Labeling Wikipedia Links with Wikidata Properties

Hayden Freedman
UC Irvine
San Miguel de Allende,
Guanajuato, Mexico
hfreedma@uci.edu

Qingyu Song
UC Irvine
Wantai Haojinng,
Hebei, China
qingys1@uci.edu

Tarun Sai Ganesh Nerella
UC Irvine
Palo Verde,
Irvine, California
nerellat@uci.edu

Abstract

Both Wikipedia and Wikidata are crowd-sourced, publicly-available knowledge bases; while Wikipedia is widely-used and presents information as human-readable encyclopedia articles, Wikidata is lesser known and structures information as a computationally-processable knowledge graph. This project seeks to use existing NLP techniques and models to systematically improve Wikidata based on content from Wikipedia. We establish a baseline for assigning Wikidata labels between linked pages in Wikipedia, taking advantage of the text surrounding page links in Wikipedia articles as well as the relationships already present in Wikidata. We also provide details of our process for building the training and test datasets, and plan to make both our codebase and the datasets we generated publicly available. Our initial results indicate that it is plausible to generate high-quality statements for Wikidata using our approach; however, as the percentage of statements appropriate for inclusion in Wikidata in our model's output is low, at this point we require a human in the loop to prune low-quality statements before making the contribution to Wikidata.

Introduction

Many rich public bases of knowledge exist online, yet often they are not well-integrated with each other. The Semantic Web is an attempt to standardize structured data in order to enable the integration of various databases and services; however, one of the most popular public knowledge sources, Wikipedia, is text-based and structured as human-readable encyclopedia articles that include a network of links to other pages. On the other hand, Wikidata is a public source of structured knowledge and contains entities that map directly to Wikipedia articles. Due to its graph-like structure, Wikidata can also directly plug into Linked Data initiatives on the Semantic Web. Our effort in this project is an initial step towards automatically building structured relationships between entities in Wikidata based on the text and link structure of Wikipedia. We feel that such an initiative, if successful, could help make the textual information in Wikipedia more accessible by Semantic Web-based applications.

Previous work on extracting relationships between entities from Wikipedia has tended to focus on tasks to parse Wikipedia articles, identify entities using Named Entity Recognition (NER) techniques, and then deduce the relationship between the entities based on probabilistic models. However, this approach fails to take advantage of the parts of Wikipedia that are in fact structured. At the heart of Wikipedia is its network of page links; like the World Wide Web, Wikipedia is greatly enriched by its ability to efficiently route users from page to page via links embedded in article text. Article titles and page links form the basis of an unlabeled, directed graph, where articles are nodes and links are edges between them. This paradigm of using the latent structured knowledge in Wikipedia provides an alternative, and potentially simpler, approach to text-based parsing using NER.

The primary assumption made in order to conduct the analysis described in this paper is that the Wikipedia link network can be used as a source of ground truth for the existence of relationships between pairs of concepts; that is, we assume pairs of linked pages mean that the real-world entities represented by those pages are indeed associated in reality, and that this relationship can be easily described. This assumption is worth examining further; anyone can edit Wikipedia, meaning that at times the existence or non-existence of a link between two pages may be arbitrary, unpredictable, or biased based on the opinions of the editors. However, the assumption gives us a useful philosophical basis on which to work from.

Our work in this paper focuses on dataset creation and defining a new NLP problem, as well as establishing a baseline. We describe our processes for building the training and test datasets, for our initial experimentation with several different model architectures, and for conducting our human evaluation on a subset of our model's predictions on the test dataset. We also present our initial results, including baseline validation accuracy scores and human-evaluated accuracy scores for our model, and discuss where future work based off our initial research may lead.

Related Work

There has been a significant amount of machine learning research using both Wikipedia and Wikidata to automatically enhance semantic knowledge graphs. Heiko Paulheim [5] provides an excellent survey of various knowledge graph refinement techniques using Wikipedia, including efforts to parse text in Wikipedia abstracts or entire Wikipedia articles to determine relations between concepts, extract relationships from Wikipedia tables, and using Wikipedia lists to infer classifications of concepts.

