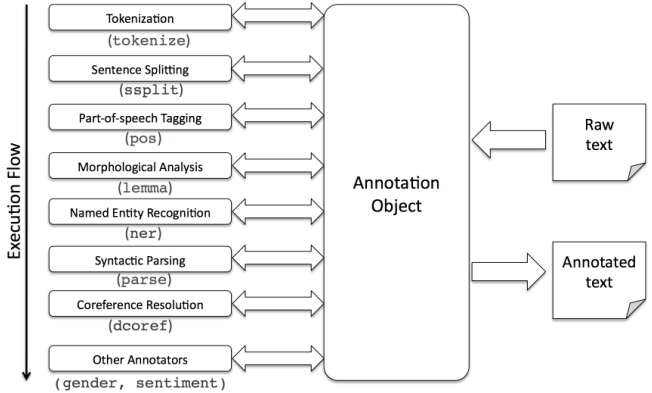
FIGURE 5—1–7, ?*Shansiella tongxinensis* Guo;

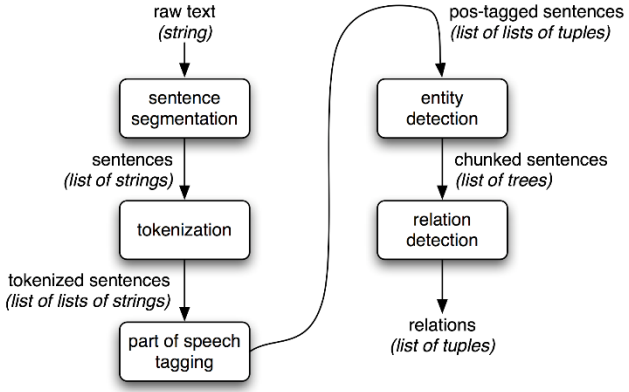
Taxon-Real Size

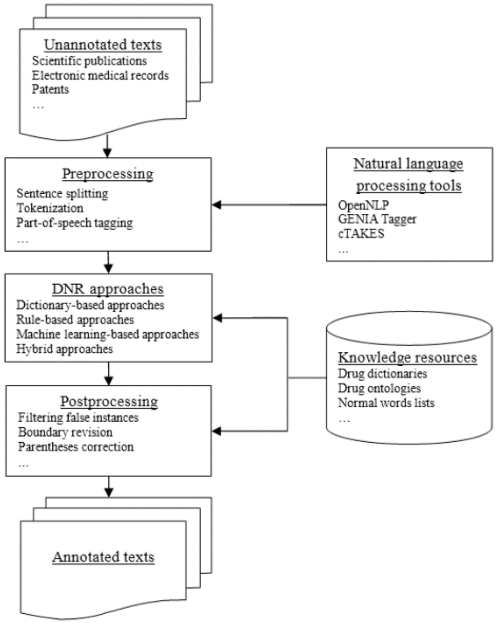
Taxon	Real Size
Shansiella tongxinensis	5cm x 5cm

contentSkip to site indexPoliticsSubscribeLog InSubscribeLog InToday's PaperAdvertisementSupported ORG byF.B.I. Agent Peter Strzok PERSON , Who Criticized Trump PERSON in Texts, Is FiredImagePeter Strzok, a top F.B.I. GPE counterintelligence agent who was taken off the special counsel investigation after his disparaging texts about President Trump PERSON were uncovered, was fired. CreditT.J. Kirkpatrick PERSON for The New York TimesBy Adam Goldman ORG and Michael S. SchmidtAug PERSON . 13 CARDINAL , 2018WASHINGTON CARDINAL — Peter Strzok PERSON , the F.B.I. GPE senior counterintelligence agent who disparaged President Trump PERSON in inflammatory text messages and helped oversee the Hillary Clinton PERSON email and Russia GPE investigations, has been fired for violating bureau policies, Mr. Strzok PERSON 's lawyer said Monday DATE .Mr. Trump and his allies seized on the texts — exchanged during the 2016 DATE campaign with a former F.B.I. GPE lawyer, Lisa Page — in PERSON assailing the Russia GPE investigation as an illegitimate "witch hunt." Mr. Strzok PERSON , who rose over 20 years DATE at the F.B.I. GPE to become one of its most experienced counterintelligence agents, was a key figure in the early months DATE of the inquiry.Along with writing the texts, Mr. Strzok PERSON was accused of sending a highly sensitive search warrant to his personal email account.The F.B.I. GPE had been under immense political pressure by Mr. Trump PERSON to dismiss Mr. Strzok PERSON , who was removed last summer DATE from the staff of the special counsel, Robert S. Mueller III PERSON . The president has repeatedly denounced Mr. Strzok PERSON in posts on

Stanford CoreNLP Solution Overview







KEEP UP **ON** *Vietnam* **YOUR** **READING** *UK* **WITH AUDIO** **BOOKS** *Louisiana, USA*

Audio **books** *Louisiana, USA* are highly **popular** *S.Carolina, USA* with **library** *Pennsylvania, USA* patrons in the **town** *Mass., USA*

of *Turkey* **Springfield,** *Virginia, USA* **Greene** *Maine, USA* County, **MO.** *Norway* "People are **mobile** *Alabama, USA*

and busier, and audio **books** *Louisiana, USA* fit into that lifestyle" says **Gary** *Indiana, USA*

Sanchez, *Dominican Republic* who oversees the **library's** *Pennsylvania, USA* $2 **million** *Kentucky, USA* budget...

