Increasing Sentence-Level Comprehension Through Text Classification of Epistemic Functions

Maria Berger  
Ruhr University Bochum  
maria.berger-a2l@rub.de

Elizabeth J. Goldstein  
Carlisle, PA, USA  
lizgoldstein15@gmail.com

Abstract

Word embeddings capture semantic meaning of individual words. How to bridge word-level linguistic knowledge with sentence-level language representation is an open problem. This paper examines whether sentence-level representations can be achieved by building a custom sentence database focusing on one aspect of a sentence’s meaning. Our three separate semantic aspects are whether the sentence: (1) communicates a causal relationship, (2) indicates that two things are correlated with each other, and (3) expresses information or knowledge. The three classifiers provide epistemic information about a sentence’s content.

1 Introduction

In reading comprehension, the sum is greater than the parts. A strong reader combines the reader’s prior knowledge, reasoning ability, and the text’s substance to reason about the events, entities, and their relations across a full document (Kočiský et al., 2018). In essence, reading comprehension requires the reader to develop high levels of abstraction. (Kočiský et al., 2018)

Each classification task introduced in the paper, if mastered by a computer model, allow a model to comprehend a sentence’s meaning at greater level of abstraction than comparing word meaning similarity between sentences. With this new ability, the model should be able to master more complex downstream tasks than a model limited to word-level comprehension (Kim et al., 2019).

1.1 Causal and Correlating Relations

Causality: Causal relationships are a key to critical reasoning (Magliano and Pillow, 2021; Asghar, 2016). Causal relationship knowledge allows people to control their environment and make predictions about the future (Magliano and Pillow, 2021). It links one’s actions with their probable consequences. Text causality recognition is crucial for natural language information retrieval, event prediction, question answering, generating future scenarios, decision processing, medical text mining, and behavior prediction (Li et al., 2021).

For purposes of this paper, a sentence containing a causal relationship expresses a cause and effect relationship. Following (Li et al., 2021), the expression can be explicit such as, “Financial stress is one of the main causes of divorce.” This paper also recognizes cause and effect relationships, which are implicit such as, “He could choose to go on a diet, but this would bring significant muscle loss.”

Correlation: Correlations are patterns where changes in one variable are associated with changes in a second variable. In other words, one variable’s changes are statistically dependent on the other variable. For purposes of this paper, causal relationships are not considered correlations because we have developed a separate model for causal relationships. Examples of sentences containing correlations are, “Roosters always crow before the sun rises.” and “Chocolate sales were 30% higher in communities with a high number of Covid-cases”. Since there is no causal link, correlative relationship identification allows one to anticipate the future, but not to control one’s surroundings. (Meehan, 1988) For example, prohibiting chocolate sales cannot reduce a community’s Covid-19 rates.

1.2 Knowledge and Information

Knowledge: Humans are master pattern recognizers. Humans automatically and mostly without conscious effort turn experiences into simple patterns that they use to anticipate what will happen, to make things happen, and to choose among options. Spence (Spence, 2005; Meehan, 1988) writes that these simple patterns are generalized, meaning that they are assumed to transcend time. These patterns are what the authors define as knowledge.

Information: Information is knowledge’s building
blocks. It consists of specific instances (Spence, 2005; Meehan, 1988) of, say, descriptions of events, objects, and relations among things. (Spence, 2005; Meehan, 1988) Facts, when used to mean an actual occurrence at a specific time and place, fit within information’s scope. Data, on the other hand, is a smaller unit contained within facts. Bouthillier and Sheare define information as data in context. They identify “–10 degrees” as data and “it is -10 degrees outside” as information. (Bouthillier and Shearer, 2002)

John Dewey explains the concept of knowledge as follows: To run against a hard painful stone is not of itself, I should say, an act of knowing; but if running into a hard and painful thing is an outcome predicted after inspection of data and elaboration of a hypothesis, then the hardness and the painful bruise which define the thing as a stone also constitute it emphatically an object of knowledge. (Ratner et al., 1939, p. 932)

In other words, data about past individual experiences with the pain caused by the skin’s contacts with particular stones at a particular instance is information. One can aggregate these individual data points into a hypothesis, which is referred to in this paper as knowledge.

2 Related Work

2.1 Epistemic Classification
(Marco and Navarro, 1993) define epistemology as, “the study of the process of human knowledge, its logic, origins and basis” and recognize that its study is essential to “the design and implementation of better cognitive strategies for guiding the process of documentary analysis”. (Marco and Navarro, 1993, p. 126–132) The authors, following Meehan’s applied, tactical approach to epistemology, define epistemic as describing the patterns used by humans to organize their experience thereby allowing them to generate future expectations. (Meehan, 1988) Humans create patterns by aggregating specific experiences into general rules to address the future. These patterns include: causation, correlation, and knowledge. Causal and correlative relationships divined from aggregating individual experiences allow humans to anticipate future events. Both causal and correlative relationships which are generalized from individual events fit within the scope of knowledge. Knowledge encompasses all generalized patterns, which humans use to create future expectations. Text classification relating to epistemic distinctions apply universally. (Meehan, 1988) Thus, epistemic text classifications can be utilized by models across disciplines to solve problems presented by downstream tasks.

(Meehan, 1988) introduces an epistemic framework for computational knowledge generation. Humans construct knowledge to serve as a tool to accomplish three basic purposes: anticipating the future, causing change, and choosing among options. (Meehan, 1988) Others refer to computational knowledge engineering as computational epistemology. (i Segura, 2009) This paper’s classification tasks fit within (Meehan, 1988)’s epistemic framework.