Xi Yang et.al [6] provide a novel way of linking Wikidata relations to plain text, using bag of distribution modelling. For example, mentions in an article such as "is born in", "is the hometown of", and "comes from" can be linked to the Wikidata property "place of birth" (P19).

The paper called OpenTapioca: Lightweight Entity Linking for Wikidata by Antonin Delpeuch [1] says that Named Entity Linking is the task of detecting mentions of entities from a knowledge base in free text. It provides a simple Named Entity Linking system that can be trained from Wikidata that demonstrates the strengths and weaknesses of this data source for this task and provides an easily reproducible baseline to compare other systems.

From the Professor's suggestion, we read the paper BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding by Jacob Devlin [2] which describes a system designed to pre-train deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers. The reason why we read this is that Bidirectional Encoder Representations from Transformers (BERT) can be used to perform sentence classification which may be useful for us since our idea is essentially a classification problem.

The paper Introducing Wikidata to the Linked Data Web by Fredo Erxleben [3] introduces the use of Resource Description Framework (RDF) to connect Wikidata with the Semantic Web. This paper describes how using the common RDF framework can help connect Wikidata data with other public resources. The paper provides a useful explanation of the Wikidata data model. It also proposes methods of extraction using class hierarchies and ontological axioms to create partial exports to RDF.

Overall we have observed a significant amount of past research using Wikipedia and Wikidata as source data for NLP tasks. However, our approach differs in a few key ways that make it novel. Past research has focused on performing NER over Wikipedia text and then extracting the relationship between entities from the result; however, we have not seen efforts to identify the relationship between entities based on the links that are already present between pages in Wikipedia. Furthermore, we did not see that past research has leveraged Wikidata concepts and properties as a source of training data.

We feel that our approach is in some ways more simple than previous approaches, as we are taking advantage of pre-existent components of publicly available content; namely, the network of links within Wikipedia and the ontology and structured knowledge of Wikidata.

Approach

Dataset Creation

A large chunk of our effort for this project has involved researching and developing software tools to extract the necessary data from both Wikipedia and Wikidata. We used a pre-existing library called Graphipedia (https://github.com/mirkonasato/graphipedia) for extracting the network of articles and page links from Wikipedia into a graph database service. Once we had this network available, we were able to query the graph database to gather lists of all Wikipedia pages linked to the page of 3 central domains which we selected based on variety and personal interest: Sustainability, Roman History, and Basketball. One of the authors had previously written code to extract the relevant subgraph of Wikidata based on a provided list of Wikipedia article titles (https://github.com/greenguy33/wikidata-subgraph-builder). We used this tool to extract the relevant subsections of Wikidata based on our Wikipedia domain lists.

We also required the functionality to extract text surrounding relevant links from Wikipedia, as we intended to use the text surrounding the link as the feature input to the machine learning problem. We wrote custom code that parses Wikipedia HTML to extract data about each page link, including the destination page, the sentence containing the link, and the link text itself. In an effort to obtain links that connected entities with a high degree of relevance, we limited our parsing of Wikipedia HTML to just the "abstract" section of each page, meaning the paragraphs coming before the "Table of Contents" box.

Table 1: Example Training Data

	Origin Page	Destination Page	Link Text	Sentence Text	Wikidata Property
Sustainability	Sustainable Development Goals	United Nations	UN	Though the goals are broad and interdependent two years later 6th of July 2017 the SDGs were made more actionable by a UN Resolution adopted by the General Assembly	P170 ("creator")
Roman History	Rome	1960 Summer Olympics	1960 Summer Olympics	The host city for the 1960 Summer Olympics Rome is also the seat of several specialised agencies of the United Nations such as the Food and Agriculture Organization FAO the World Food Programme WFP and the International Fund for Agricultural Development IFAD	P276 ("location")
Basketball	Midnight Basketball	United States	United States	Midnight basketball is an initiative which developed in the 1990s to curb innercity crime in the United States by keeping urban youth off the streets and engaging them with alternatives to drugs and crime	P17 ("country")

Table 2: Basic Data Statistics

	Labeled Rows	Unlabeled Rows	% Labeled	Average Sentence Length	Training Classes
Sustainability	4,034	14,328	28.2	25.4	163
Roman History	12,349	37,698	32.8	24.6	234
Basketball	5,195	8,661	60.0	17.7	93

