Graduate Qualifying Project & Professional Journey

Vandana Anand
October 27th, 2021

Agenda

(1) Introduction

(2) GQP Project Overview

(3) Key Takeaways & Learnings from GQP

(4) Verizon Position, Work Responsibilities, How GQP Helped

(5) Tips and Skills for Data Science Interviews and Jobs

(6) Questions & Answers

Vandana Anand

Bachelor's in Computer Science '19

Master's in Data Science '20

Digital Marketing Catalyst VLDP (Data Science Analyst) on Business Enablement team

Hometown: Boston MA

Love travelling: 11 countries & counting!

Acapella

Badminton

GQP Project

Entity Linking
WPI-Basis Technology Project

Sept 2020-Dec 2020

Team members: Kratika Agrawal, Vandana Anand, Xinlu He, Min Huang, Soumya Joshi, Jing Yu

Basis Technology Mentors: Gil Irizarry, Kfir Bar, Lital Ravid, Karin Lin, Zachary Yocum

WPI Mentors: Prof. Chun-Kit Ngan, Prof. Fatemeh Emdad

Project Team Introduction

Sponsors & Mentors

Gil Irizarry
Vice President - Engineering
gil@basistech.com

Kfir Bar
Chief Scientist
kfir@basistech.com

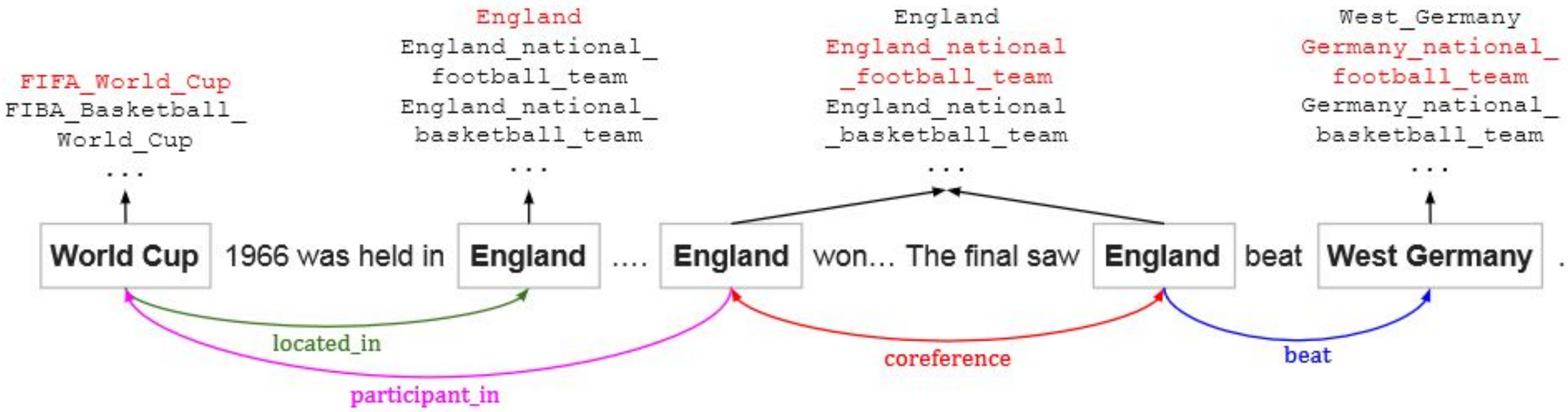
Latent Relations		End2End		DeepType	
					
Soumya Joshi sjoshi2@wpi.edu	**Min Huang** mhuang3@wpi.edu	**Vandana Anand** vanand@wpi.edu	**Xinlu He** xhe4@wpi.edu	**Kratika Agrawal** kagrawal@wpi.edu	**Jing Yu** jyu5@wpi.edu

Project Goal

Goal: Improve BT's current entity linking tool **Rosette** by applying a novel model to increase entity linking performance.

ROSETTE
TEXT ANALYTICS

Our Entity Linking System



- <u>Named Entity Linking / Entity disambiguation method</u> explained above includes the **co-reference** which is a relation between two or more mentions in a text when they refer to the same entity

