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

Team Members

					
Vandana Anand vanand@wpi.edu 978-427-5212	Kratika Agrawal kagrawal@wpi.edu 774-701-9288	Jing Yu jyu5@wpi.edu 608-886-8578	Soumya Joshi sjoshi2@wpi.edu 508-615-8354	Min Huang mhuang3@wpi.edu 774-253-5410	Xinlu He xhe4@wpi.edu 774-253-8221

Project Goal

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

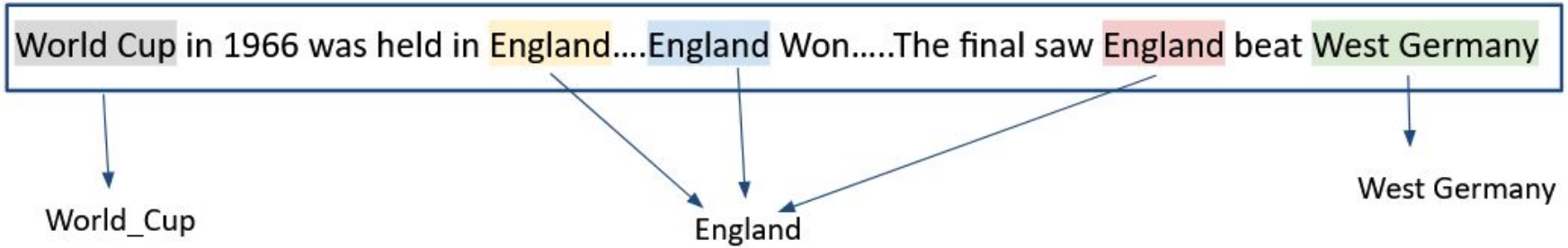
Project Timeline:

TASK	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
	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 Paper

Overview of the Latent Relations Paper

Other Entity Linking Systems:

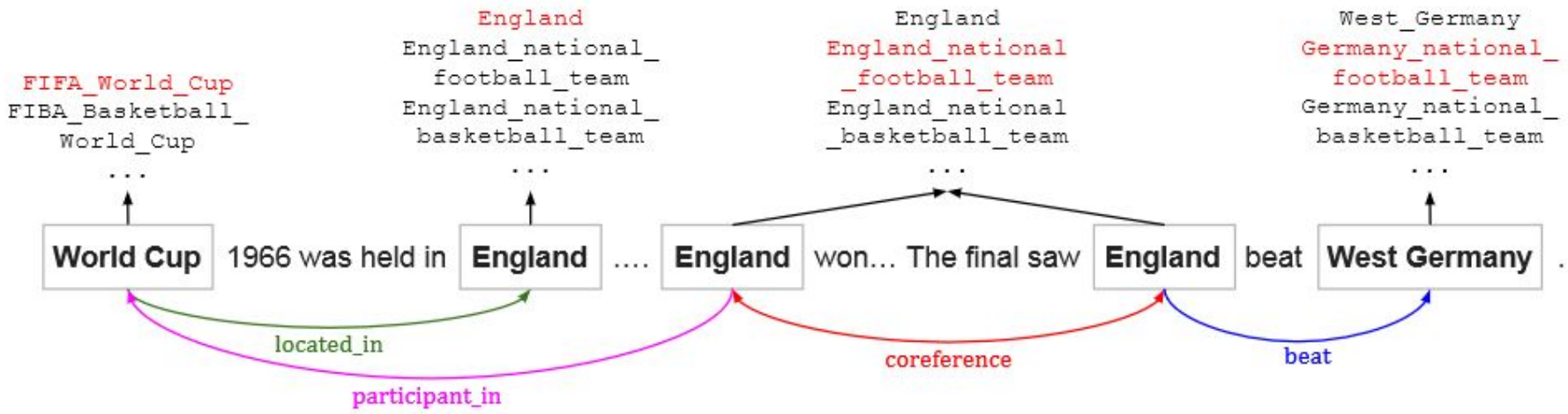


Problems:

1. Here World_Cup refers to FIFA_World_Cup
2. The 1st entry England refers to the England(The country)
3. The 2nd entry England and 3rd entry England refers to the England Football Team

Overview of the Latent Relations Paper

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

Named Entity Recognition:

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

1. **Entity Recognition**: Given a text, identify all Named Entities (Nouns)

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

2. **Entity Linking/Entity Disambiguation**:

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

Kashmir (Region) or **Kashmir (Song)** or **Kashmir (Band)**

Larry Page or **Page, Arizona** or **Jimmy Page**

Robert Plant or **John Plant** or **Plant (film)**

Gibson Les Paul or **Gibson, Missouri**

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

Kashmir (Song) Jimmy Page Robert Plant Gibson Les Paul

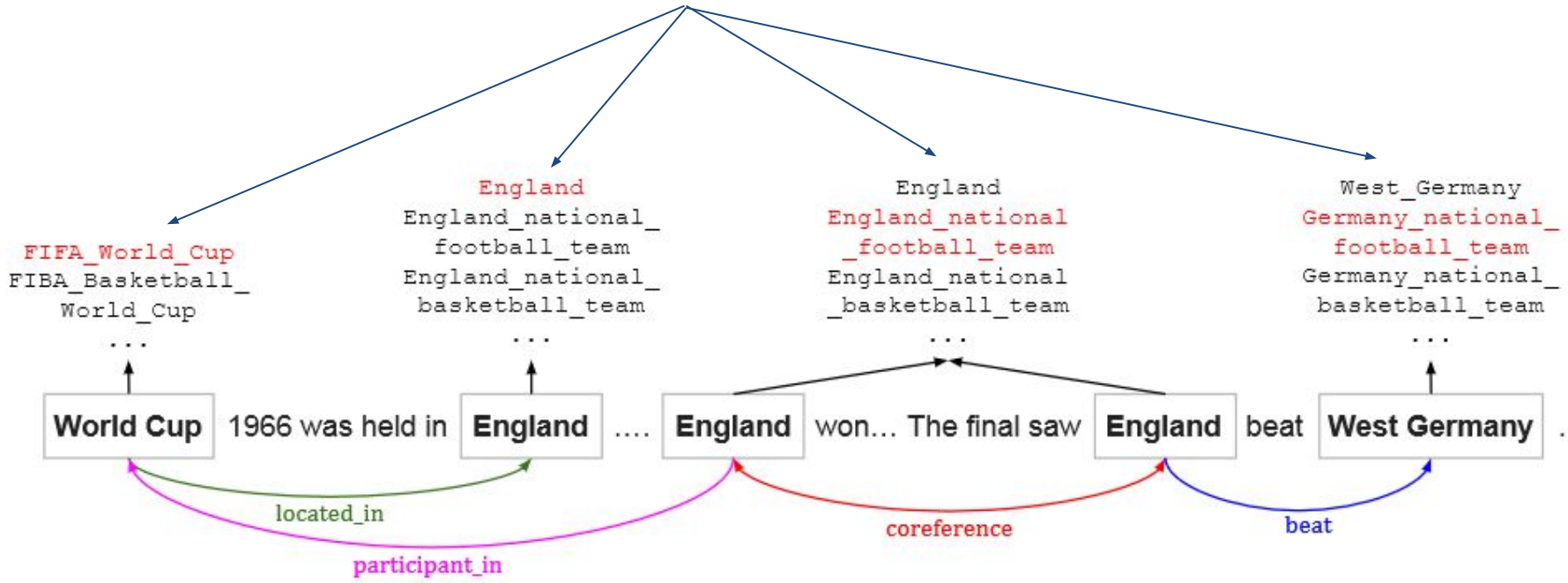
Latent Relations Paper:

Paper Hypothesis:

1. Relations used in Named Entity Linking don't need domain expertise.
2. They can be recognised by using these relations as latent variables and optimising the entity linking model
3. The paper uses Representation learning to learn mention embeddings, contexts and NEL relations.
4. The paper uses the publicly available dataset Aida-ConLL to train the model
5. Candidate Selection is a major part of modelling relations as latent variables

Selection of Candidates:

All Candidates including gold/selected candidates in red



Selection of Candidates:

Local Model:

- It considers only the local context of a mention
- Excludes any inter-dependency among mentions

$$e_i^* = \arg\max_{e_i \in C_i} \Psi(e_i, c_i)$$

Entry i in the KB

Local score of a mention mi with local context ci

Global Model:

- The global model considers both the local score and weighted sum of all the pairwise scores
- The pairwise score are the based on the relations between mentions

Local score

$$E^* = \arg\max_{E \in C_1 \times \ldots \times C_n} \sum_{i=1}^{n} \Psi(e_i, c_i) + \sum_{i \neq j} \Phi(e_i, e_j, D)$$

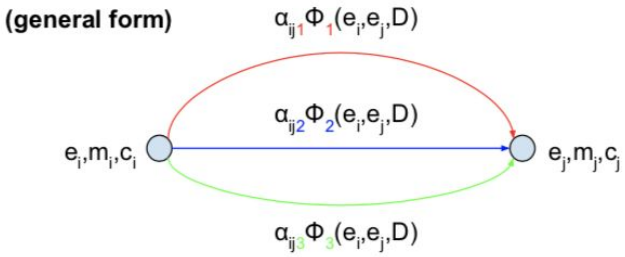
Pairwise scores

Selection of Candidates:

Pairwise scores:

- We assume there are K latent relations between mention mi and mj:
- Each K is assigned to a mention pair (mi,mj) with a weight aijk

$$\Phi(e_i, e_j, D) = \sum_{k=1}^{K} \alpha_{ijk} \Phi_k(e_i, e_j, D)$$

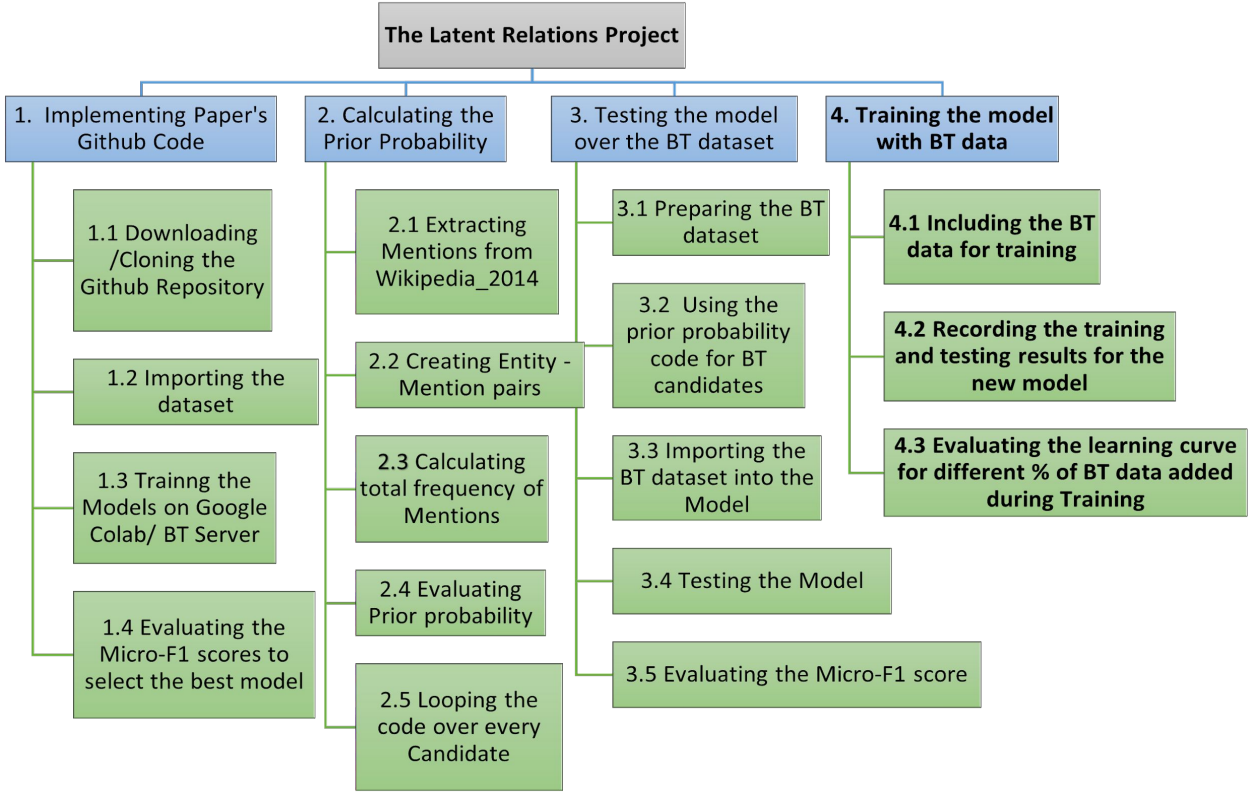


= Weighted sum of relation-specific pairwise scores

$$\Phi_k(e_i, e_j, D) = \mathbf{e}_i^T \mathbf{R}_k \mathbf{e}_j$$

*The red , blue and green line represents the relations between the mention pair (mi,mj) along with their specific weights and pairwise scores