LexNLP: Natural language processing and information extraction for legal and regulatory texts

Michael J Bommarito II, Daniel Martin Katz, Eric M Detterman

LexPredict, LLC

```
>>> from lexnlp.extract.en.definitions import
 ↪ get_definitions
>>> list(get_definitions(text2))
['First Deposit', 'Deposit', 'Second Deposit']
```

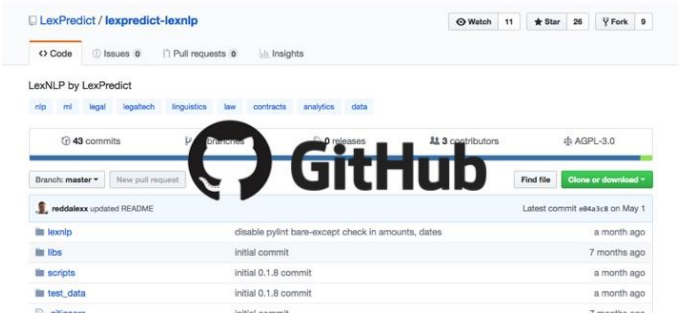
Finally, in addition to sentence boundary pitfalls, Example 3 also demonstrates the usage of constraints and regulatory references.

```
>>> from lexnlp.extract.en.constraints import
 ↪ get_constraints
>>> constraints = list(get_constraints(text3))
```

```
deed and closing the
byer, on behalf of itself and
 and assigns, shall thereby
f the seller parties from,
nd all liabilities against
ller parties for,
o, or in connection with the
her arising or accruing',
```

LEXNLP LEXPREDICT®

python™



GitHub

#OpenSource #OpenSourceLegal

Vertical	Typical use case
Finance	Search; Compliance; Entity matching; Call center analytics; Risk management; Anti-money laundering;
Insurance	Sentiment of customer interactions; Problem topics identification; Claim adjuster notes extraction
Media	Social media analytics; Audio/video broadcast content analysis
Retail	Brand/product analytics based on customer feedback
Process control	Problem/cause augmentation from operator notes
Energy	Price and demand forecasting
Oil and gas	Operator comments analysis; Drilling efficiency
Legal	Search; Relationship extraction (e.g., Fred sued Carl); Document clustering (clustering documents with similar complaints for class action suit discovery)
Health care	Medical record content extraction; Drug interaction discovery from PubMed articles; Disease outbreak monitoring and control from social media data
Security	Log analysis, NLP techniques translate to security models
Government	Disaster scoping and damage assessment from social media data

NLP Platform Features

Vectors
representing
Phrases and Sentences
that do not ignore word order
and capture semantics for NLP tasks

Single Word Vectors

- Distributional Techniques
- Brown Clusters
- Useful as features inside models, e.g. CRFs for NER, etc.
- Cannot capture longer phrases

Documents Vectors

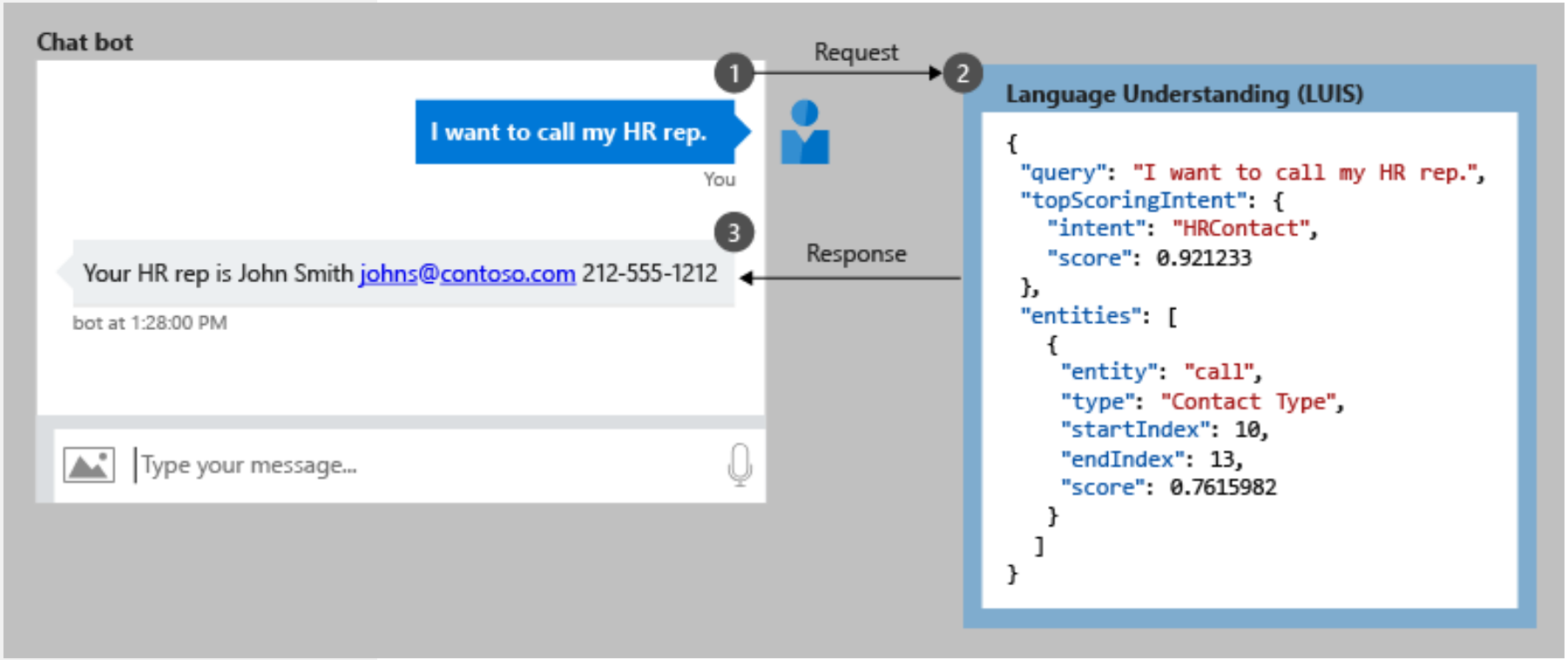
- Bag of words models
- LSA, LDA
- Great for IR, document exploration, etc.
- Ignore word order, no detailed understanding