Project Timeline

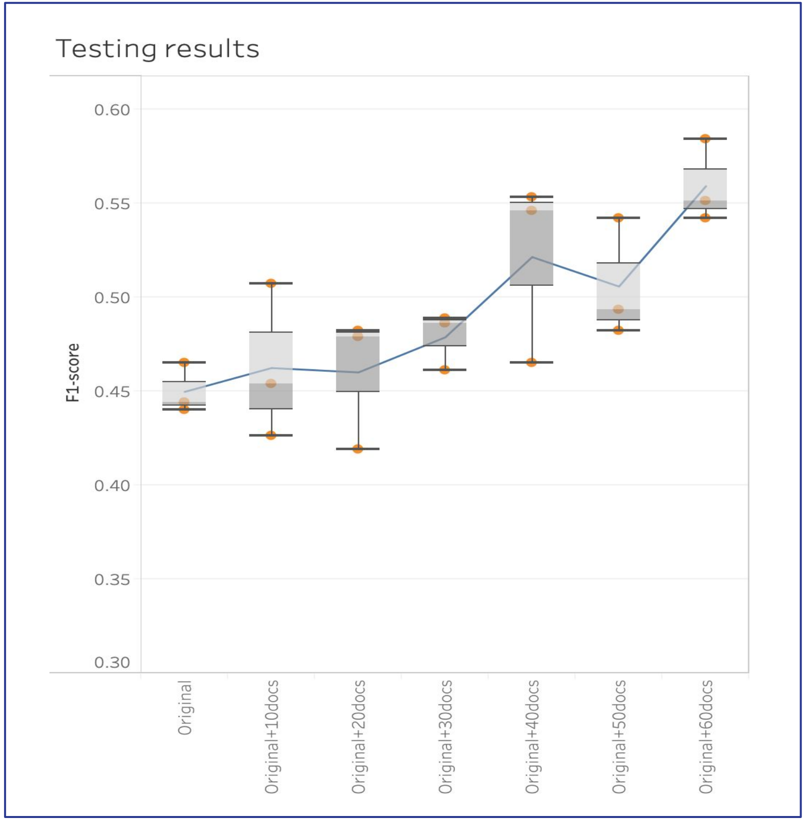
	Week 1	Week 2	Week 3	Week 4	Week 5	Week 6	Week 7	Week 8	Week 9	Week 10	Week 11	Week 12	Week 13	Week 14
TASK	8-Sep-20	14-Sep-20	21-Sep-20	28-Sep-20	5-Oct-20	12-Oct-20	19-Oct-20	26-Oct-20	2-Nov-20	9-Nov-20	16-Nov-20	23-Nov-20	30-Nov-20	7-Dec-20
Team & Project Introduction														
Exploring tool Rosette														
Annotation Tool Review														
Read relevant papers														
Paper Presentations														
Compare Evaluation Metric														
Implementation: 3 Teams														
Transforming BT Dataset														
Evaluation on BT Dataset														
Discussion of Novel Approach														
Novel Approach Implementation														

Latent Relations Model

Research Problems Addressed

- Tackled candidate selection problem which is to choose best entity from a list of candidates by
  - local and global modeling
  - Innovatively including latent relations between mentions into the global model
- Eliminated the need for extensive feature engineering by using representation learning to learn relation embeddings

Results



- Successfully applied BT dataset into the latent relations model
- Better F1 score with larger BT dataset

Conclusion & Future Scope

Conclusion:

- Successfully applied BT dataset into latent relations model
- Better F1 score with larger BT dataset

Future Scope:

- More training rounds are needed to reduce the randomness
- Larger BT dataset is required
  - Generate candidates based on latest version wikipedia KB
  - Might bring higher F1 score

Project References

- **Improving Entity Linking by Modeling Latent Relations between Mentions:** https://arxiv.org/pdf/1804.10637.pdf
- **Latent Relations Model Github:** https://github.com/lephong/mulrel-nel
- **Deep Joint Entity Disambiguation with Local Neural Attention:** https://arxiv.org/pdf/1704.04920.pdf
- **Deep Joint Model Github:** https://github.com/dalab/deep-ed