Paper Implementation Workflow



Input Information

Input information	Details	Does BT .json file contain?
Mention	Text linking to entities	✅
Context	Left context, Right context	✅
Candidates	Entity, Prior Probability	❌
Gold	Correct entity	✅
Secondary Context	Sentence(mention in), Mention position	✅

BT dataset processing

- Mention
- Gold
  - Keep entity_id starts with Q
  - Apply wbgetentities API
- Context
  - Extract the right and left context
  - Clean the context
  - Limit at most 100 words

- Secondary Context
  - Split each sentence to words list
  - Store each word lists into Dict[‘Sentences’]
  - Use list.index function to get mention’s position
  - Store each mention position into Dict[‘Mentions’]


```
{
    "version": "1.1.0",
    "data": "On Wednesday, the total number of confirmed deaths linked to
SARS-CoV-2 coronavirus infections surpassed 100,000 in the United States,
Johns Hopkins University data indicated. The coronavirus causes COVID-19,
a sometimes-fatal disease. The milestone came just under a month after
the total number of confirmed infections in the United States surpassed
one million on April 28.\n\nAs of Wednesday, the United States had the
highest number of known infections, accounting for around 30% of world-
wide coronavirus infections with over 1.6 million confirmed cases. The
United States likewise had the highest number of confirmed deaths linked
to the coronavirus, with the next highest country, the United Kingdom,
reporting 37,542 deaths as of Wednesday.\n\nThe milestone came as states
began to relax restrictions put in place during the COVID-19 pandemic.
According to the University of Washington's Institute for Health Metrics
and Evaluation, the coronavirus may cause roughly 32,000 more deaths in
the United States by August 4.\n",
    "attributes": {
        "entities": {
            "type": "list",
            "itemType": "entities",
            "items": [
                {
                    "mentions": [
                        {
                            "startOffset": 44,
                            "endOffset": 50
                        }
                    ],
                    "type": "SYMPTOM",
                    "entityId": "Q4"
                },
```

Candidates Generation

Dataset: textWithAnchorsFromAllWikipedia2014Feb.txt

e.g. Séamus O\'Doherty (11 June 1882 - 23 August 1945) was an Irish republican.\n\nSéamus O\'Doherty was born on 11 June 1882 in Derry.

Mention: Irish republican, Entity_name: Irish republicanism

Mention: Derry, Entity_name: Derry

Calculating the Prior Probability

- Calculating prior probability P(e|m)
  - [mention,entity] Frequency: The times that the mention linked to the entity in Wikipedia KB
  - P(e|m) = [mention,entity] Frequency / All the [mention, entity] pairs Frequency for each mention
  - <u>Frequency.json</u> Format:

    data[mention] = {‘total_freq’: counts, ‘entities’:{‘entity1’:counts,’entity2’:counts}}

e.g. data:

```
{'political philosophy': {'total_freq': 453,
  'entities': {'political philosophy': 448,
   'political philosophy of Muammar Gaddafi': 1,
   'political science': 1,
   'liberalism': 1,
   'Political Philosophy': 2}},
```

Candidates Generation

Dataset: textWithAnchorsFromAllWikipedia2014Feb.txt

Final BT dataset

Input BTdataset Format:
{‘mention’, ‘context’:[leftctx,rightctx], ‘candidates’:[entityname, priorprob], ‘gold’, ‘conll_doc’:{ ‘sentences’: splitted lists of sentences in the article, ‘mentions’: the position of all the mentions in the list}, ‘conll_m’: the position of each mention in the list}

e.g. {"1.json": [{"mention": "United States", "context": ["A ", " federal air marshal shot dead on Wednesday an American Airlines passenger named Rigoberto Alpizar on American Airlines Flight 924 a Boeing 757 at Miami International Airport in Miami Florida USA …. three shots"], "gold": ["United States of America", 1e-05, -1], "candidates": [["United States", "0.944"], ["United States men's national soccer team", "0.009"], ["United States of America", "0.007"], ["United States women's national soccer team", "0.004"], ["United States national rugby union team", "0.002"], ["United States men's national ice hockey team", "0.002"], ["Secondary education in the United States", "0.002"], ["United States national rugby league team", "0.001"], ["United States men's national basketball team", "0.001"], ["United States national cricket team", "0.001"], ….."conll_doc": {"sentences": [["A", "United", "States", "federal", "air", "marshal", "shot", "dead", "on", "Wednesday", "an", "American", "Airlines", "passenger", "named", "Rigoberto", "Alpizar", "on", "American", "Airlines", "Flight", "924", "a", "Boeing", "757", "at", "Miami", "International", "Airport", "in", "Miami", "Florida", "USA"], ["The", "44yearold", "passenger", "ran", "out", "of", "the", "door", "of", "the", "airplane", "after", "he", "reboarded", "the", "plane", "following", "a", "customs", "check", "in", "Miami"], ["He", "was", "intercepted", "by", "the", "marshals", "before", "reaching", "the", "jetway", "and", "told", "to", "get", "on", "the", "ground"], ["According", "to", "Air", "Marshal", "Service", "spokesman", "Dave", "Adams", "the", "passenger", "did", "so", "but", "then", "reached", "for", "a", "bag", "at", "which", "point", "a", "marshal", "fired", "two", "or", "three", "shots", "and", "killed", "the", "passenger"], ["The", "passengers", "recall", "that", "they", "heard", "up", "to", "six", "shots"],.....], "mentions": [{"sent_id": 0, "start": 1, "end": 2}, {"sent_id": 0, "start": 11, "end": 12}, {"sent_id": 0, "start": 15, "end": 16}, {"sent_id": 0, "start": 11, "end": 12}, {"sent_id": 0, "start": 23, "end": 24}, {"sent_id": 0, "start": 26, "end": 28}, …….]}, "conll_m": {"sent_id": 0, "start": 1, "end": 2}},

Training and testing the model

Training data: AIDA dataset (953docs)

Testing dataset:

- ❑ AIDA-A
- ❑ AIDA-B
- ❑ MSNBC
- ❑ AQUAINT
- ❑ ACE2004
- ❑ CLUEWEB
- ❑ BT dataset(70docs)

BT Problem:

- New - context
- Small dataset

```
-----------------------------------------------
277
recall 0.9730745147150908
aida-A #dev docs 218
108
114
recall 0.9828316610925306
aida-B #dev docs 232
recall 0.9847560975609756
msnbc #dev docs 20
recall 0.9436038514442916
aquaint #dev docs 50
recall 0.9066147859922179
ace2004 #dev docs 35
recall 0.9169804554419939
clueweb #dev docs 320
recall 0.923418095801301
wikipedia #dev docs 318
130
126
recall 0.7347560975609756
btdata #dev docs 62
aida-A micro F1: 0.9214069512576976
aida-B micro F1: 0.9337792642140468
msnbc micro F1: 0.9349655700076511
aquaint micro F1: 0.8713286713286713
ace2004 micro F1: 0.8812877263581488
clueweb micro F1: 0.768505210204815
wikipedia micro F1: 0.7776052067154796
btdata micro F1: 0.3554878048780487
```