- **Language Modeling** (Speech Recognition, Machine Translation)
- **Word-Sense Learning** and Disambiguation
- **Reasoning over Knowledge Bases**
- Acoustic Modeling
- Part-Of-Speech Tagging
- Chunking
- Named Entity Recognition
- Semantic Role Labeling
- Parsing
- Sentiment Analysis
- Paraphrasing
- Question-Answering

Python packages for NLP

- NLP Focus Packages
  - **NLTK**
  - **Spacy**
  - **Gensim**
  - Textblob
  - **Scikit Learn**
  - Stanford NLP (java)
  - WordNet, SentiWordNet
  - **FastText / MUSE / Faiss**
- Deep Learning Frameworks
  - **Tensorflow / Keras**
  - Pytorch

- Other Noteworth
  - **Scrapy**
  - Newspaper
  - nlp-architect


Chat bot
1
Request
2
Language Understanding (LUIS)
I want to call my HR rep.
You
{
"query": "I want to call my HR rep.",
"topScoringIntent": {
"intent": "HRContact",
"score": 0.921233
},
"entities": [
{
"entity": "call",
"type": "Contact Type",
"startIndex": 10,
"endIndex": 13,
"score": 0.7615982
}
]
}
3
Response
Your HR rep is John Smith johns@contoso.com 212-555-1212
bot at 1:28:00 PM
Type your message...

DAWN

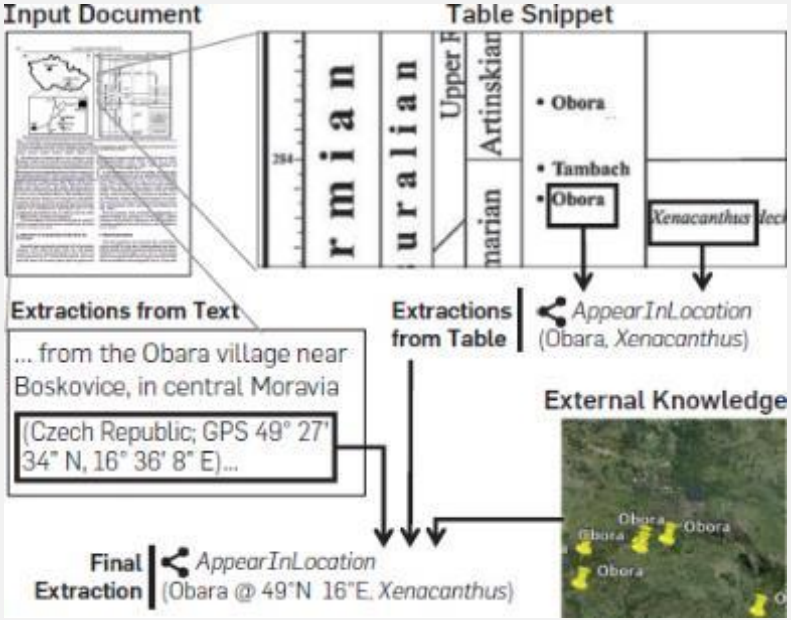
- **Dawn is an end to end data science solution categorized into four main products**
- **Macrobase: Analytics and real time monitoring of streams**
- **Snorkel: Learning from weakly labeled data**
- **Weld and Delite: 100-1000x faster Data Science**
- **Eagle: Virtualized Sensing of the physical world.**