Table 3: Class Reduction

	Labeled Rows after Class Reduction	% Labeled Rows Removed
Sustainability	2,437	42.8
Roman History	6,276	49.2
Basketball	4,314	17.0

The final step in the data collection process was to perform a Wikidata search for each retrieved data row from Wikipedia. Pairs of entities in our Wikipedia dataset that already had been assigned a relationship in Wikidata became part of our training dataset (labeled data), whereas pairs without a relationship in Wikidata were integrated into the test dataset (unlabeled data). In order to make sure we accounted for the inconsistent use of reciprocal relationships in Wikidata (such as "part of" vs "has part"), we searched for Wikidata relationships in both directions between the pair of Wikipedia entities.

Table 1 shows a sample of labeled and unlabeled data for each of our domains. Table 2 shows the quantity of labeled and unlabeled data for each of our 3 domains, as well as the average sentence length and number of training classes in each domain.

Models

After collecting our datasets, the next step was the model selection. We tried out 3 different models with varying degrees of complexity: a simple Logistic Regression classifier, a classifier using the Long short-term memory (LSTM) architecture [4], and a classifier using the BERT architecture. For each of the models, we used a feature set consisting of label-encoded representations of the Wikipedia origin and destination pages as well as the text of the sentence that includes the relevant page link.

We also trained each of the models using both the full training dataset, and a reduced training dataset which only included the top 10 most commonly ap-

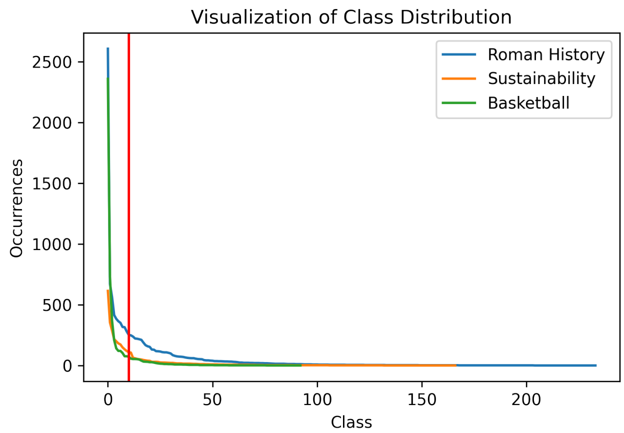

Figure 1: Class Distribution for each Domain

pearing classes for each domain. Table 3 shows how much data was removed from the full to the reduced dataset in each of the three domains, and the plot in Figure 1 visualizes the distribution for each domain and shows how much data was removed from each domain after the reduction (the vertical red line represents the 10-class threshold).

LogReg

Building our Logistic Regression model involved a fairly straightforward modification of the HW1 code to accept our data format. We kept all of the default settings as we wanted to use this model to provide a simple baseline, and in order to avoid over-fitting on the training data. We also modified the code to return accuracy-per-label in addition to just overall accuracy.

Simple LSTM

In our LSTM model, which we previously used to report the baselines in the status report, we built a standard pipeline consisting of a Vocabulary, a Dataset, a Dataloader, and a Model, exposing some useful functions like map_tokens_to_ids, and map_ids_to_tokens, which are used to locate the position of the word in the vocabulary formed. Our WikiDataset class inherits pyTorch's Dataset class. It is used to tensorize the given input sequence by mapping each of its words to their corresponding token IDs. We also have torch's Dataloader which is useful to sample batches of data from the dataset and pass it to the model. We drew inspiration for the model from the neural tagger architecture used in HW3 and adapted it to suit our classification task. Our basic LSTM model consists of an embedding layer, and an LSTM encoding layer, as shown in Figure 2. We extract the encoding of the last word (because this will contain the entire information present in the sentence) and pass it to two linear layers at the top which produce the output vector of size = number of classes.

