Theory Entity Extraction for Social and Behavioral Sciences Papers using Distant Supervision

Xin Wei, Lamia Salsabil, Jian Wu
Old Dominion University
Norfolk, VA, USA
{xwei001,lsals002,j1wu}@odu.edu

ABSTRACT

Theories and models, which are common in scientific papers in almost all domains, usually provide the foundations of theoretical analysis and experiments. Understanding the use of theories and models can shed light on the credibility and reproducibility of research works. Compared with metadata, such as title, author, keywords, etc., theory extraction in scientific literature is rarely explored, especially for social and behavioral science (SBS) domains. One challenge of applying supervised learning methods is the lack of a large number of labeled samples for training. In this paper, we propose an automated framework based on distant supervision that leverages entity mentions from Wikipedia to build a ground truth corpus consisting of more than 4500 automatically annotated sentences containing theory/model mentions. We use this corpus to train models for theory extraction in SBS papers. We compared four deep learning architectures and found the RoBERTa-BiLSTM-CRF is the best one with a precision as high as 89.72%. The model is promising to be conveniently extended to domains other than SBS. The code and data are publicly available at https://github.com/lamps-lab/theory.

CCS CONCEPTS

• **Computing methodologies** → **Information extraction**.

KEYWORDS

NLP, NER, big data, deep learning, distant supervision

ACM Reference Format:
Xin Wei, Lamia Salsabil, Jian Wu. 2022. Theory Entity Extraction for Social and Behavioral Sciences Papers using Distant Supervision. In *DocEng2022: September 20th, 2022 to September 23rd, 2022, Virtual Event (Hosted from San Jose, CA, USA).* ACM, New York, NY, USA, 4 pages.

1 INTRODUCTION

An exponential growth of scientific literature is prominently observed over the past decades [1]. It has become increasingly challenging for researchers to familiarize related works by reading relevant papers of a particular topic in a certain field because of the large number of papers published. Abstracts and high-level key phrases can help researchers to understand the main ideas of a paper [2], but they are often insufficient to assist researchers to delve into details.

Theory and model (hereafter "theory" to refer to theory and model unless otherwise notified) names are ubiquitous in scientific papers, usually used as foundations of further derivations, basis of hypothesis, and justifications of claims. They can assist researchers capture the reasoning process. They can also be incorporated into to the faceted search feature of an academic search engine. Automatic extraction of theories can facilitate building knowledge graphs that connect publications, which further powers graph embedding and novelty analysis. Extracted theory mentions can also be used for literature analysis, such as domain development, innovation composition and other innovation-related topics.

Due to the lack of large-scale training corpus, theory entity extraction has not been extensively explored. The traditional approach to obtain training data is to let humans manually annotate a set of documents. However, annotating theory mentions is time-consuming. In addition, the annotation task requires annotators to have sufficient domain knowledge to understand the context. Crowdsourcing is not an appropriate solution here since annotation of theory entities is constrained by recruiting enough expertise in specific domains from a pool of qualified researchers. To the best of our knowledge, there is no existing labeled data for extracting theory entities in SBS domains.

Distant supervision has been applied in many tasks as a solution to overcome the challenge of learning with relatively small data, e.g., relation classification [3]. We use distant supervision to address the data sparsity problem of theory extraction, which can save human labor on annotating thousands of sentences. With distant supervision, we make use of an already existing database, such as Wikipedia, to collect instances of entity mentions. We then use these instances to automatically generate our training data. The deep learning model trained on the data generalized well and extracted a significant fraction of *new* theory entity mentions.

The contributions of this paper are summarized as follows:

(1) We proposed a framework that extracts theory entity mentions from scientific papers using a distantly supervised method.
(2) We created a new benchmark corpus from SBS papers, which consists of 4534 sentences with 550 unique theory mentions automatically annotated. This new dataset fills the gap in the availability of datasets for theory entity extraction in SBS domains. The data and source code are publicly available.
(3) We performed a comparative study of state-of-the-art (SOTA) Named Entity Recognition (NER) methods and found that

DocEng2022, September 20th, 2022 to September 23rd, 2022, Virtual Event (Hosted from San Jose, CA, USA)

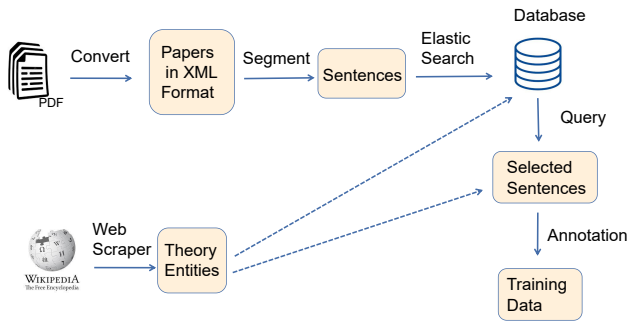

Figure 1: Pipeline for construction ground truth data.