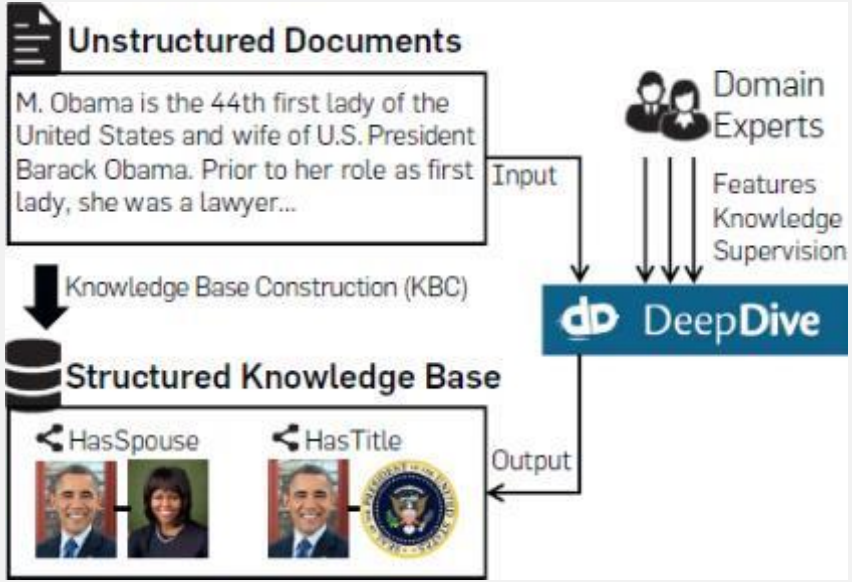
Deepdive

- **Dark Data is data that resides in text, tables, and figures.**
- **This is not easily processed by existing software and thus needs a special way to be "unlocked"**
- **Deep Dive solve this problem, while maintaining a high quality standard.**
- **Data can than be used in both AI components and Visualization tools such as Tableau.**
- **Can be considered a Very Intelligent data extractor at scale**

- PaleoDB is a Volunteer data store where Scientists contribute data from research papers into a structured format inside a database.
- One grad student at Stanford was able to create a dataset with the same papers and with an accuracy rate of 90%, beating hundreds of scientists and hours of labor.
- https://cs.stanford.edu/people/chrismre/papers/dd.pdf

Knowledge & Entity Extraction from PDFs and other Unstructured Sources using Deep Dive

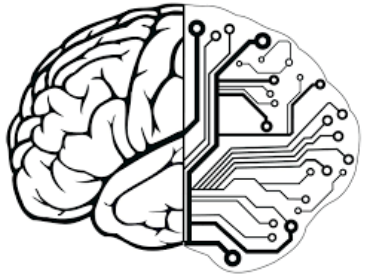




Macrobase

- **Human attention is limited and Data sizes are ever increasing at an exponential speed. I.E It's becoming more difficult to keep track of everything**
- **Macrobase is a pipeline management tool that keeps track of data flow while at the exact same time surfacing insights you may have missed.**
- **Given Labeled data or supervision rules, Macrobase executes a set of supervised and unsupervised models to leverage both domain knowledge and learn unknown behaviors.**

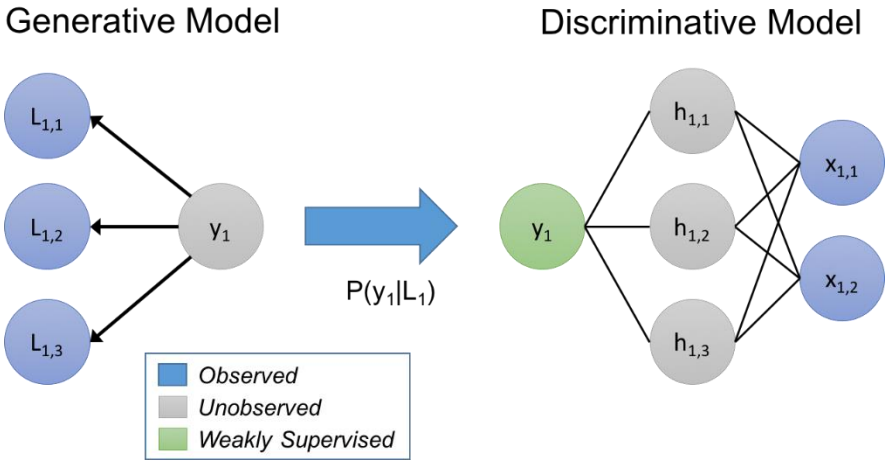
- Can be used to fuse multiple data sources together, commonly seen in internet of things projects.
- It is an open source project allowing for implementation to be easily fitted for the given usecase.
- https://github.com/stanford-futuredata/macrobase
- https://columbiaviz.github.io/files/papers/macrobase.pdf



Snorkel

- **Snorkel is an open source system for creating, modeling, and managing training data.**
- **Based on the same techniques in Deepdive, Snorkel is an open source implementation**
- **Deepdive was further commercialized into a product called Lattice**
- **Snorkel is still in the early days but the open source community has been heavily involved growing the project.**
- **https://github.com/HazyResearch/snorkel**

The *New* New Oil



DIFFERENT TECHNIQUES OF TEXT ANALYSIS

- WORD EMBEDDING
- TF-IDF
- WORD2VEC
- DOC2VEC
- LATENT DIRECHLET ALLOCATION (LDA)
- ELMO
- BIDAF

WORD EMBEDDING

- **Word embeddings are a set of feature engineering techniques widely used in predictive NLP modeling.**
- **Word embeddings transform sparse vector representations of words into a dense, continuous vector space.**
- **It identifies similarity between words and phrases on a large scale based on their context.**

EXAMPLES OF USES OF WORD EMBEDDING

Cat and dog: Both cute animals, can be pets, have 2 eyes, 4 legs, and one nose.