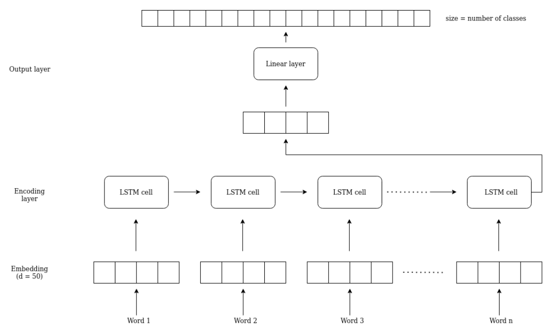

Figure 2: LSTM model

BERT

For the BERT model, we used the pre-trained model "BERT-base-uncased" from the BERT website and token embedding method shown in Figure 3. To handle the variable sizes of the sentence text, we used a padding function to standardize sentence text size. First, we calculated the average sentence length of each data set and set it as the threshold for the padding. If the input vector length is longer than the threshold, the function will cut it into the threshold length, and if the length is shorter than the threshold, the function will fill the input using 0 until it has the length equal to the threshold. In our model, the word vector dimension is 768 and the output dimension is 10 (when using reduced labels), and there are 12 encoder layers that each have 12 Attentions. The training epoch we set is 10 because we found that the models of the 3 datasets can reach the optimal stage by the 10th epoch. Figure 4 shows the high-level structure of the BERT architecture, and Figure 5 shows the configuration that we applied, which was mostly the defaults.

Human Evaluation

Our next step was to perform the human evaluation on the predicted test datasets. Since we had limited human bandwidth, we were unable to review every predicted row from each of the models. We chose to review predictions from the BERT model trained

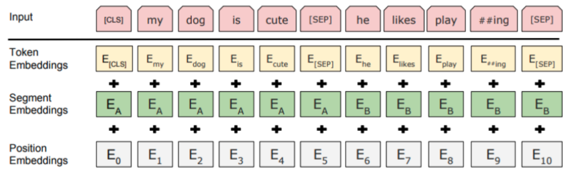

Figure 3: Embeddings

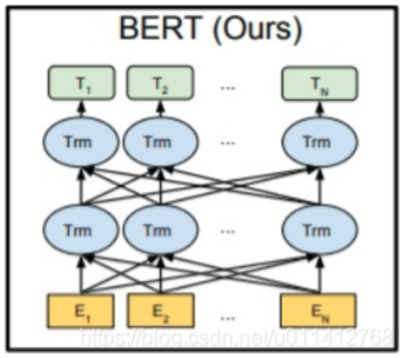

Figure 4: Bert Structure

with the reduced training dataset, as this combination gave us our highest validation accuracy for all three domains. To complete the human evaluation, we wrote code to aggregate all predictions into a single list, then break it into three chunks which were exported as spreadsheets and sent to each member of our team. Each spreadsheet had a column for "agree" or "disagree" which we used to indicate whether each prediction was appropriate for Wikidata. We applied a fairly high standard for marking "agree", conducting research on the concepts presented in each row and the real-life relationship between them when necessary.

In total, we collectively evaluated 950 of the many thousands of predictions generated. A partial example of one of our filled-in human evaluation spreadsheets is shown in Figure 5.

Domain	Index	Page1	Page2	Predicted \	Predicted label	Agree	Disagree
Sustainability	8978	Quaternary extinction	Holocene extinction	P279	subclass of		X
Roman History	32867	Ferrara	Bologna	P131	located in the administrative territorial entity		X
Roman History	33341	The Decline of the West	Classical antiquity	P361	part of		X
Roman History	16918	Bar Montenegro	Italian language	P17	country		X
Roman History	13529	Battle of Zama	Battle of Canusium	P361	part of		X
Basketball	3730	December 2009 in sports	Chile	P17	country		X
Roman History	34170	Sundgau	France	P131	located in the administrative territorial entity	X	
Roman History	7192	List of ancient Illyrian peoples	Religion in ancient Rome	P361	part of		X
Roman History	11822	Hostis humani generis	Ancient Rome	P361	part of		X
Basketball	7931	Gilman School	American football	P118	league		X
Roman History	4525	Lucius Caesetius Flavus	Roman Senate	P39	position held		X
Sustainability	11294	Externality	Economics	P527	has part	X	
Roman History	10639	Victor Emmanuel II of Italy	Capture of Rome	P710	participant	X	

Figure 5: Part of a Human Evaluation Spreadsheet

Results and Analysis

Accuracy on Validation Dataset

We ran each of the three models 6 times; once for each domain with the full training dataset, and once for each of the domains with the reduced training dataset.