Training with BT dataset

Final BT dataset: 70docs (60docs for training, 10docs for testing)

Training dataset:

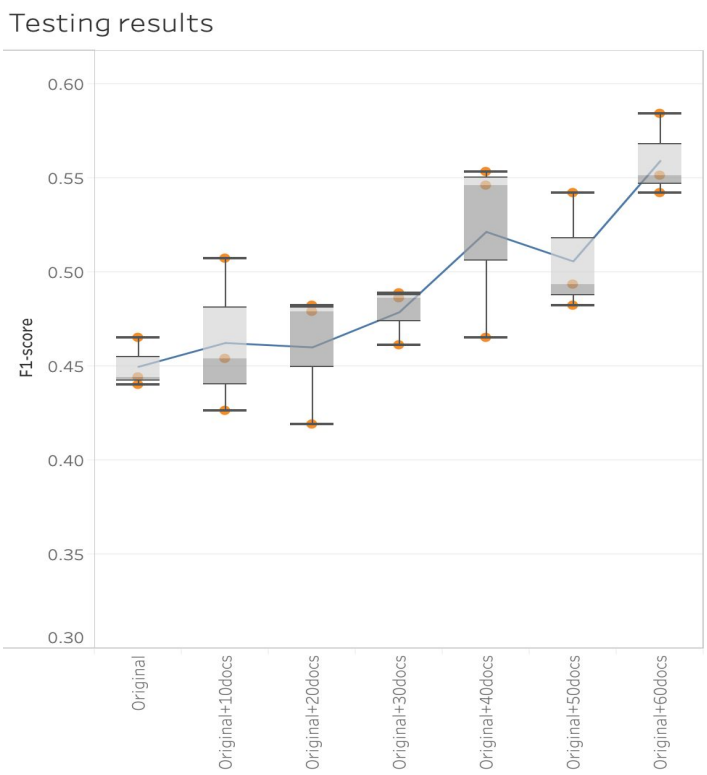
- Original AIDA dataset (953docs)
- Original AIDA + BT 10docs
- Original AIDA + BT 20docs
- Original AIDA + BT 30docs
- Original AIDA + BT 40docs
- Original AIDA + BT 50docs
- Original AIDA + BT 60docs

Testing the Model with BT dataset

Testing results:

Training docs	Original	Original+ 10docs	Original+ 20docs	Original+ 30docs	Original+ 40docs	Original+ 50docs	Original+ 60docs
Testing1 Results	0.444	0.426	0.479	0.489	0.546	0.482	0.584
Testing 2 Results	0.440	0.507	0.482	0.461	0.553	0.542	0.528
Testing 3 Results	0.465	0.454	0.419	0.486	0.465	0.493	0.542
Avg Testing Results	0.449	0.462	0.460	0.479	0.521	0.506	0.551

Testing Results



- Better performance
- Large Intervals

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

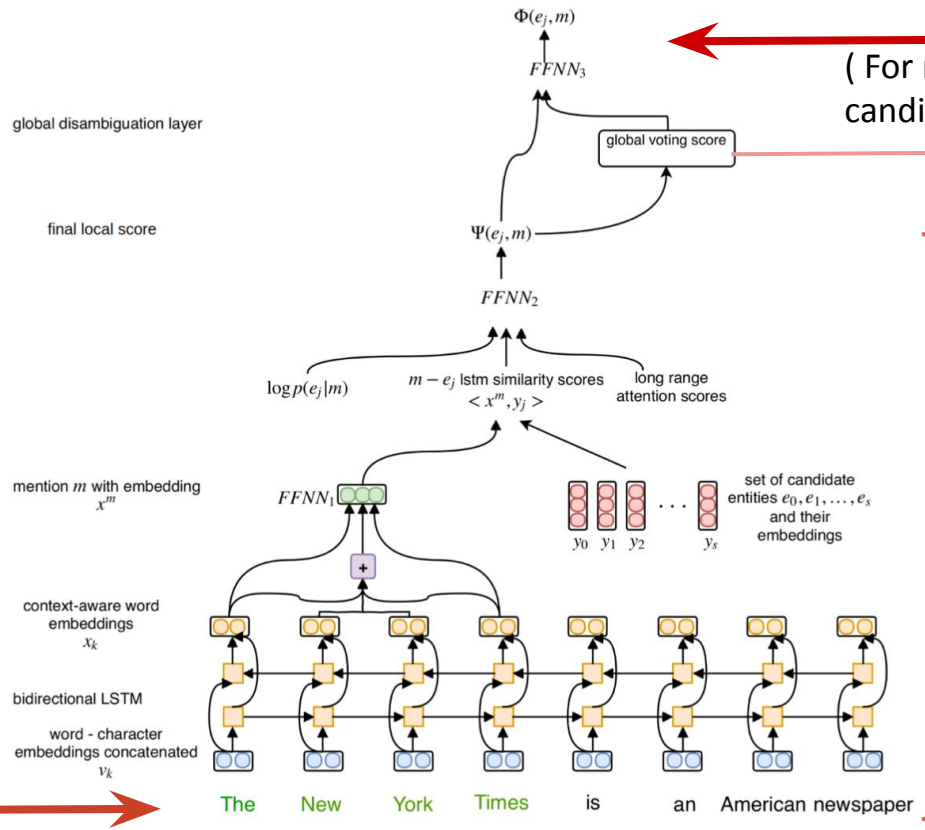
Advantage of the End2End Model

	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 Disambiguation	W Worcester W Massachusetts	W Worcester_Polytechnic_Institute W Massachusetts

Overview of the End2End Paper

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

local score



local score

try different spans--mention

Input: the Whole Text

End2End Paper-- Candidate Selection

- Mention--Entity Candidate pair:
  e.g:
  Discovery

  Discovery Channel--basic cable and satellite television channel
  Star Trek: Discovery--American television series
  Discovery:--space shuttle orbiter
  Discovery--2001 album by Daft Punk