Audi and BMW: Both powerful expensive German automobile companies.

Word embeddings can also be trained to identify relations such as:

KING - MAN + WOMAN = QUEEN

WORD2VEC

- Word2Vec is a word embedding model created by google.
- It is a predictive embedding model which means it is trained to predict a target word from the context of its neighboring words.
- The model first encodes each word using one-hot-encoding, then feeds it into a hidden layer using a matrix of weights; the output of this process is the target word.

Word2Vec utilizes two different types of model architecture for computing vector representations of words:

- A) Continuous Bag-of-Words (CBOW)
- B) Skip-gram
- It is used for:
- 1) Compare word similarity between 2 sentences.
- 2) Compare query vector with document vector and retrieve documents.
- 3) Recommend music/videos basing on user likes.

WORD2VEC

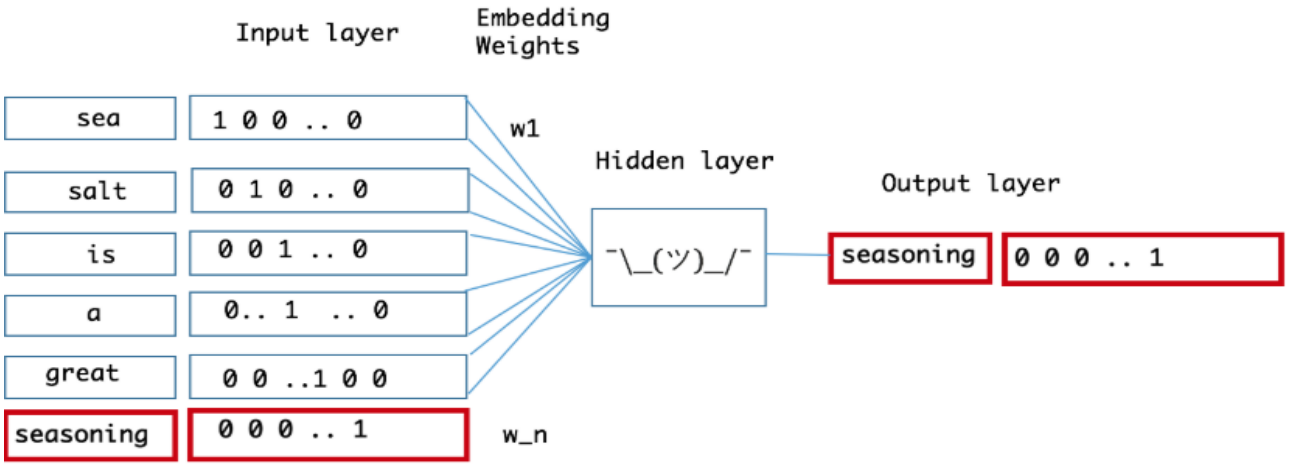
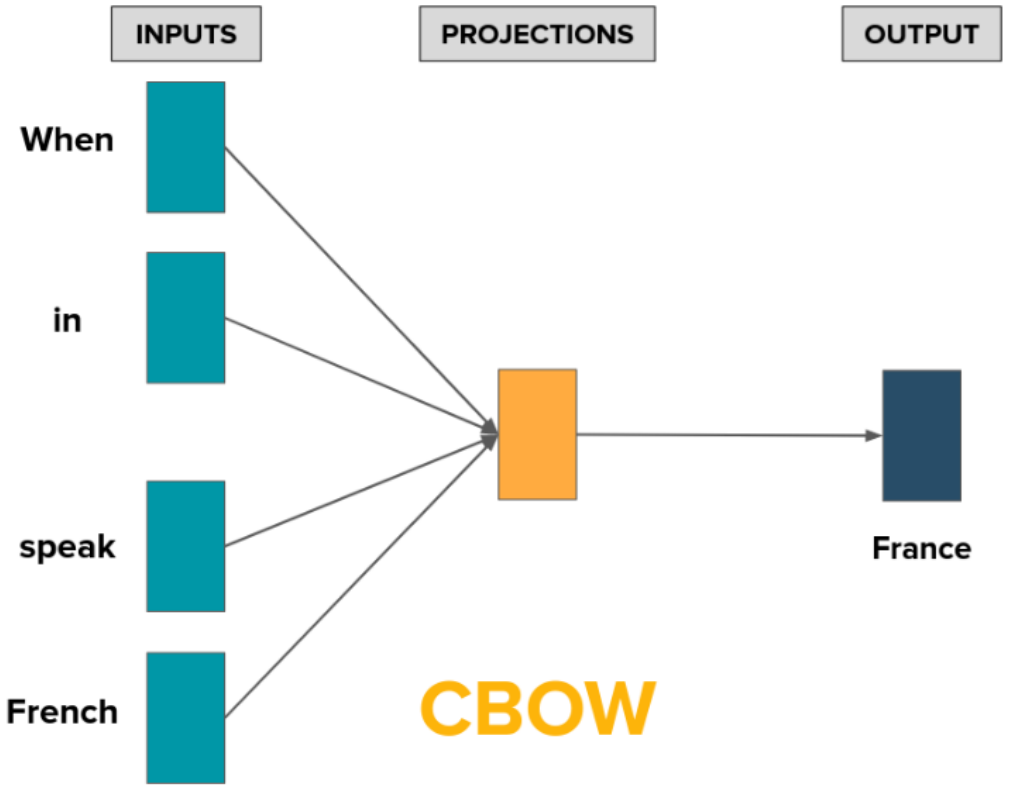