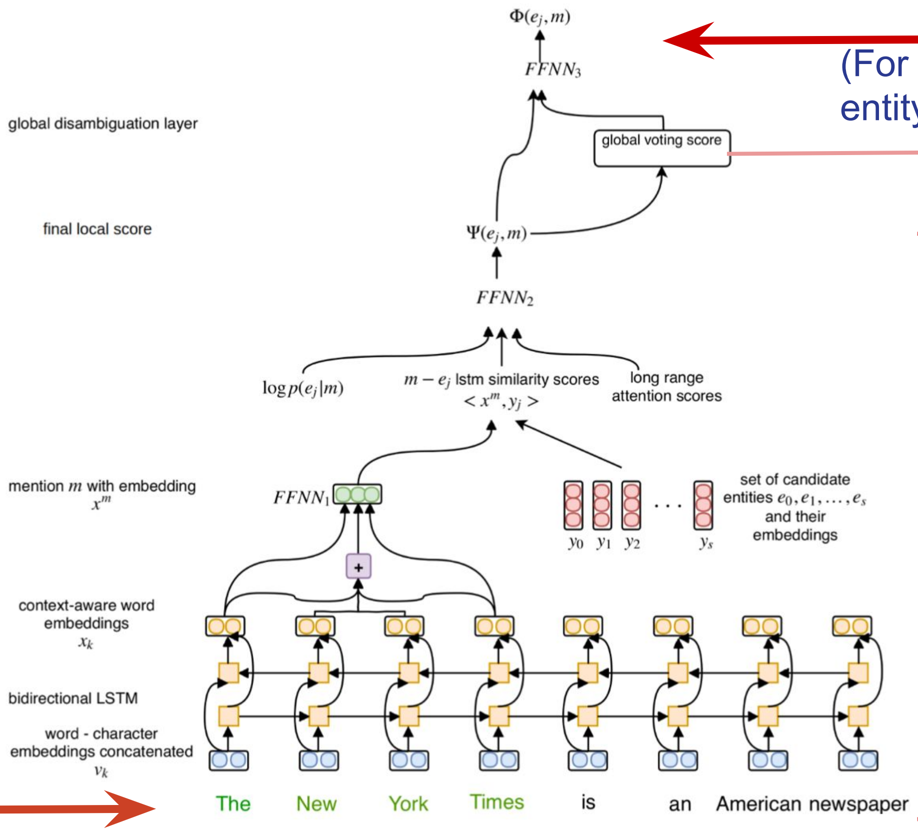
End2End Model

Advantage of the End2End Model

	Other Research	End2End Model
Original Text	Worcester Polytechnic Institute is a private research university in MA.	
	Entity Extractor	End2End Model
Mention Detection	Worcester MA	
	Entity Disambiguation Model	
Entity Disambig-uation	Worcester Massachusetts	Worcester_Polytechnic_Institute Massachusetts

Overview of the End2End Paper

Output: Final Score
(For mention-entity pair: Is this entity candidate suitable here)



local score

try different spans--mention

local score

Input: the Whole Text

End2End Paper - Candidate Selection

- Mention--Entity Candidate pair:
  e.g:
  Discovery channel — Discovery Channel--basic cable and satellite television
  Star Trek: Discovery--American television series
  Discovery--space shuttle orbiter
  Discovery--2001 album by Daft Punk

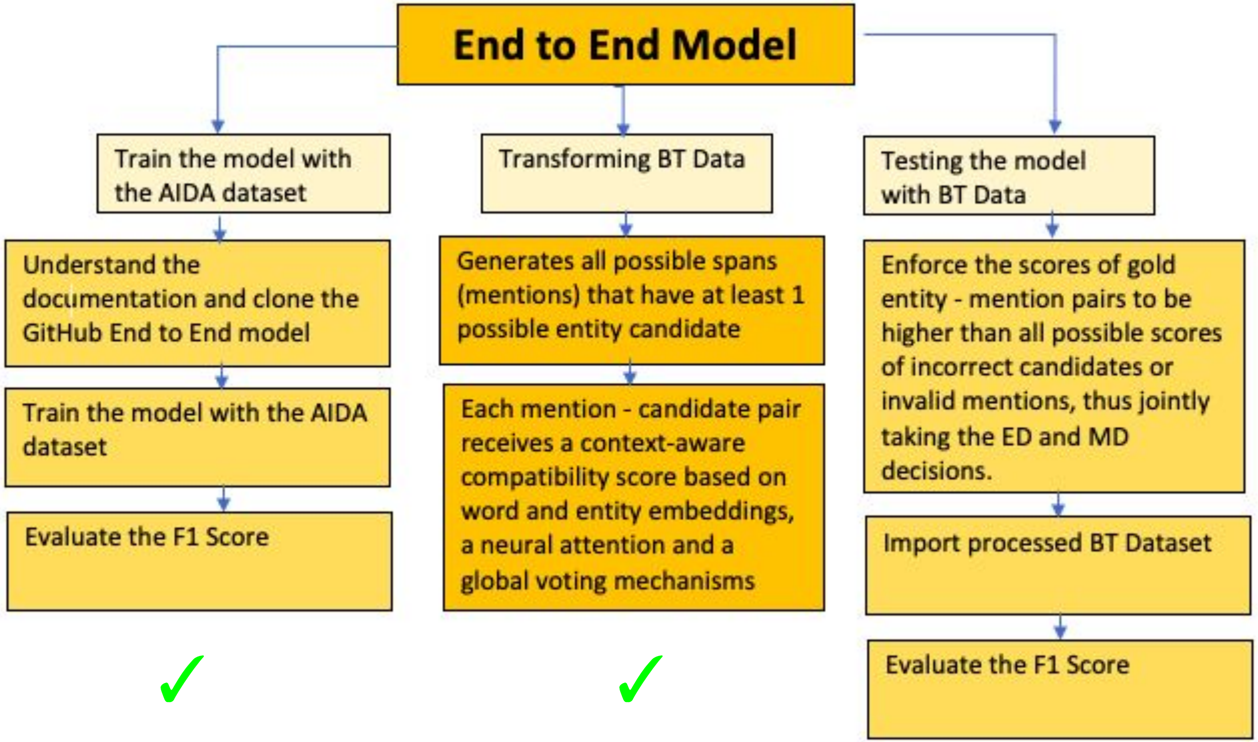
- How to get candidate-- Prior Possibility
  - The ratio of the times the pair appears to the times the mention appears in KB
  - KB: a lot of text sample--Wikipedia2014
  - 30 candidates