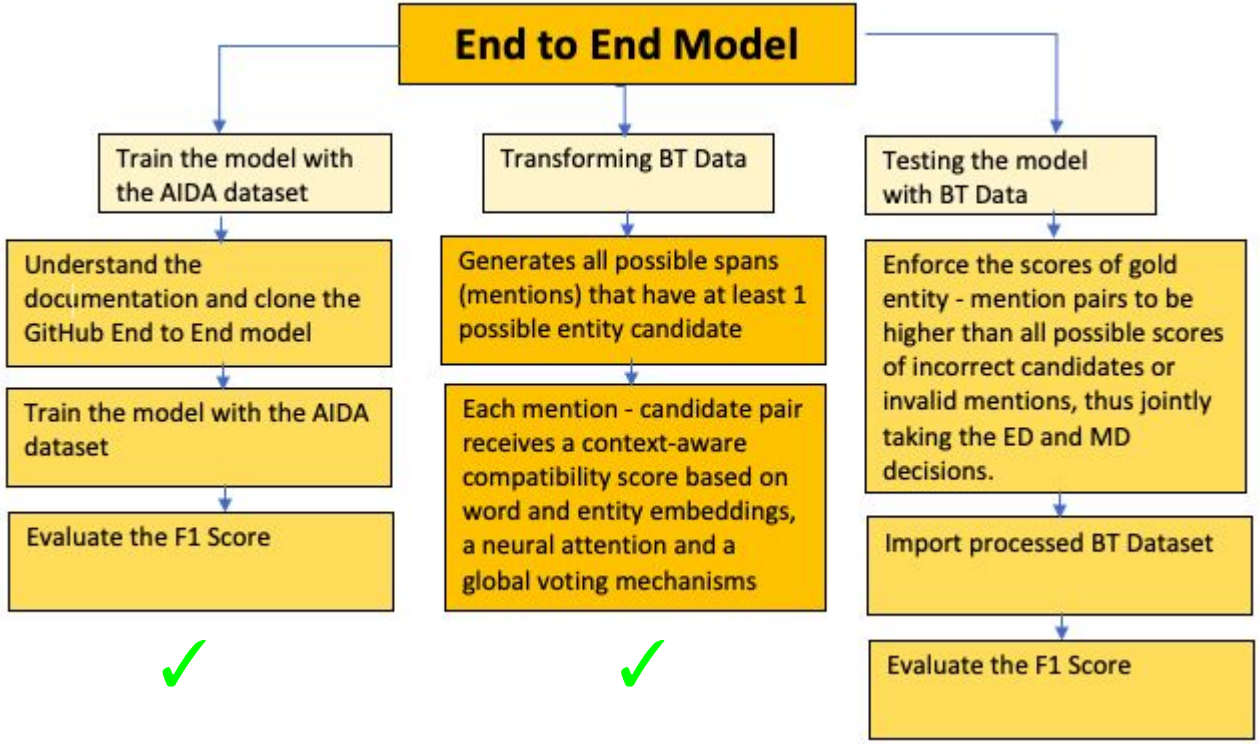
- How to get candidate-- Prior Possibility
  - The ratio of the times the pair appears to the times the mention appears in KB
  - KB: a lot of text sample--Wikipedia2014
  - 30 candidates

- 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 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

```
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
```

Wikipedia

```
-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
```

AIDA

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_deaths
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

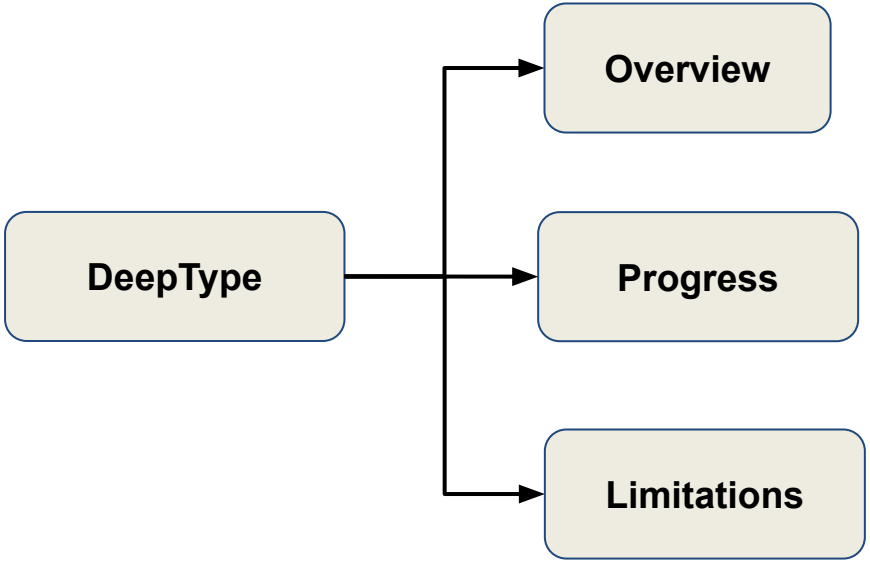
- End2End: if only the word disambiguation process, the result cannot take much of an advantage of this model.
- End2End gave reasonable accuracy on the AIDA testing set because the number of mentions is so big. Since BT's data is lower we expect this number to be lower as well
- We learned:
  - Entity 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

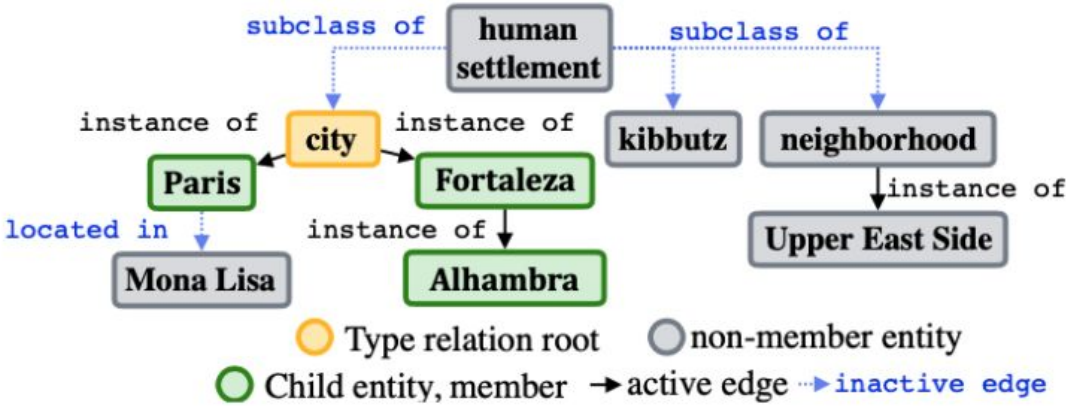
DeepType Model



Deep Type: Overview

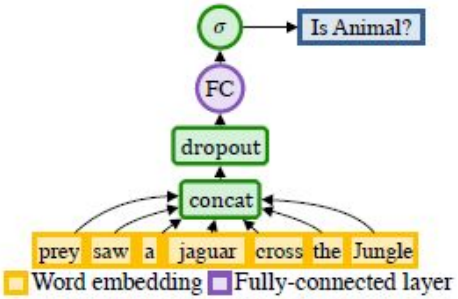
- Two-step process:
  - Construct a Type System
  - Build a Type Classifier