FIG. IMAGE OF WORD2VEC MODEL

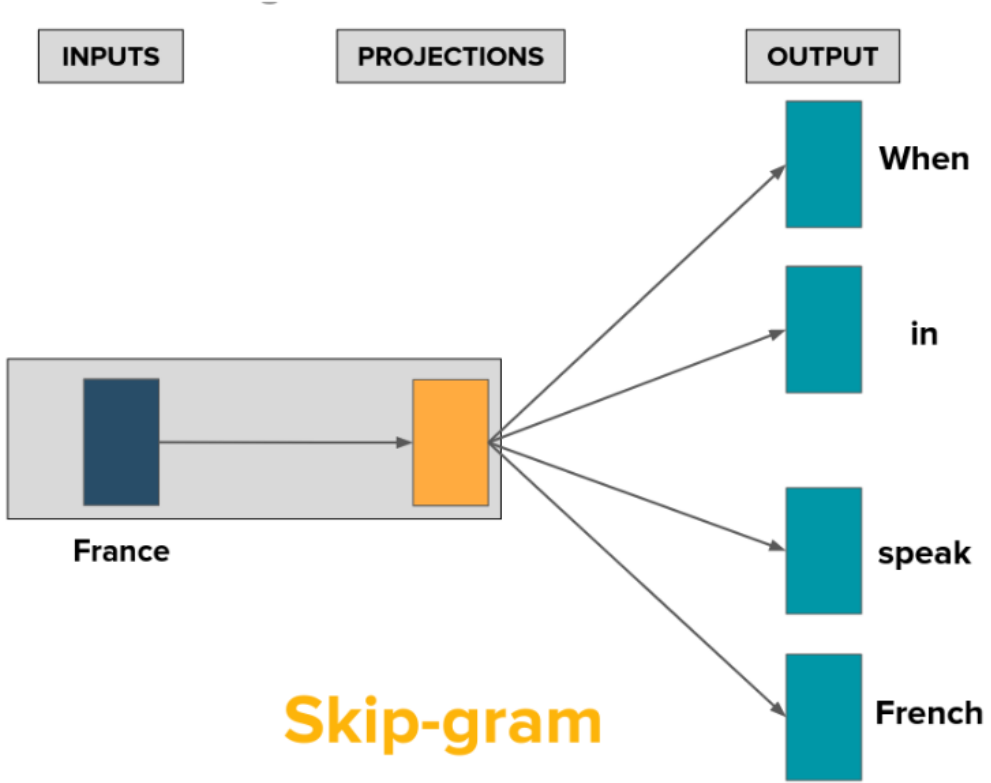
CONTINUOUS BAG OF WORDS(CBOW)

- **In the CBOW model, the aim is to fill in the missing word given its neighboring context.**



SKIP-GRAM

- In the skip-gram model, the aim is to predict the context.
- Skip-gram is slower while it does a better job for infrequent words.



DOC2VEC

- Doc2vec is an unsupervised algorithm to generate vectors for sentence/paragraphs/documents.
- Doc2vec is an adaptation of word2vec which can generate vectors for words.
- The vectors generated by doc2vec can be used for tasks like finding similarity between sentences/paragraphs/documents.
- For sentence similarity tasks, doc2vec vectors may perform reasonably well. However if the input corpus is one with lots of misspellings like tweets, this algorithm may not be the ideal choice.

DOC2VEC

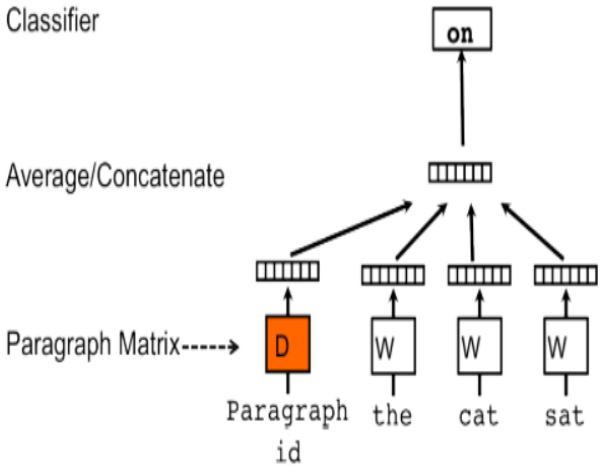


fig 3: PV-DM model

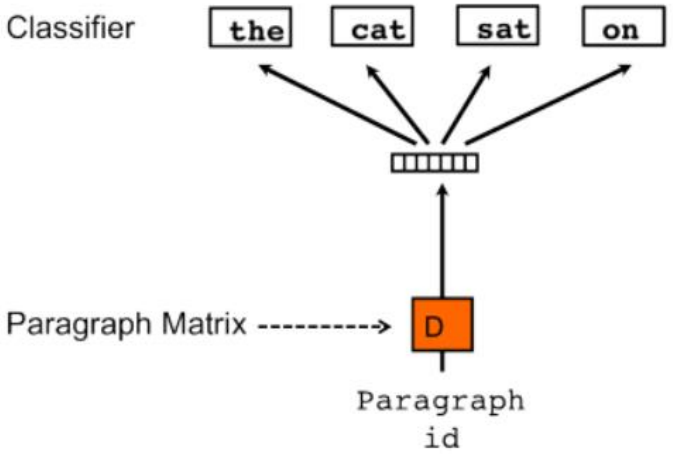


fig 4: PV-DBOW model

Term Frequency Inverse Document Frequency (TF-IDF)