RoBERTa-BiLSTM-CRF was the best architecture for our task, achieving a precision of 89.72%.

2 RELATED WORK

The problem of theory mention extraction can be treated as an NER task, which is fundamental to building knowledge bases and search engine repositories and thus has been extensively studied (see a recent survey [4]). The SOTA models include the Bidirectional Long Short-Term Memory (BiLSTM) model e.g., [5, 6], the Transformer [7] and the Graph Convolutional Network (GCN). However, which one performs the best seems task-dependent.

Distant supervision was introduced to natural language processing (NLP) tasks by Mintz et al. [3], who used a semantic database of relations to obtain the entity pairs in a relation and then used them to match texts and trained a relation classifier. The work by Fries. et al. [8] used distantly supervised NER models for extracting disease and chemical names from biomedical papers. They used a collection of lexicons, ontologies, and optional heuristic rules as knowledge base to build their training corpus. Another work [9] used distant supervision on Aspect Term Extraction and showed that it generated much higher F1 scores than rule-based baseline methods. A more recent work [10] generalized the distant supervision NER models to open domains.

There are few existing works relevant to theory extraction. One early work [11] used heuristic methods to retrieve mathematical theorem statements in the domain of mathematics and physics, aiming at obtaining long text spans instead of theory phrases. Recently, both computer vision and NLP methods are tested in [12] to find theorem-like parts in a paper. A recent work [13] proposed a sequence tagging model comprised of a parallel structure of CNN and BiLSTM layers to extract method and dataset mentions using a corpus of manually annotated training data consisting of 2800 samples in the computer science domain. The first two papers [11, 12] aim at long theorem statements other than theory names. The last paper [13] trained its model particularly for the computer science domain. Therefore, they can not be directly applied to our task.

3 GROUND TRUTH DATA CONSTRUCTION

We use distant supervision to generate our ground truth data automatically and then this data is used to train a supervised model. Examples of ground truth data are shown in Figure 2. The parent sample is obtained by the Defense Advanced Research Projects Agency (DARPA) programme 'Systematizing Confidence in Open Research and Evidence' (SCORE) project [14], containing approximately 30,000 articles published from 2009-2018 in 62 major SBS journals in psychology, economics, politics, management, education, etc. We obtain the text for labeling from a random sample of 2400 SBS papers. The theory names are obtained from Wikipedia as knowledge base (KB). The automated pipeline used to generate ground truth data is shown in Figure 1. The pipeline is composed of the following five modules:

Web Scraping: We select 10 hand-curated Wikipedia webpages containing SBS theory names. Then we utilize a web scraper to obtain those entities. Examples of webpages are shown here[1].

The web scraper is inclusive but not selective. Some of the phrases in the preliminary list such as"working class", "Colony" or "third camp" are not theory mentions, which were excluded by a heuristic filter. Specifically, we keep phrases ending with the following head words, including "theory", "model", "concept", "phenomenon", "effect", "principle", "hypothesis", "bias", and "correlation". The final seed list is comprised of 550 unique theory names.

Obtaining Body Text: GROBID [15] is a machine learning library for extracting and re-structuring raw documents. We adopt GROBID to convert PDF documents into XML format because its performance was shown to be better than many other methods [16]. Sections of papers are parsed and marked in XML files, making it straightforward to obtain body text.

Sentence Segmentation: We use Stanza[2] to segment the body text of papers into sentences. Stanza achieved the best performance among 4 segmentation tools [16]. We extracted in total of 870,000 sentences from the 2400 papers.

Elasticsearch: The sentences are indexed by Elasticsearch. Elasticsearch is an industry-quality search platform based on the Lucene library, providing full-text search with a web interface. It uses BM25 as the default retrieval model. In our case, each sentence is indexed along with paper number, sentence index, and other metadata.