Table 4: Overall Validation Accuracies

	Basketball (full)	Basketball (reduced)	Sustainability (full)	Sustainability (reduced)	Roman History (full)	Roman History (reduced)
LSTM	76.0	92.4	28.6	44.4	40.2	64.0
LogReg model	81.8	94.1	38.3	51.6	45.1	74.4
BERT	84.4	96.2	41.3	57.0	55.1	77.8

Table 4 shows the overall validation accuracy for each of these outputs; we can see that in every case the model trained on the reduced label training set performed better than the model trained on the full training set. This improvement was largest on the Roman History domain, which may be explained by the observation that the full training set for this domain had significantly more classes than either of the other two domains.

Overall, our best results for all 3 domains were obtained using the BERT model. We suspect that the complexity of the BERT embedding structure, which analyzes the position of each word in a sentence in relation to all other words, allowed the model to capture the relevant parts of each training sentence better than the other models.

All 3 models performed significantly better on the Basketball domain than on the others. This is unsurprising, as our Basketball training dataset had a significantly higher percentage of labeled training data than the other two domains, as well as the fewest classes in the training data, meaning that a smaller percentage of data was lost when the dataset was reduced. A shorter average sentence length in this domain's dataset may have also contributed to helping the model find signal.

The 10 "Per label accuracy" scores for the reduced datasets are shown in Tables 5, 6, and 7. The top performing labels in Basketball domain across all three models are "school district", "participant in", and "located in the administrative territorial entity". Such high accuracy might be due to the presence of an explicit structure or set phrase in the words surrounding the link which indicates the relationship. The top performing labels in the Sustainability domain are "country of citizenship", "member", "country", and in the Roman history domain are "position held", "located in the administrative territorial entity", and "participant".

Table 5: Basketball: Reduced dataset accuracy per label

Class (relationship)	LSTM	LogReg	BERT
league	75	72.4	100
located in the administrative territorial entity	75	99.8	94.7
participant in	100	100	92.8
follows	40	99.4	0
followed by	88.2	75.4	100
country	60	100	93.6
country of citizenship	82.35	93.75	92.8
position played on team	83.3	9.09	100
school district	100	100	100
sport	100	18.1	99.6

Table 6: Sustainability: Reduced dataset accuracy per label

Class (relationship)	LSTM	LogReg	BERT
located in the administrative territorial entity	12.5	58.75	43.4
contains administrative territorial entity	70.15	43.05	0
country	28.5	62.26	70
country of citizenship	40.9	78.8	88.8
subclass of	6.25	15.15	86.6
part of	45.8	67.8	43.4
member of	54.5	32.8	95.2
shares border with	8	4	27.5
has part	66.6	3.7	5.5
diplomatic relation	11.7	70.2	87.5

Table 7: Roman history: Reduced dataset accuracy per label

Class (relationship)	LSTM	LogReg	BERT
located in the administrative territorial entity	90.9	73.5	69.8
country	11.1	90.8	92.6
place of birth	18.18	96.2	26.4
place of death	34.37	47.2	41.9
country of citizenship	15.3	35.3	75
location	60	14.6	34.3
part of	100	54.8	57.5
position held	76.4	93.4	97.4
shares border with	50	75.2	95.6
participant	54.5	80.9	73

Accuracy from Human Evaluation

Our human evaluation scores were significantly lower than the validation accuracy scores across all three domains, as shown in Table 8. Basketball also performed the best in the human evaluation. Within the basketball domain, two labels in particular achieved very high accuracy: "position played on team / specialty" and "participant in". We hypothesize that the model's success in predicting these relationships is due to the formulaic nature of Wikipedia pages for basketball players. For example, some such pages have a short, clear sentence stating "At a height of {player height}, he/she played at the {position} position." The predictable structure of such sentences apparently gave the model a strong signal to predict player position with a high degree of accuracy.