- TF-IDF is a numerical statistic that is intended to reflect how important a word or n-gram is to a document in a collection or corpus.
- TFIDF provides some weighting to a given word based on the context it occurs.
- The tf–idf word app document

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

of times a r of

$tf_{i,j}$ = number of occurrences of i in j
df_i = number of documents containing i
N = total number of documents

Term Frequency Inverse Document Frequency (TF-IDF)

- **Tf–idf weighting scheme are often used by search engines as a central tool in scoring and ranking a document's relevance given a user query.**
- **Tf–idf can be successfully used for stopwords filtering in various subject fields, including text summarization and classification.**
- **However even though tf-idf BoW representations provide weights to different words they are unable to capture the word meaning.**
- **Practical Usage: Used by search engines for scoring and ranking documents basing on user query.**

Latent Dirichlet Allocation (LDA)

- LDA is an example of a topic model(statistical model for discovering abstract topic).
- In LDA, each document is viewed as a mixture of topics that are present in the corpus.
- The model proposes that each word in the document is attributable to one of the document's topics.
- Thus, LDA is a mathematical method for finding the mixture of words that is associated with each topic, while also determining the mixture of topics that describes each document.
- Input: Collection of documents
- Output: i) word to topic probablilty matrix
- ii) document to topic probability matrix

Latent Dirichlet Allocation (LDA)

- LDA is useful when you have a set of documents, and you want to discover patterns within, but without knowing about the documents themselves.
- LDA can be used to generate topics to understand a document's general theme.
- LDA is used in recommendation systems, document classification, data exploration, and document summarization.
- LDA is also used for dimensionality reduction in Machine Learning models or methods.

Bi-Directional Attention Flow (BIDAF)

- **BIDAF is a multi-stage process representing the context at different levels of granularity and uses a bi-directional attention flow mechanism to achieve a query aware context representation.**

1. Character Embedding Layer - maps each word to a vector space using character-level CNNs.

2. Word Embedding Layer - maps each word to a vector space using a pre-trained word embedding model.

3. Contextual Embedding Layer - utilizes contextual cues from surrounding words to refine the embedding of the words. These first three layers are applied to both the query and context.

4. Attention Flow Layer - couples the query and context vectors and produces a set of query aware feature vectors for each word in the context.

5. Modeling Layer - employs a Recurrent Neural Network to scan the context.

6. Output Layer - provides an answer to the query.

Bi-Directional Attention Flow (BIDAF)

- **Attention Flow Layer - In this layer, attentions are computed in two directions: from context to query as well as from query to context.**
- **Context-to-query (C2Q) attention signifies which query words are most relevant to each context word.**
- **Query-to-context (Q2C) attention signifies which context words have the closest similarity to one of the query words and are hence critical for answering the query.**

- **Potential Uses – BIDAF can be widely used for question answering data and can also be used in modern chatbots or comer support system to detect potential queries and significant answers.**