Automatic Annotation: The seed theory mentions obtained in *Web Scraping* are used to query the Elasticsearch index. Finally, we obtained 4534 sentences as the ground truth. The automatic annotation is conducted in the following procedures: First, we substitute certain punctuation marks such as "-", "/", "+", etc. appearing in sentences and theory mentions by spaces. Then we tokenize each sentence by the NLTK tokenizer. Next, we query Elasticsearch with the theory mentions one by one and keep sentences with at least one theory name. Then we identified text spans of theory mentions in the sentence and represent the sentences into BIO (Begin, Inside, Outside) schema.

4 THEORY NAME EXTRACTION MODELS

As claimed in a paper about mathematical theorem extraction [11], sentences containing theory names usually have similar syntactic and semantic features. Empirically, sentences in SBS papers associated with a theory appear to show such similarities. For example, sentences similar to the text spans such as "as attribution theory

[1]https://en.wikipedia.org/wiki/List_of_social_psychology_theories;
https://en.wikipedia.org/wiki/Category:Political_science_theories;
https://en.wikipedia.org/wiki/Category:Statistical_tests;
https://en.wikipedia.org/wiki/Category:Econometric_models
[2]https://stanfordnlp.github.io/stanza/

Due to confirmation bias, individuals do not fully analyze evidence that contradicts their preconceived notions of a current situation.

The conformity of risk preference across methodologies is what I term having no endowment effect for risk.

As attribution theory predicted, individuals exhibited ingroup favoritism when interpreting the cause of climate change.

Figure 2: Sentences containing highlighted theory names in ground truth data.

predicted", and "according to attribution theory" are more likely to be associated with theory mentions. We hypothesize that these syntactic and semantic features can be captured by latent representations output by neural networks and *new* theory names could be identified by the deep neural models.

Deep Neural Network Architectures. We compare four deep neural network architectures, including BiLSTM, BiLSTM-CRF, Transformer, and GCN. We also investigate the performance dependencies on the input language models.

The BiLSTM architecture analyzes the contextual dependency for each token from both forwards and backwards simultaneously, and then assigns each token a label based on probability scores for each tag. This model has shown effectiveness in capturing sequential dependency between tokens within a sentence. BiLSTM can work together with a Conditional Random Field (CRF) layer [17], which labels a token based on its own features, features and labels of nearby tokens [18].

Transformer [19] is widely used in NLP tasks such as machine translation and pre-training language models. The transformer model has also been used in NER tasks and achieved SOTA or sub-SOTA performance, e.g., [7]. A transformer model predicts labels of tokens based on features of neighboring tokens simultaneously using a multi-head attention mechanism.

GCN [20] is a type of CNN that processes graph-like data structures. GCNs have been widely used in computer vision, knowledge graph representation, social networks mining, and NER tasks in which text is represented as graphs. The architecture we used (called GCN-BiLSTM) contains a BiLSTM layer stacked on top of a GCN[3].

Distributed Text Representations. We investigate the performance of models by comparing several representative distributed text representations, including FastText [21], GloVe [22], ELMo [23], BERT [24] (the "bert-base-uncased" version), RoBERTa [25] (the "roberta-base" version), and GPT [26] (the "GPT-1" version).

Experiment Setup. The ground truth samples were split into training, validation, and testing sets, consisting of 3934, 200, and 400 sentences, respectively. The experiments were conducted on a rack server with 24 Intel Xeon Silver Cores, 380GB RAM, and 4 Nvidia GTX 2080 Ti GPUs.

5 EVALUATION

Table 1 summarizes the performance of each architecture for extracting the theory mentions using different word embeddings. The BiLSTM-CRF architecture with the RoBERTa embedding achieved the highest performance with an F1=77.21% and a precision=89.72%. All models achieve higher precision than recall by up to 25%.

[3]Implemented at https://github.com/graph4ai/graph4nlp

	Precision	Recall	F1
BiLSTM-CRF			
FastText	68.05	50.12	57.72
Glove	81.99	60.00	69.29
Elmo	84.98	62.59	72.09
GPT	88.20	63.29	73.70
BERT	81.54	**69.65**	75.13
RoBERTa	**89.72**	67.76	**77.21**
BiLSTM			
FastText	47.71	24.47	32.35
Glove	49.54	25.18	33.39
Elmo	59.65	35.88	44.81
BERT	63.66	61.41	62.51
GPT	68.70	60.94	64.59
RoBERTa	68.33	64.47	66.34
Transformer			
RoBERTa	69.43	66.98	68.18
BERT	74.67	66.59	70.39
GCN	87.12	54.12	66.76
MDER [13]	76.23	64.20	69.63
Wu et al. [27]	60.0	48.00	53.00

Table 1: A comparison of neural network architectures with various text embeddings. The highest values of each metric are indicated in bold. The results from MDER and Wu et al. were directly quoted from their papers.