- If already know the mention (eg: The New York)
  - lose advantage: loose part of the context information
  - don’t have to analyze all the spans

End2End Paper - Other Skills

- Word2Vec
  - transfer the word into vector
  - the cosine similarity indicates the level of semantic similarity between the words
  - pre-trained
- Bi-derection LSTM
  - both inside a word and the mention between word
  - lexical information in Character Embeddings
  - context- aware

Paper Implementation Workflow



Preparing the dataset

Datasets

aida_train
aida_test
aida_dev
ace2004
aquaint
clueweb
msnbc
wikipedia

Format

Text

Electoral_division_of_Apsley

```
The Electoral division of Apsley is one of the 15 electorates
or seats in the Tasmanian Legislative Council. It is the
second-largest upper house electorate in the state by area
after Murchison.
```

XML

```xml
<?xml version="1.0" encoding="UTF-8"?>
<wikipediaData.entityAnnotation>
    <document docName="Electoral_division_of_Apsley">
        <annotation>
            <mention>Tasmanian Legislative Council</mention>
            <wikiName>Tasmanian Legislative Council</wikiName>
            <offset>78</offset>
            <length>29</length>
        </annotation>
        <annotation>
            <mention>Murchison</mention>
            <wikiName>Electoral division of Murchison</wikiName>
            <offset>184</offset>
            <length>9</length>
        </annotation>
    </document>
```

Preprocessing Output

Wikipedia

```
DOCSTART_Electoral_division_of_Apsley
The
Electoral
division
of
Apsley
is
one
of
the
15
electorates
or
seats
in
the
MMSTART_579457
Tasmanian
Legislative
Council
MMEND
.
It
is
the
second-largest
MMSTART_579457
upper
house
MMEND
```

AIDA

```
-DOCSTART- (1 EU)
EU      B       EU      --NME--
rejects
German  B       German  Germany http://
en.wikipedia.org/wiki/Germany   11867   /m/
0345h
call
to
boycott
British B       British United_Kingdom
http://en.wikipedia.org/wiki/United_Kingdom
        31717   /m/07ssc
lamb
.

Peter   B       Peter Blackburn --NME--
Blackburn       I       Peter Blackburn --
NME--

BRUSSELS        B       BRUSSELS
Brussels        http://en.wikipedia.org/
wiki/Brussels   3708    /m/0177z
1996-08-22
```

Training the Model

Entity Linking Training

```
Evaluating EL datasets
Best validation threshold = 0.053 with F1=90.1
aida_dev
micro P: 90.1    R: 90.2          F1: 90.1
macro P: 88.1    R: 88.3          F1: 88.2
aida_test
micro P: 83.6    R: 82.1          F1: 82.8
macro P: 83.8    R: 84.3          F1: 84.1
aida_train
micro P: 96.0    R: 95.4          F1: 95.7
macro P: 95.3    R: 95.2          F1: 95.3
ace2004
micro P: 19.3    R: 69.0          F1: 30.2
macro P: 21.3    R: 61.4          F1: 31.6
aquaint
micro P: 38.2    R: 43.3          F1: 40.6
macro P: 40.1    R: 41.8          F1: 40.9
clueweb
micro P: 45.7    R: 49.1          F1: 47.3
macro P: 53.6    R: 49.0          F1: 51.2
msnbc
micro P: 78.1    R: 76.3          F1: 77.2
macro P: 79.5    R: 74.5          F1: 76.9
wikipedia
micro P: 40.9    R: 42.8          F1: 41.8
macro P: 44.1    R: 43.4          F1: 43.7
(end2end_neural_el_env) [wpi-project@wpi-gcp-2021-2
```

Model Performance

Entity Disambiguation Training

```
Evaluating ED datasets
Best validation threshold = -0.037 with F1=93.8
aida_dev
micro P: 94.5    R: 93.1          F1: 93.8
macro P: 93.1    R: 91.9          F1: 92.5
aida_test
micro P: 89.2    R: 85.4          F1: 87.2
macro P: 90.0    R: 88.1          F1: 89.1
aida_train
micro P: 97.3    R: 96.1          F1: 96.7
macro P: 97.0    R: 95.9          F1: 96.5
ace2004
micro P: 92.6    R: 83.9          F1: 88.1
macro P: 93.8    R: 85.1          F1: 89.2
aquaint
micro P: 92.4    R: 87.2          F1: 89.7
macro P: 92.4    R: 86.7          F1: 89.4
clueweb
micro P: 83.2    R: 72.3          F1: 77.3
macro P: 82.8    R: 72.6          F1: 77.4
msnbc
micro P: 94.4    R: 90.3          F1: 92.3
macro P: 95.4    R: 91.1          F1: 93.2
wikipedia
micro P: 78.2    R: 70.9          F1: 74.4
macro P: 78.7    R: 72.2          F1: 75.3
(end2end_neural_el_env) [wpi-project@wpi-gcp-2021-2
```