Table 8: Human Evaluation Results

	Accuracy %	Highest Accuracy per Label %	Label 1 Text	Second Highest Accuracy per Label %	Label 2 Text
Overall	23.2	83.3	position played on team / speciality	81.8	participant in
Sustainability	25.3	54.5	member of	44.4	has part
Roman History	19.1	33.7	part of	29.0	located in the administrative territorial entity
Basketball	36.1	83.3	position played on team / speciality	81.8	participant in

Table 9: Examples of High Quality Outputs

Domain	Origin Page	Relation	Target Page
Roman History	Victor Emmanuel II of Italy	participant	Capture of Rome
Roman History	Saint Domnius	place of birth	Syria
Sustainability	Mennonites	subclass of	Plain people
Sustainability	Sustainable Tourism	part of	Sustainable Development Goals
Basketball	Stéphane Ostrowski	position played on team / specialty	Power forward
Basketball	Étienne Onimus	participant in	1936 Summer Olympics

However, a confounding factor is that the best performing labels of our human evaluation differed significantly from the best performing labels on the validation data. This may indicate significant differences between the validation data and the test data, which will need to be explored further.

Despite our low human evaluation accuracy scores, we still identified more than 200 relationships generated by our model that we felt would be appropriate to contribute to Wikidata, from the small percentage that we looked at. Table 9 shows a small sample of some of these relationships for each domain.

Discussion and Future Work

Our initial results suggest that the NLP technique identified in this paper may differ in efficacy depending on Wikipedia domain. Our model was able to process the Basketball domain data with much higher accuracy than either Sustainability or Roman History. Basketball's superiority was corroborated by both the validation accuracy and the human evaluating results. Future work could involve analyzing the domain datasets to see what aspects of the Basketball dataset caused a stronger signal to be generated. We hypothesize that a smaller number of classes, shorter average sentence length, and a relatively small number of different types of entities with well-defined relationships between them (i.e. players, coaches, teams, positions, etc.) all contributed to better predictions in this domain than the others.

We were very surprised by the huge drop in accuracy from our validation accuracy results to our human evaluation results, as well as the notable differences between the highest performing labels in the validation results and the human evaluation results. However, we believe there is a reasonable explanation for this. Our validation dataset consisted of data that had already been labeled in Wikidata, meaning that for each pair of concepts, someone had previously decided that there was a clearly expressible relationship between the two. However, our test dataset, over which we performed the human evaluation, is composed of pairs of concepts that have not been associated in Wikidata but have a Wikipedia page link connecting them.

We suspect our relatively poor human evaluation accuracy results give credence to the notion that Wikidata simply does not have properties fit to express the relationship between many of the pairs of concepts in our test dataset.

As such, our original assumption that such a relationship exists between any pair of linked Wikipedia pages may not be entirely correct, and should be reevaluated in future iterations of this work.

We feel that the datasets and initial baseline models we have described here leave a lot of areas for potential improvements. One potential future direction is to incorporate the Wikidata ontology into the model. Our current model allows freeform predictions for any combination of pages and Wikidata properties, but Wikidata properties are actually bound by certain constraints limiting what types of entities they can associate. Constraining the model to only make predictions allowed by the ontology could potentially help the predictions be more reasonable. Another potential direction of future work would be to factor the results of our human evaluation back into the model, potentially even down-weighting predictions that we explicitly flagged as incorrect.

In order to make this work more accessible and solicit contributions from others, we plan to publish our datasets and all project code to an open source Github repository. We also plan to contribute the high-quality statements that our process has created to Wikidata, in order to have a small real-world impact from our class project.

References

[1] Antonin Delpeuch. Opentapioca: Lightweight entity linking for wikidata. *arXiv preprint arXiv:1904.09131*, 2019.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[3] Fredo Erxleben, Michael Günther, Markus Krötzsch, Julian Mendez, and Denny Vrandečić. Introducing wikidata to the linked data web. In *International semantic web conference*, pages 50–65. Springer, 2014.

[4] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[5] Heiko Paulheim. Knowledge graph refinement: A survey of approaches and evaluation methods. *Semantic web*, 8(3):489–508, 2017.

[6] Xi Yang, Shiya Ren, Yuan Li, Ke Shen, Zhixing Li, and Guoyin Wang. Relation linking for wikidata using bag of distribution representation. In *National CCF Conference on Natural Language Processing and Chinese Computing*, pages 652–661. Springer, 2017.