- Type System:

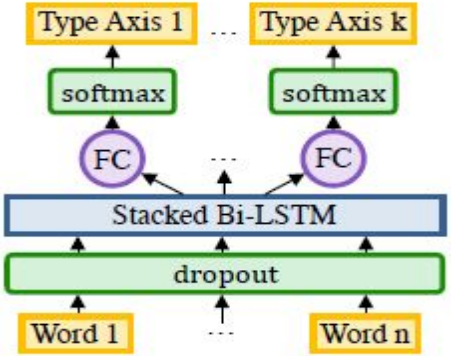


Deep Type: Overview

- Learnability:



- Type Classifier:



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 $p < 0.01$.

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

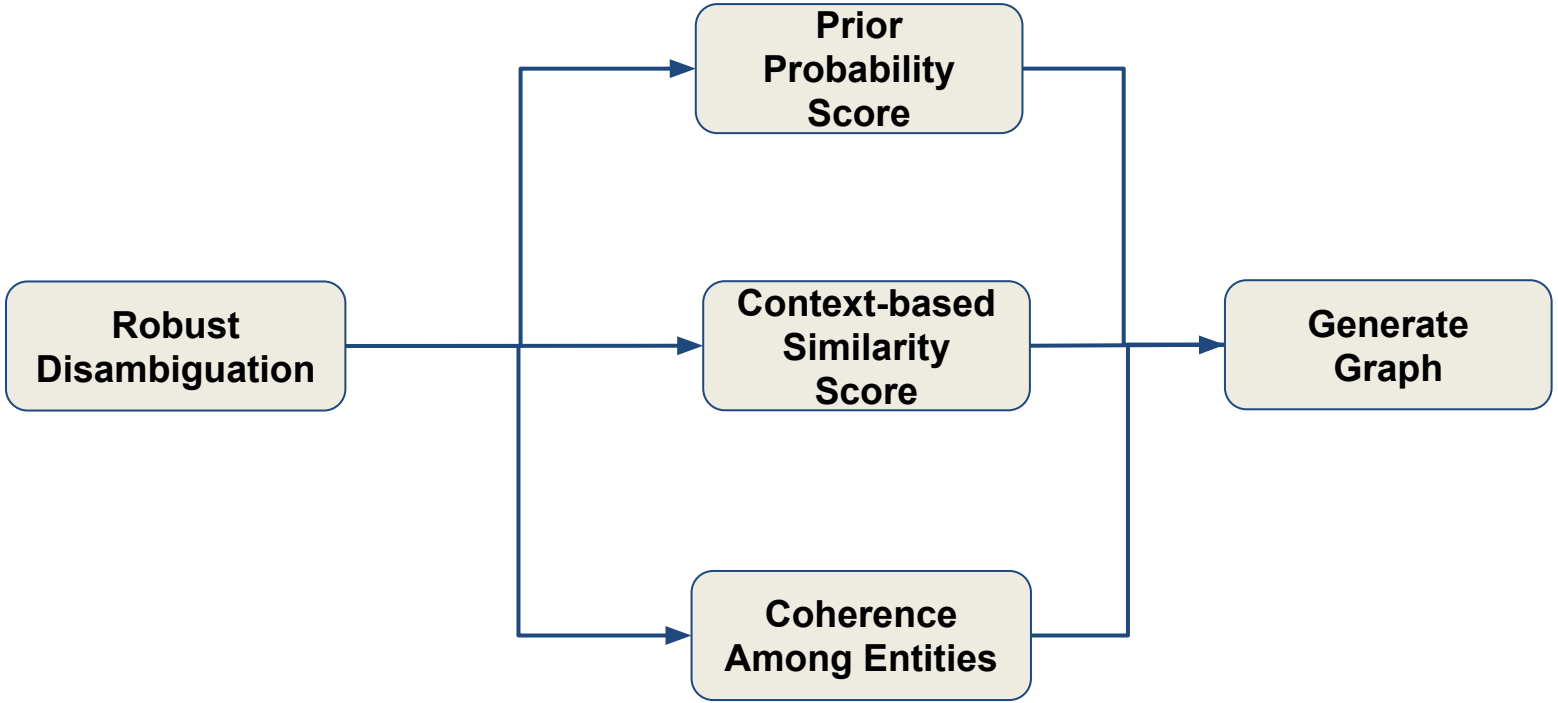
- **Lesson Learnt:**
  - Assumes strong coherence among entities
  - Entities can be easily disambiguated using Type System graph
  - More accuracy if learnability is more for Type System knowledge base

Project References

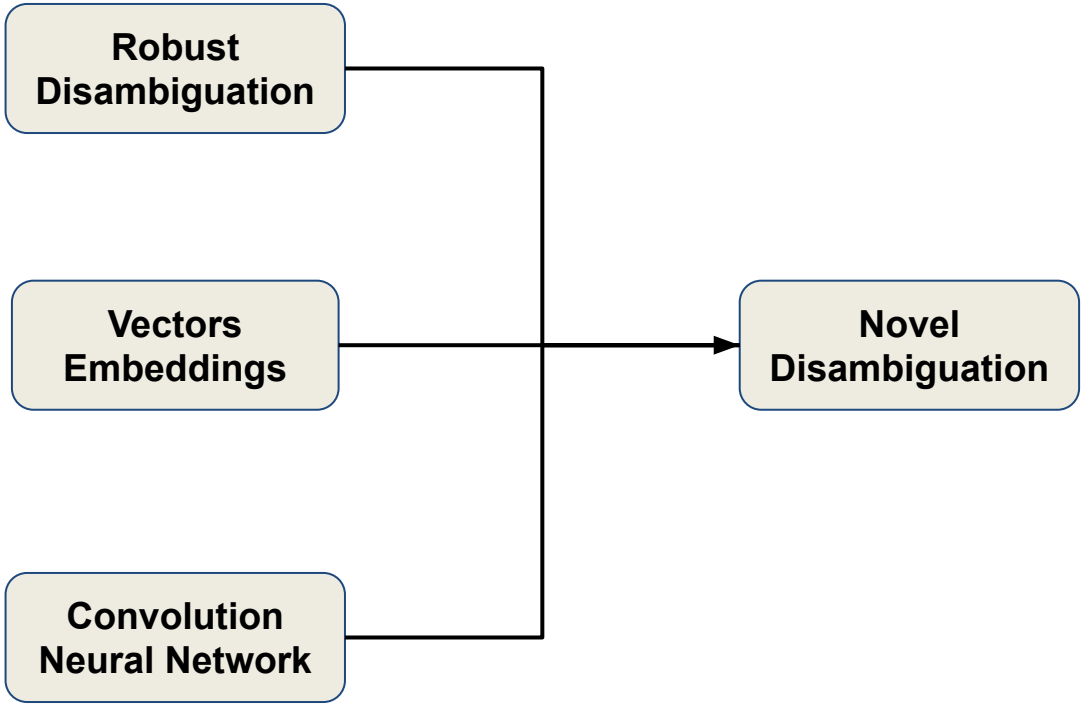
- **DeepType: Multilingual Entity Linking by Neural Type System Evolution**
  https://arxiv.org/pdf/1802.01021.pdf
- **DeepType Source Code**: https://github.com/openai/deeptype