Testing the Model w/ AIDA and other datasets

Entity Disambiguation	**Train Score**	**Test Score**	**Test Precision**	**Test Recall**
Best Scores	92.2%	85.7%	88.5%	83%

Entity Linking	**Train Score**	**Test Score**	**Test Precision**	**Test Recall**
Best Scores	62.9%	56.8%	57.5%	56.1%
Using all Spans Best Score	89.1%	80.1%	83.5%	76.9%

Preparing the BT dataset


```
{
  "version": "1.1.0",
  "data": "On Wednesday, the total number of confirmed deaths linked to SARS-CoV-2
  "attributes": {
    "entities": {
      "type": "list",
      "itemType": "entities",
      "items": [
        {
          "mentions": [
            {
              "startOffset": 44,
              "endOffset": 50
            }
          ],
          "type": "SYMPTOM",
          "entityId": "Q4"
        }
      ]
    }
  },
  "documentMetadata": {
    "title": [
      "SARS-CoV-2 surpasses 100,000 confirmed deaths in the United States"
    ],
    "language": [
      "en"
    ],
    "direction": [
      "ltr"
    ],
    "published_date": [
      "2020-05-29"
    ],
    "accessed_date": [
      "2020-08-19T17:57:25.224462"
    ],
    "url": [
      "https://en.wikinews.org/wiki/SARS-CoV-
2_surpasses_100,000_confirmed_deaths_in_the_United_States"
    ],
    "domain": [
      "https://en.wikinews.org"
    ],
    "wikinews_categories": [
      "https://en.wikinews.org/wiki/Category:Archived",
      "https://en.wikinews.org/wiki/Category:COVID-19",
      "https://en.wikinews.org/wiki/Category:Coronavirus",
      "https://en.wikinews.org/wiki/Category:Disease",
      "https://en.wikinews.org/wiki/Category:Health",
```


JSON

→

```
DOCSTART_SARS-CoV-2_surpasses_100,000_confirmed_death
On
Wednesday
,
the
total
number
of
confirmed
MMSTART_4
deaths
MMEND
linked
to
MMSTART_84263196
SARS
-
CoV
-
2
MMEND
MMSTART_89469904
coronavirus
MMEND
MMSTART_166231
infections
MMEND
surpassed
100
```

Doc start: DOCSTART_fileName
row: word or punc
mention: start with "MMSTART_Wikiid"
end with "MMEND"
between paragraph: *NL*

```
processBT   pair_num: 3114
processBT   BTDB_num 207
processBT   unknown_id 0
processBT   file_num 70
```

Future Scope

- Testing the End2End Model with the BT dataset (cannot plug in directly)
- Rearranging the implementation code to be compatible with BT's data structure
- Evaluating the accuracy of the model on BT's dataset
- Running the model on BT's server and integrating with existing processes to improve entity procedures

Project Conclusion

- End2End: if only the word disambiguation process, the result cannot take much of an advantage of this model.
- End2End gave reasonable accuracy on the AIDA testing set because the number of mentions is so big. Since BT's data is lower we expect this number to be lower as well
- We learned:
  - Entity linking and entity disambiguation processes
  - Researching and understanding how the End2End model works
  - Preprocessing various datasets and transforming BT's data to be compatible with the model

Project References

- **Research Paper:** https://arxiv.org/pdf/1808.07699.pdf
- **End2End Github**: https://github.com/dalab/end2end_neural_el
- https://github.com/lephong/mulrel-nel (Entity Linking)
- https://github.com/basis-technology-corp/annotated-data-model
- https://github.com/basis-technology-corp/wpi-gqp-2020 (BT data)

DeepType Model:

Multilingual Entity Linking by Neural Type System Evolution

BASIS TECHNOLOGY | **Worcester Polytechnic Institute**

Deep Type: Implementation

- **Progress:**
  - Downloaded Wikidata & Wikipedia latest dumps
  - Building Type System
  - Followed varied approaches in building type system
- **Limitations:**
  - Large files, need even larger space to run the whole process
  - Incompatible packages and data dump
  - Needed more time for building Type System through WikiAPI
- **Lesson Learnt:**
  - Assumes strong coherence among entities
  - Entities can be easily disambiguated using Type System graph
  - More accuracy if learnability is more for Type System knowledge base