Under the BiLSTM-CRF architecture, the transformer-based word embedding models such as BERT, RoBERTa, and GPT achieves superior performance compared with other word embedding models. A similar pattern is seen for BiLSTM model. The CRF layer improves performance significantly when added to the BiLSTM layer. The GCN-BiLSTM architecture shows marginal improvement compared with BiLSTM.

Due to lack of existing work on theory extraction, we compare our results with similar works to demonstrate the effectiveness of our proposed method. In Wu et al. [27], the authors trained several sequence tagging models to extract domain knowledge entities and achieved an F1 of 53% when working together with a heuristic classifier. In Hou et al. [13], the authors manually annotated about 5400 sentences containing both method and dataset mentions. They achieved an overall F1=69.63% using BiLSTM-CRF. Six graduate students were hired to annotate all sentences.

6 RESULT DISCUSSION

The distantly supervised method significantly reduces the amount of time used for building the annotated data compared with human annotation. It takes less than half an hour to compile a list of sentences by going through 870,000 sentences and checking whether they contain any of the 550 theory phrases. The overhead to index all sentences is negligibly small. The most time-consuming part is "Automatic Annotation", which takes two hours to annotate the 4534 sentences, but it still takes much less time than human annotation, which may take days to weeks. The distant supervision approach

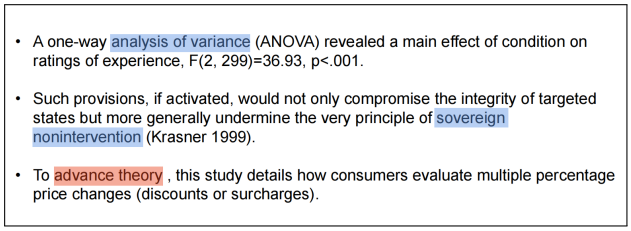

Figure 3: Examples of theory names (highlighted) extracted.

also allows us to transfer the model to other domains, given a new list of theory mentions. Human supervision only occurs when compiling the seed theory names by first selecting Wikipedia webpages and excluding unqualified terms using a heuristic filter.

The limitation of the method was a potentially reduced coverage of theory entities due to limited coverage of Wikipedia pages. It is possible that there are unlabeled entities in the training sentences, which would affect the recall of the model when evaluated on a human-annotated dataset. However, we hypothesize that the deep learning architecture we employed was able to capture the latent representations and language patterns and extract *new* theory mentions that do not exist in the training corpus. To test the hypothesis, we randomly sampled 428 sentences from 10 SBS papers that were not used for generating the training data and extracted theory names using our best model. All theory names extracted from these sentences were *new*. In particular, about 42% contain head words that were *not* in the heuristic filter. The test indicates that the deep learning model was able to generalize to unseen theory names. Figure 3 illustrates the results of extraction, among which red color indicates possible errors. We can combine distantly supervised method with semi-supervised models to further improve performance.

7 CONCLUSION

We proposed a trainable framework that extracts theory and model mentions from scientific papers using distant supervision. The framework automatically generates annotated text based on seed entity names adopted from Wikipedia, which mitigates the data scarcity problem in neural NER models. We have created a new benchmark corpus consisting of 4534 annotated sentences from papers in SBS domains. This dataset can be used for future models on theory extraction. We compared several NER neural architectures and investigated their dependency on pre-trained language models. The empirical results indicated that the RoBERTa-BiLSTM-CRF architecture achieved the best performance with an F1 score of 77.21% and a precision of 89.72%. Moreover, the automatic ground truth data generating framework can be potentially transferable to other domains with sparse annotated data.

ACKNOWLEDGMENTS

This work was partially supported by the Defense Advanced Research Projects Agency (DARPA) under cooperative agreement No. W911NF-19-2-0272. The content of the information does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