Novel Approach with Robust Disambiguation

Robust Disambiguation Approach:

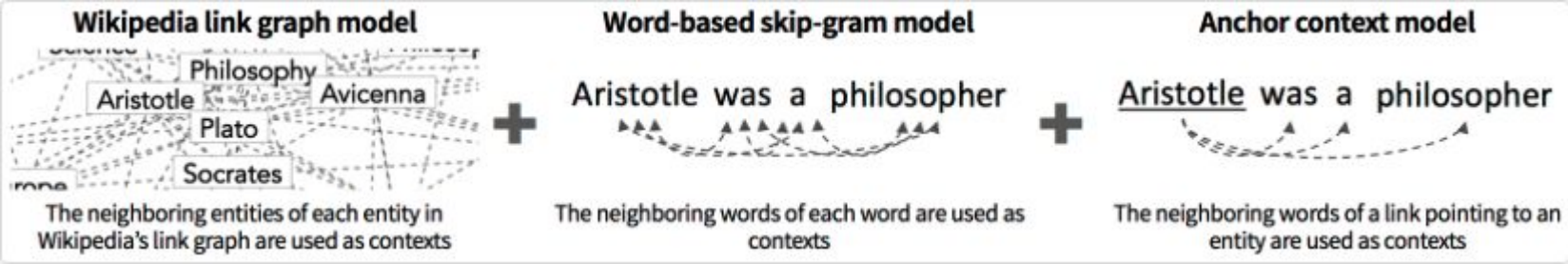


Novel Disambiguation Approach:

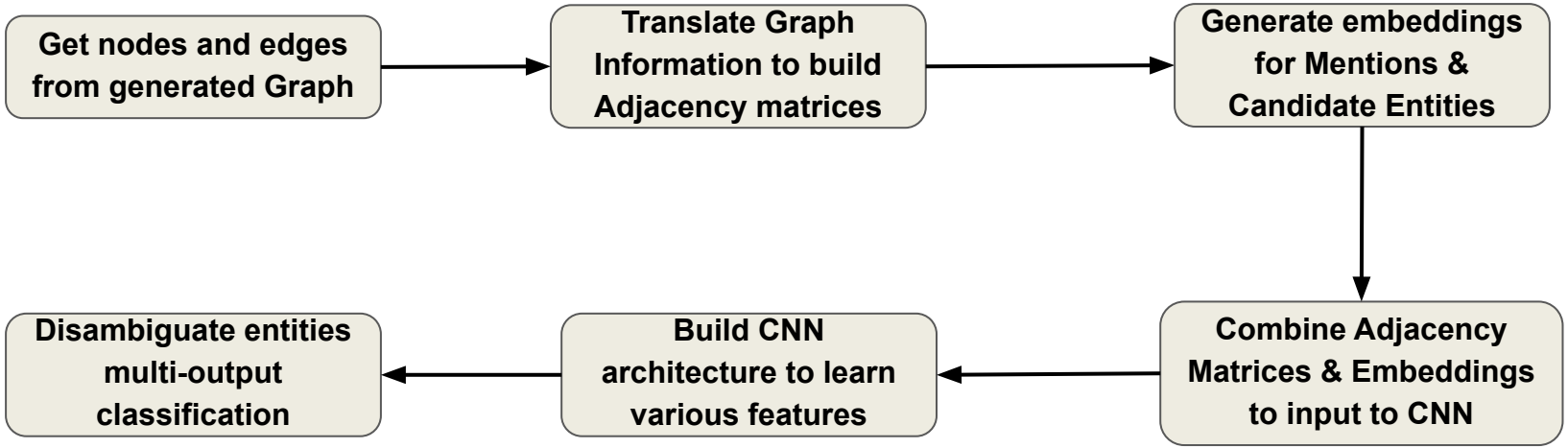


Novel Model: Mention and Entity Embedding

- **Wikipedia2Vec Embedding:**
  - Mention -> Word2Vec Embedding
  - Entity -> Wikipedia2Vec Embedding



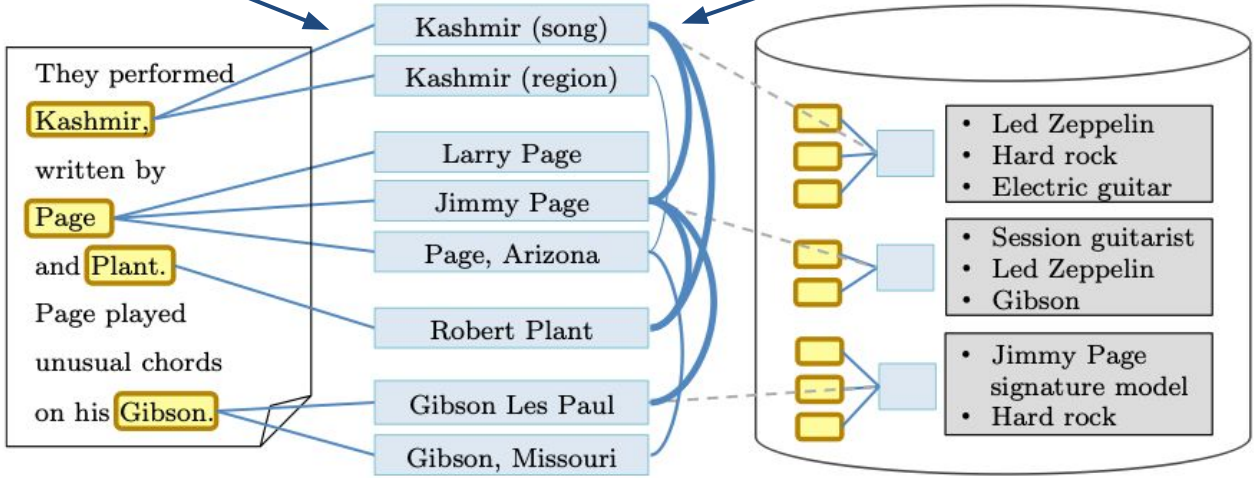
Novel Model:


Get nodes and edges from generated Graph
Translate Graph Information to build Adjacency matrices
Generate embeddings for Mentions & Candidate Entities
Combine Adjacency Matrices & Embeddings to input to CNN
Build CNN architecture to learn various features
Disambiguate entities multi-output classification

Robust Disambiguation: Graph Network

Mention-Entity Links: Based on Prior-Popularity score & Context Similarity score

Entity-Entity Links: Based on Entities Coherence score



Novel Model: Adjacency Matrix

Mention-Entity Adjacency Matrix

	Kashmir (song)	Kashmir (region)	Larry Page	Jimmy Page	Page, Arizona	Robert Plant	Gibson Les Paul	Gibson, Missouri
Kashmir	1	1	0	0	0	0	0	0
Page	0	0	1	1	1	0	0	0
Plant	0	0	0	0	0	1	0	0
Gibson	0	0	0	0	0	0	1	1

Novel Model: Adjacency Matrix

Entity-Entity Adjacency Matrix

	Kashmir (song)	Kashmir (region)	Larry Page	Jimmy Page	Page, Arizona	Robert Plant	Gibson Les Paul	Gibson, Missouri
Kashmir (song)	1	0	0	1	0	1	0	0
Kashmir (region)	0	1	0	0	1	0	0	0
Larry Page	0	0	1	0	0	0	0	0
Jimmy Page	1	0	0	1	0	0	1	0
Page, Arizona	0	1	0	0	1	0	0	1
Robert Plant	1	0	0	0	0	1	0	0
Gibson Les Paul	0	0	0	1	0	0	1	0
Gibson, Missouri	0	0	0	0	1	0	0	1

Novel Model: Graph Generation

Done:

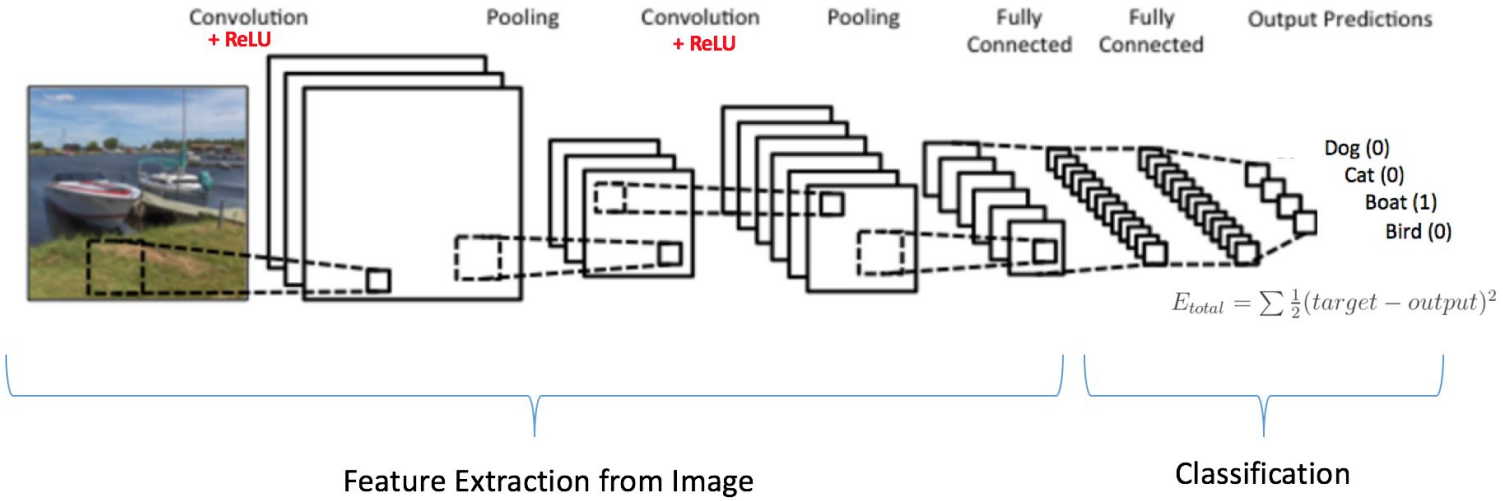
- Use Robust Disambiguation technique code to produce graph https://github.com/codepie/aida
- Installed PostgreSQL 9.6.0
- Downloaded Yago dumps
- Imported Yago dumps in PostgreSQL
- Extract Mention Nodes, Entity Nodes and edge information from the generated graph

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 = {}
```

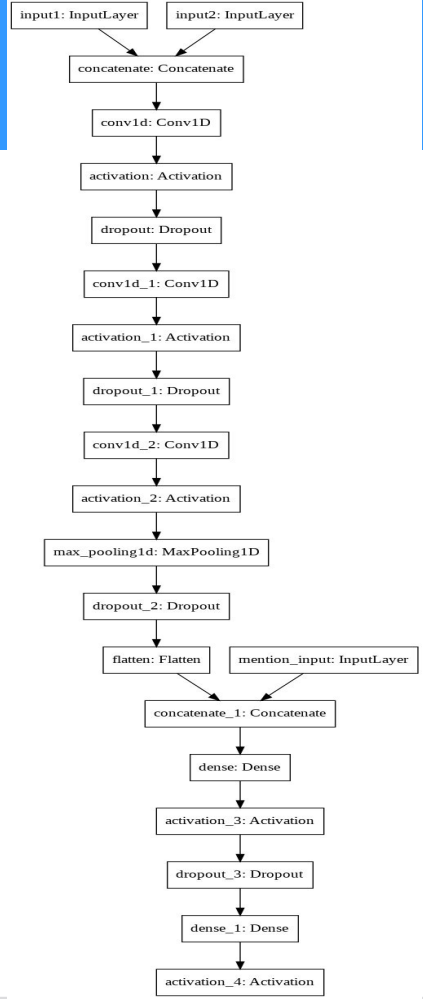
Convolutional Neural Network

- **Convolutional Neural Network (ConvNet or CNN)**
  - Deep Learning Neural Network
  - Generally applied on Images for identifying patterns
  - Shared weights: learnt through training on large dataset



Novel Model: CNN Architecture

- **Novel Model Input:**
  - Combine: Adjacency Matrices & Embedding Matrices
  - Input to CNN model
  - Input particular mention Embedding again
  - Perform Disambiguation



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

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

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

Any Questions?