Deep Type: Reported Results in the paper

Model		enwiki	frwiki	dewiki	eswiki	WKD30	CoNLL	TAC 2010
M&W(Milne and Witten 2008)						84.6	-	-
TagMe (Ferragina and Scaiella 2010)		83.224		80.711		90.9	-	-
(Globerson et al. 2016)						-	91.7	87.2
(Yamada et al. 2016)						-	91.5	85.2
NTEE (Yamada et al. 2017)						-	-	87.7
LinkCount only		89.064**	92.013	92.013**	89.980	82.710	68.614	81.485
Ours	manual	**94.331****	92.967			91.888**	93.108**	90.743*
	manual (oracle)	97.734	98.026	98.632	98.178	95.872	98.217	98.601
	greedy	93.725**	**92.984**			**92.375****	94.151**	**90.850***
	greedy (oracle)	98.002	97.222	97.915	98.246	97.293	98.982	98.278
	CEM	93.707**	92.415			92.247**	93.962**	90.302*
	CEM (oracle)	97.500	96.648	97.480	97.599	96.481	99.005	96.767
	GA	93.684**	92.027			92.062**	**94.879****	90.312*
	GA (oracle)	97.297	96.783	97.408	97.609	96.268	98.461	96.663
	GA (English only)	93.029**				91.743**	93.701**	-

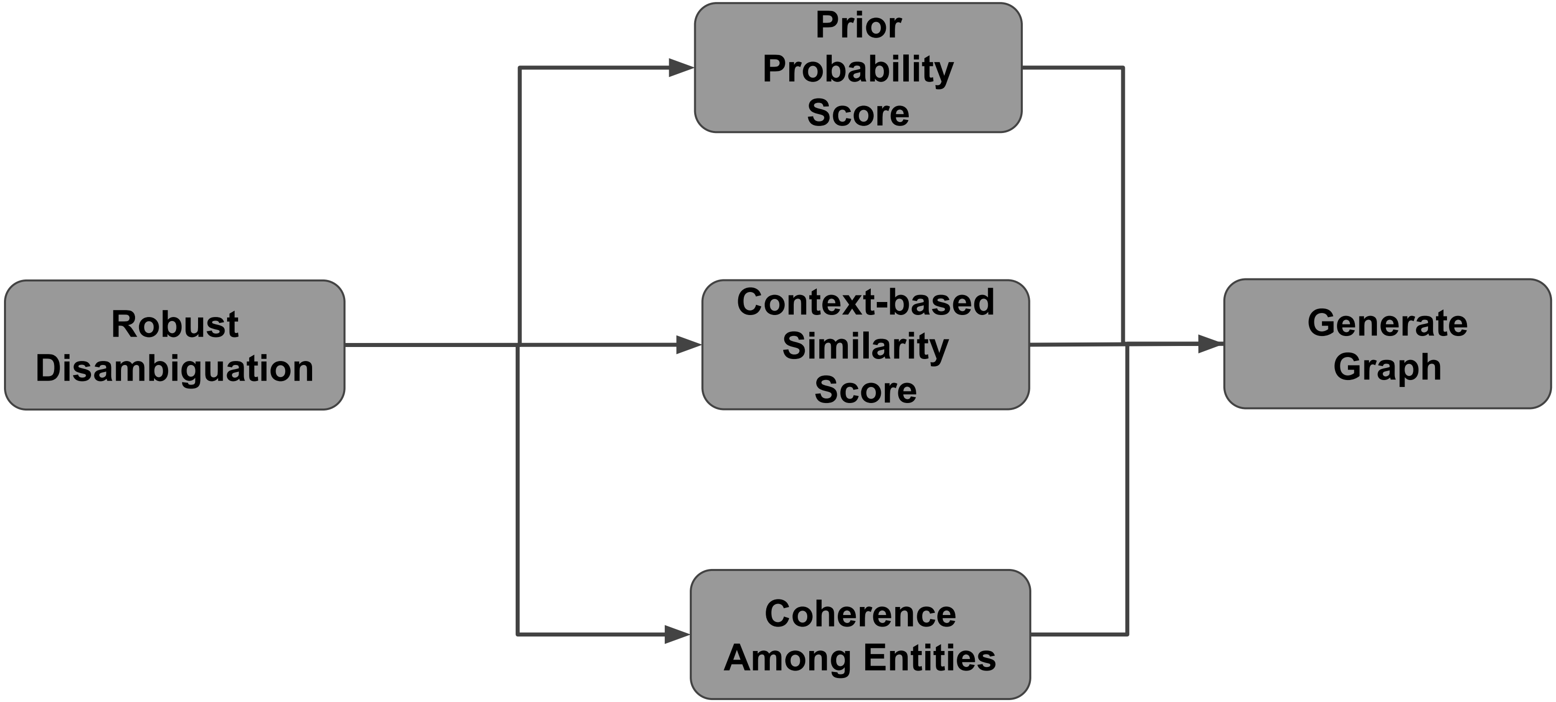
Significant improvements over prior work denoted by $*$ for $p < 0.05$, and $**$ for $p < 0.01$.

