Graduate Qualifying Project
(DS 598)

Master of Science in Data Science Capstone

Entity Linking Prediction

Sponsors

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

Introduction

Basis Technology (BT) is a company that aims to create a safer environment and more productiveness in the world by building Artificial Intelligence solutions for analyzing text, connecting data, and discovering digital evidence. For over twenty years, Basis Tech has provided solutions for businesses and the government to overcome some of the biggest challenges such as verifying identity, understanding customers, predicting global events, and even uncovering crime. Among the analytical software they developed include Rosette, a text analytics machine learning and deep neural networks hybrid to extract meaningful information from unstructured data (Basis Tech, 2020).

Unstructured data does not have an apt data model structure or a pre-defined organization method. They cannot be stored in a traditional database. Therefore, it is more difficult to analyze and not easily searchable. Data analytics tools such as Rosette aim to ease this problem and produce more valuable insights from the data. Rosette overall is an adaptable platform for text analytics & discovery. It offers various capabilities to efficiently analyze documents in more than 30 languages. Tasks supported by Rosette include chat translation, document categorization, parts of speech tagging, entity extraction, entity linking, and more.

Project Summary

In the project affiliated with Worcester Polytechnic Institute (WPI) and BT called Graduate Qualifying Project (GQP), the focus will be on the entity linking aspect of Rosette. The Named-Entity Linking task aims to extract all named entities (ie. person, organization, location, etc) from a text document, called Entity Extraction/Entity Detection/Named-Entity Recognition, and link the identified mentions to an entity record in WikiData database.

It is important to link the entity mentioned in the document to one and only one record of the WikiData. However, it poses a difficulty when more than one record in the WikiData database has the same name. For example: “Washington” in a text document can be referred to as either a US President George Washington or the US State Washington. Thus, an aspect of entity linking is to disambiguate the mention based on the entire text reference and then link it to a particular record in WikiData database.

Although BT has Rosette to efficiently perform entity detection and entity linking to the database, the goal is to improve and optimize the current procedure. This will be achievable by performing entity linking along with the utilization of entity disambiguation (ED) to achieve the best possible F1 or accuracy score.

DataSet

- AIDA CoNLL-YAGO dataset
    - This is a public dataset offered by NIST (National Institute of Standards and Technology)
    - This dataset is to be used during the current stage of the project to implement various algorithms and compare their accuracy.
- Basis Technology Dataset
    - Internal data collected and prepared by BT
    - Already Pre-processed following the pre-processing guidelines
    - Annotations provided by BT using their in-house annotation tool
    - Entity Extraction is followed as per the algorithm already built by BT

Data Pre-processing

Pre-processing of the text document is required to handle inconsistencies in the data and normalize it to a standard format to reduce the problem of potential losses.

Data Cleaning:

We will first focus on cleaning the textual data:

- Convert the text to a standard case (ie. converting to lowercase)
- Tokenize the text
- Lemmatize to find the core word. It includes removing ‘s as well.

Data Pre-processing:

- Extract entities from the text document
- Classify entities among categories like Person, Organization, Location, Time, etc.
    - **AIDA dataset** - Create YAGO2, Freebase and Wikipedia entity annotations of the original dataset by applying the tool from BT
    - **BT dataset** - Annotate using data annotation rules and standards
- Partition into training and testing datasets with 70% training and 30% testing data

Methodology

Our overall project plan is to gather data and build a knowledge base, annotate the data, evaluate the current entity linking algorithms, design and build neural network models for entity linking as well as compare the performance of different neural network algorithms versus the current SOTA algorithms. There are two different approaches called End-to-End and Disambiguation-Only. End-to-End uses Named Entity Recognition and then disambiguates these extracted entities to the correct entry in a given knowledge base such as Wikidata, DBpedia, or YAGO. Disambiguation-Only directly takes gold standard named entities as input and only

disambiguates them to the correct entry in a given knowledge base..There are three different models we are planning to explore that includes End-to-End Entity Linking which utilizes the End-to-End approach, as well as DeepType and Entity Linking via Latent Relations which both use the Disambiguation-Only process.

1. End-to-End Entity Linking Model

End-to-End Entity Linking Model emphasizes the importance of the mutual dependency between Mention Disambiguation (MD) and ED so that errors in one stage could be covered by the next.

The research covers the following process:

1. Generate all possible spans that have at least one possible entity candidate.
2. Mention - Candidate pairs receive a context aware compatibility score based on word and entity embeddings coupled with a neural attention and global voting mechanisms.
3. During training, they enforce the scores of gold entities - mention pairs to be higher than all possible scores of incorrect candidates or invalid mentions, thus jointly taking the ED and MD decisions.

2. DeepType: Multilingual Entity Linking by Neural Type System Evolution

It is a Disambiguation Only type technique which explicitly integrates the symbolic information related to the identified mention by building a type system for them and choosing a type that best represents the entity corresponding to the text.

DeepType system suggests a two-step solution to the problem:

1. Heuristic search for constructing a Type System
2. Fit a classifier system to the type system

3. Improving Entity Linking by Modeling Latent Relations between Mentions

Entity linking systems often exploit relations between mentions in a document to decide if the linking decisions are compatible. This paper innovatively treats relations between mentions as latent variables in Entity Linking.

The research paper focuses on the following points:

1. Many other relations besides coreference should be taken into account
2. Treat relations as latent variables in their neural entity-linking model instead of supervised systems
3. Apply representation learning but considering relation embeddings

Expected Results

The following results are what we expect to produce and present after completion:

- Evaluation for each model including F1 scores, pros and cons, etc.
- Implementation and experiments on novel ideas based on each model
- Evaluation for improved models
- Summary of using neural networks in entity linking jobs
- Suggestions on future work

Related Materials and Information Learned

- End-to-End Neural Entity Linking, https://arxiv.org/pdf/1808.07699.pdf
    - Combine the Mention Detection and Entity Disambiguation in one model
    - Consider both local score and global score, according to which to decide the final entity candidate
- End to End Github Link: https://github.com/dalab/end2end_neural_el
- DeepType: Multilingual Entity Linking by Neural Type System Evolution, https://arxiv.org/pdf/1802.01021.pdf
    - Builds the Type System for each Mention in the text document. Eg. Washington {Person, Politics, Place}
    - Classify to find the best Type System based on the reference in the text
    - Use the identified Type to disambiguate among Entities
- Deeptype Github Link: https://github.com/openai/deeptype
- Latent Entity Modelling 1: https://arxiv.org/pdf/1804.10637.pdf
    - Apply representation algorithm to extract feature
    - Innovatively take into account latent relations between entities
- Latent Entity Modelling 2: https://arxiv.org/pdf/2001.01447.pdf
    - Utilizing the pre-trained BERT model to produce similarity scores between entities and their mention context
- Latent Github Link: https://github.com/izuna385/Entity-Linking-Recent-Trends
- Training Entity Embeddings: https://github.com/dalab/deep-ed/
- Repository to track the progress in Natural Language Processing (NLP), http://nlpprogress.com/english/entity_linking.html
    - Gain insight into the NLP procedures and understand best practices as well as how they are implemented since it is necessary for this project
- Basis Technology Company Info: https://www.basistech.com/about/
    - Knowledge of what the company does and how our project will fit into their business needs to solve a challenge

We extracted our three baseline models from the above papers, and will go through their official implementation codes in Github to help us on our own implementation. Based on these materials, we can adjust these models and come up with new ideas to make a comparison or an improvement in entity linking.