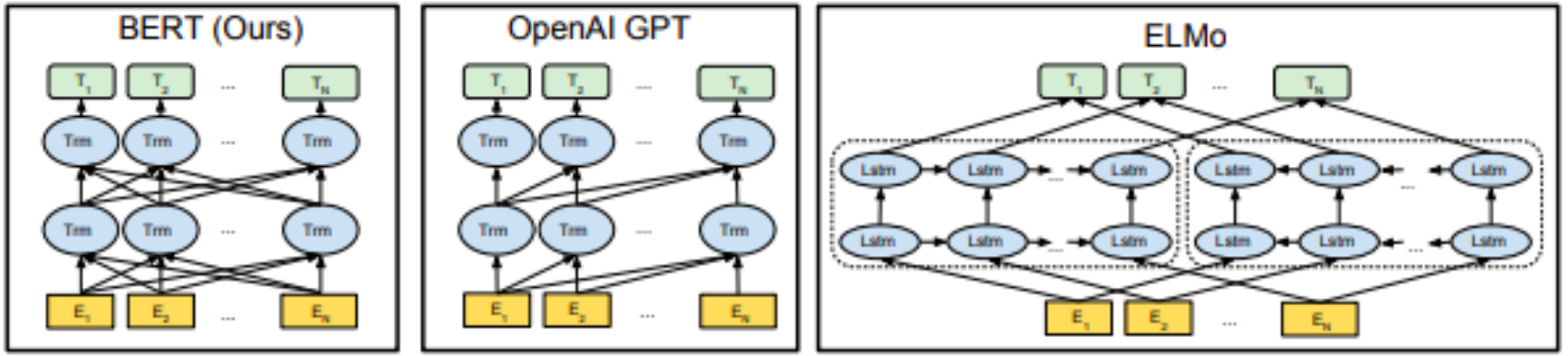
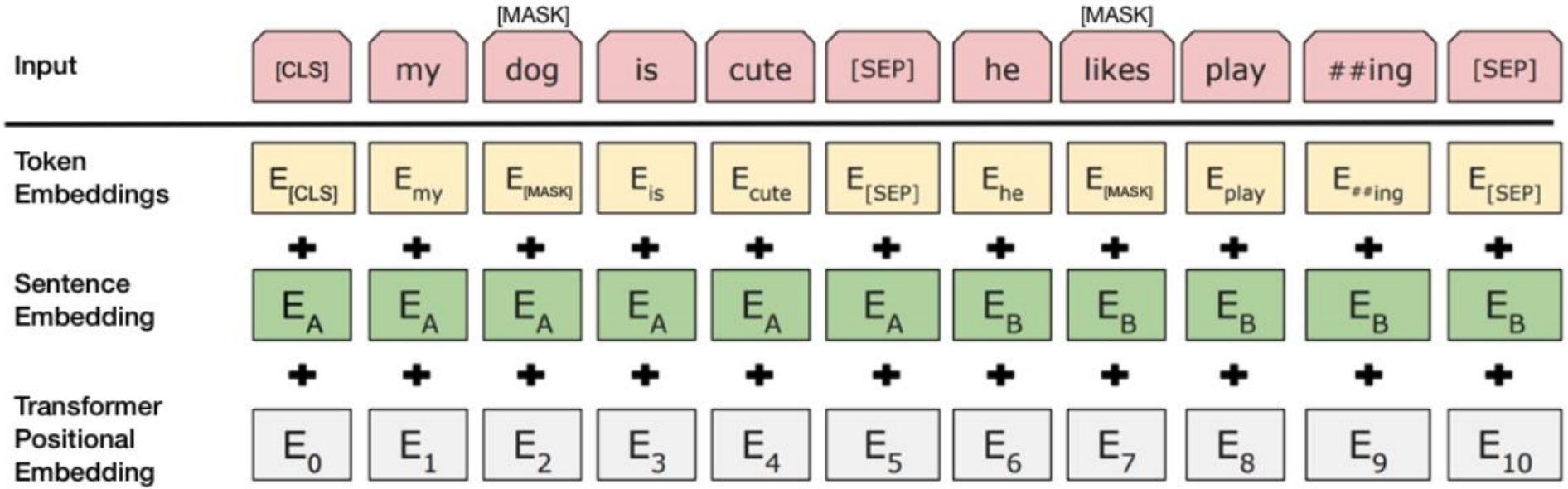
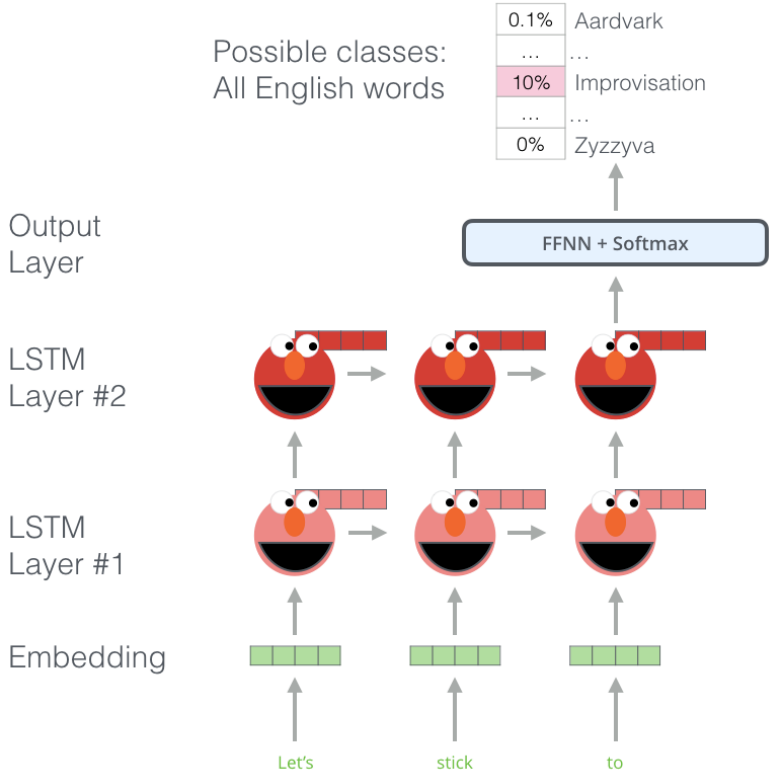


Figure 1: Differences in pre-training model architectures. BERT uses a bidirectional Transformer. OpenAI GPT uses a left-to-right Transformer. ELMo uses the concatenation of independently trained left-to-right and right-to-left LSTM to generate features for downstream tasks. Among three, only BERT representations are jointly conditioned on both left and right context in all layers.

BERT - Bidirectional Encoder Representations from Transformers



ELMo - Deep contextualized word representation



Questions?

- adnanmasood@gmail.com
- https://twitter.com/adnanmasood
- https://www.linkedin.com/in/adnano

Please use EventsXD to fill out a session evaluation.

Thank you!

Microsoft Azure
+ AI Conference
CO-PRODUCED BY
Microsoft & DEVintersection