Project References

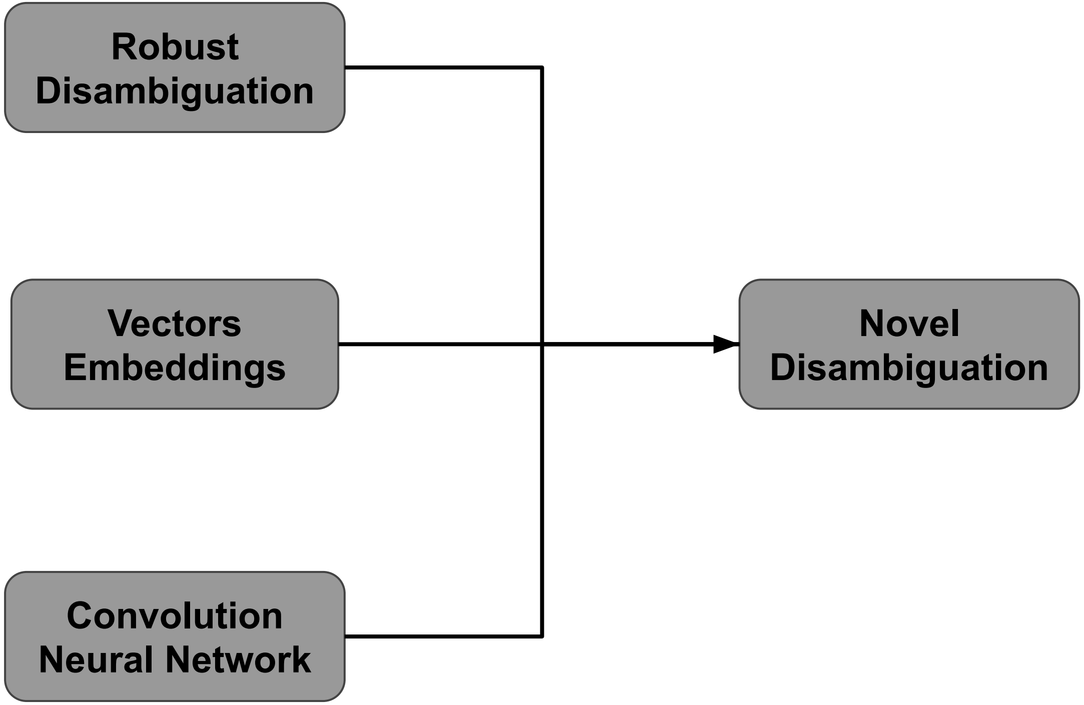
- **DeepType Multilingual Entity Linking by Neural Type System Evolution:** https://arxiv.org/pdf/1802.01021.pdf
- **DeepType Source Code**: https://github.com/openai/deeptype

Novel Approach with Robust Disambiguation

Robust Disambiguation Approach



Novel Disambiguation Approach



Novel Model: Graph Generation

Done:

- Use Robust Disambiguation technique code to produce graph
  https://github.com/codepie/aida
- Installed PostgreSQL 9.6.0
- Downloaded Yago dumps
- Imported Yago dumps in PostgreSQL
- Extract Mention Nodes, Entity Nodes and edge information from the generated graph

They performed Kashmir, written by Page and Plant. Page played unusual chords on his Gibson.

```
Generated Graph:
0
node = Kashmir, From:2/2, To:2/2, Offset: 15, Length: 7
node type = MENTION
successors = {}
1
node = Gibson, From:16/16, To:16/16, Offset: 85, Length: 6
node type = MENTION
successors = {}
```

Future Scope

- Establish connection between PostgreSQL and AIDA repository
- Develop weighted graph
- Translate weighted Graph => weighted Adjacency Matrices
- Generate Graph: AIDA dataset and Basis Technology dataset
- Get Embeddings: AIDA dataset and Basis Technology dataset
- Perform Training of CNN model
- Evaluate performance on Basis Technology dataset

Project References

- **Robust Disambiguation of Named Entities in Text:** https://www.aclweb.org/anthology/D11-1072.pdf
- **AIDA Online Demo:** https://aida.dor.ai/overview
- **AIDA Source Code:** https://github.com/codepie/aida
- **Wikipedia2Vec:** https://wikipedia2vec.github.io/wikipedia2vec/

Overall Learning

Conduct research and apply models in a hands - on approach

Exposure to various deep learning and natural language processing techniques

Preprocessing a dataset is very important and can be an involved step

Learnings

Understanding and overcoming technical challenges

Importance of formatting train/test datasets as it can alter a model’s performance

Thank You

Sponsors & Mentors:

Gil Irizarry
Vice President -Engineering
gil@basistech.com

Kfir Bar
Chief Scientist
kfir@basistech.com

Instructors:

Prof. Chun-Kit (Ben) Ngan
Professor
cngan@wpi.edu

Prof. Fatemeh Emdad
Professor
femdad@wpi.edu

Prof. Elke A. Rundensteiner
Director of Data Science
rundenst@wpi.edu

Verizon Role: Digital Marketing Catalyst

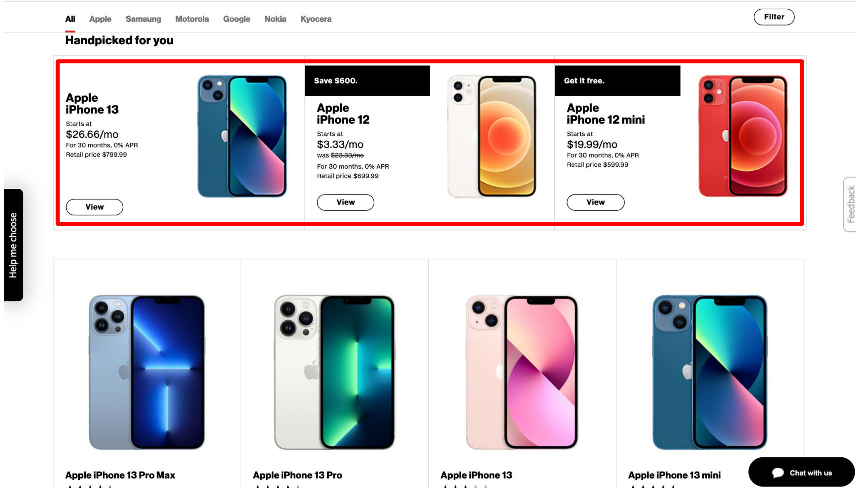
Verizon Position

Verizon Leadership Development Program (VLDP)

- Designed to drive success of Verizon's future leaders.
- Provides college graduates a 3-4 year immersive experience as the build leadership skills
- Customized job rotations, leadership development curriculum, networking activities, hands-on learning, and opportunity to present to senior leadership

Marketing Catalyst in Digital Sales

- **Digital Sales** = driving marketing, sales and growth on Verizon.com using Artificial Intelligence (AI) and Machine Learning (ML)



Marketing Catalyst in Digital Sales

What do we do?

Innovate and initiate business problems to drive AI driven solutions, derive actionable insights, and push forward marketing and sales.

- Understand the business problem
- Translate technical tasks that need to be carried out
- Connect with the right team members
- Apply problem solving techniques
- Follow through with the team to deliver results

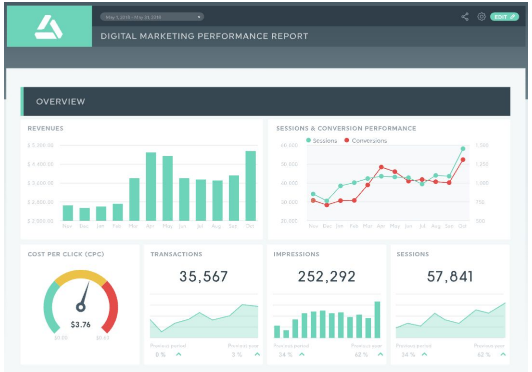
Teams we interact with:

- Business
- Marketing
- Data Engineering
- Global Technology Services (GTS) - Frontend/Backend engineers & Architects
- AI Center (Data Scientists)

Projects

Consolidate Reporting

Establish a one stop shop to evaluate models and personalization services on the digital platform



Analysis

Use reporting tools to assess model performance as well as user behavior, trends, and interaction. **Identify** areas of opportunities to implement AI/ML. **Derive** business value for model/experience implementations

CONTENT SQUARE

SQL

Adobe Analytics

Confluence Page

Document team's projects and share with the business to get up to speed on evaluating the enabled services

Confluence

How GQP Helped

- Hands on project based experience and effective teamwork skills
- Learning to run effective meetings
- Presenting to sponsors that prepared me for executive leadership level presentations
- Translating complex technical concepts into understandable bullet points
- Exposure to different type of models that I can apply at work and suggest innovative approaches to business problems

Tips & Skills

Interviews & Workplace Tips/Skills

Behavioral

- Prepare by thinking of 2-3 teamwork experiences
- Being spontaneous takes practice!
- Soft skills are *very* important in the real world
- It’s okay to take risks, we learn valuable lessons when we pick up ourselves
- Take business classes to understand the concepts

Technical

- Do your research about interesting and innovative technologies applied in the field
- Talk about your ideas and how you might integrate into the current process and develop it
- Leetcode for programming, SQL and Python are significant
- Having technical knowledge is key to communicating effectively with various teams

Q&A