REFERENCES

[1] Lutz Bornmann and Rüdiger Mutz. Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. *Journal of the Association for Information Science and Technology*, 66(11):2215–2222, 2015.
[2] Johannes Knittel, Steffen Koch, and Thomas Ertl. ELSKE: efficient large-scale keyphrase extraction. In Patrick Healy, Mihai Bilauca, and Alexandra Bonnici, editors, *ACM DocEng '21*, pages 9:1–9:4, 2021. doi: 10.1145/3469096.3474930.
[3] Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. Distant supervision for relation extraction without labeled data. In *Proceedings of ACL and AFNLP*, 2009.
[4] Jing Li, Aixin Sun, Jianglei Han, and Chenliang Li. A survey on deep learning for named entity recognition. *IEEE TKDE*, 2020.
[5] Liyuan Liu, Jingbo Shang, Xiang Ren, Frank Xu, Huan Gui, Jian Peng, and Jiawei Han. Empower sequence labeling with task-aware neural language model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.
[6] Xuezhe Ma and Eduard Hovy. End-to-end sequence labeling via bi-directional lstm-cnns-crf. *arXiv preprint arXiv:1603.01354*, 2016.
[7] Qipeng Guo, Xipeng Qiu, Pengfei Liu, Yunfan Shao, Xiangyang Xue, and Zheng Zhang. Star-transformer. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of NAACL-HLT*, 2019. doi: 10.18653/v1/n19-1133.
[8] Jason Fries, Sen Wu, Alex Ratner, and Christopher Ré. Swellshark: A generative model for biomedical named entity recognition without labeled data. *arXiv preprint arXiv:1704.06360*, 2017.
[9] Athanasios Giannakopoulos, Claudiu Musat, Andreea Hossmann, and Michael Baeriswyl. Unsupervised aspect term extraction with B-LSTM & CRF using automatically labelled datasets. In Alexandra Balahur, Saif M. Mohammad, and Erik van der Goot, editors, *Proceedings of WASSA@EMNLP*, 2017.
[10] Chen Liang, Yue Yu, Haoming Jiang, Siawpeng Er, Ruijia Wang, Tuo Zhao, and Chao Zhang. Bond: Bert-assisted open-domain named entity recognition with distant supervision. In *Proceedings of ACM SIGKDD*, 2020.
[11] Yihe Dong. Nlp and large-scale information retrieval on mathematical texts. In *International Congress on Mathematical Software*, pages 156–164. Springer, 2018.
[12] Shrey Mishra, Lucas Pluvinage, and Pierre Senellart. Towards extraction of theorems and proofs in scholarly articles. In *Proceedings of ACM DocEng*, 2021.
[13] Linlin Hou, Ji Zhang, Ou Wu, Ting Yu, Zhen Wang, Zhao Li, Jianliang Gao, Yingchun Ye, and Rujing Yao. Method and dataset entity mining in scientific literature: A CNN+ BiLSTM model with self-attention. *Knowledge-Based Systems*, 2021.
[14] Nazanin Alipourfard, Beatrix Arendt, Daniel M Benjamin, Noam Benkler, Michael M Bishop, Mark Burstein, Martin Bush, James Caverlee, Yiling Chen, Chae Clark, and et al. Systematizing confidence in open research and evidence (score), May 2021. URL osf.io/preprints/socarxiv/46mnb.
[15] Patrice Lopez. Grobid: Combining automatic bibliographic data recognition and term extraction for scholarship publications. In *Proceedings of ECDL*, 2009.
[16] Jian Wu, Pei Wang, Xin Wei, Sarah Rajtmajer, C Lee Giles, and Christopher Griffin. Acknowledgement entity recognition in cord-19 papers. In *Proceedings of the First Workshop on Scholarly Document Processing*, pages 10–19, 2020.
[17] Zhiheng Huang, Wei Xu, and Kai Yu. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*, 2015.
[18] Charles Sutton and Andrew McCallum. An introduction to conditional random fields. *Foundations and Trends® in Machine Learning*, 4(4):267–373, 2012.
[19] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.
[20] Lingfei Wu, Yu Chen, Kai Shen, Xiaojie Guo, Hanning Gao, Shucheng Li, Jian Pei, and Bo Long. Graph neural networks for natural language processing: A survey. *arXiv preprint arXiv:2106.06090*, 2021.
[21] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*, 2016.
[22] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of EMNLP*, 2014.
[23] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In Marilyn A. Walker, Heng Ji, and Amanda Stent, editors, *NAACL-HLT*, 2018.
[24] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *NAACL-HLT*, 2019.
[25] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.
[26] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. 2018.
[27] Jian Wu, Md Reshad Ul Hoque, Gunnar W. Reiske, Michele C. Weigle, Brenda T. Bradshaw, Holly D. Gaff, Jiang Li, and Chiman Kwan. A comparative study of sequence tagging methods for domain knowledge entity recognition in biomedical papers. In Ruhua Huang, Dan Wu, Gary Marchionini, Daqing He, Sally Jo Cunningham, and Preben Hansen, editors, *Proceedings of JCDL*, 2020.