2.2 Comprehension Functions
Following (Petukhova and Bunt, 2011), language can be divided into two main components: the communication’s functions and its semantic content. While we have not found earlier research on classifying text based upon whether the text contains a correlative relationship or encodes knowledge, these tasks can be viewed as examining the sentence’s function through an epistemic lens. Researchers have examined various algorithms’ ability to detect communicative functions such as identifying whether a speaker has posed a question (Zhang et al., 2015) or a sentence serves a certain rhetorical purpose in an academic paper. (Iwatsuki et al., 2020).

2.3 Causal Relation Extraction
(Asghar, 2016) divides cause-effect extraction techniques into two main categories: (i) non-statistical, pattern-matching techniques (Khoo et al. (1998), Girju et al. (2002)) and (ii) machine learning techniques Girju (2003), Girju et al. (2010), Sil et al. (2010), Rink et al. (2010), (Zhao et al., 2016)). Each approach has its limitations. The non-statistical rule based technique cannot succeed across domains, produces significantly skewed precision and recall scores, and requires subject-matter experts to craft the pattern-recognition rules. In contrast, statistical and machine learning approaches require a significant amount of time for feature engineering through experimentation (Li et al., 2021), and large, manually created, domain-specifically labeled data sets.

Most research models approach causality extraction as a two-step classification problem requiring the model to first identify cause-effect pair candidates and then to remove those candidates that do
not share a causal link. (Li et al., 2021) However, in recent work, (Li et al., 2021) represent causality extraction as a sequence labeling problem and utilize a BiLSTM-CRF model with Flair contextual embeddings to extract cause–effect pairs directly (Flair embeddings (Akbik et al., 2018)).

Work by (Mirza and Tonelli, 2016) also fits into this group. The authors present CATENA, a sieve-based system for extracting temporal and causal relations. The authors evaluating their system on TempEval-3 and TimeBank-Dense data and show that each of the sieves, the rule-based, the machine-learned and the reasoning-based one contribute to achieving state-of-the-art performance. An analysis of the interaction between the temporal and the causal components show a tight connection between the temporal and the causal dimension of texts.

2.4 Limited Data Sets and Data Augmentation Technique Guidance

Current annotated data sets are relatively limited. For cause and effect pair extraction, the largest data set is EventStoryLine (Caselli and Vossen, 2017). (Zuo et al., 2021) It only contains 258 documents, 4,316 sentences, and 1,770 causal event pairs.

It is an open question on what are the optimal data augmentation techniques. One part of our work is to experiment with data enrichment of training data by generating noisy labels based on labeling functions.

3 Study Overview

3.1 Research Question

In this paper, we examine whether machine and deep learning techniques can identify sentences which communicate causal or correlative relationships and distinguish between sentences that contain knowledge and information.

Our models do not seek to identify candidate pairs (e.g., prior candidate extraction). Pair selection can be difficult when a sentence contains a cause with multiple effects or implicit relationships. Instead we use a gold-labeled data set and test a diverse set of classifiers including a weakly supervised that enriches training data with noisy labels form generated functions. For causality relationships, our technique has the potential to identify sentences that do not utilize traditional, causal function words (e.g., “because”) or impact verbs (e.g., “to poison”). An example of a sentence implicitly indicating a causal relationship, but not utilizing such words is, “Other countries in Southeast Asia—Thailand in particular—wanted to take advantage of these expanding urban markets and jumped into the hot economic fray.”

3.2 Contribution

Following, our contributions are listed:

1. a new approximately 8,300 sentence data set (referred to as the Gold data set) derived from internet articles covering diverse subject-matter, manually labeled for the following three categories: causation, correlation, and knowledge;
2. a new 500 sentence data set (referred to as WikiGold data) derived from Wiki News, manually labeled for the same three categories;
3. experiments assessing the ability of traditional classifiers and neural networks powered by a pre-trained language model to accomplish sentence-level classification for the three classification categories; and
4. experiments evaluating whether a famous data augmentation approach (aka Snorkel, c.f., Sec. 6) can improve model training for discerning the three classification categories.

3.3 Classification Tasks

Our classification tasks are defined as follows:

1. Does the sentence communicate a causal relationship or is none communicated?
2. Does the sentence communicate either a causal or correlative relationship or is neither communicated?
3. Does the sentence communicate a causal relationship or a correlative relationship?
4. Does the sentence contain information or knowledge?

To run the experiments, we use established statistical-based algorithms and neural classifiers (c.f., Sec. 5). We also utilize Snorkel’s weakly supervised data augmentation approach (c.f., Sec. 6).

4 Text Data Used

4.1 Gold Data

To accomplish each of the classification tasks, we created a Gold data set by manually annotating articles from the internet (e.g., Reuters, The Guardian,
etc. See Appendix A for details) written for general readership on a wide range of topics, including artificial intelligence, books, finance, and Covid-19 (see Tab. 1). Fig. 1 shows the topic-wise distribution of the Gold data by sentences, articles, and classification tasks. We can see that the topics are distributed similarly in all the tasks and that—for most of the topics—the balancing in both classes is comparable too. The data set contains 8,327 labeled sentences.

Each article’s sentences were manually labeled for the three classification categories, causality, correlation, and knowledge. No pattern recognition rules were used to label the data. Thus, the labeling method captured a diverse range of sentences which met the labeling criteria.

We create task-specific balanced data sets.

4.2 Gold Data Labeling Conventions

For the first two aspects, causality and correlation, sentences expressing that one such relationship existed between two things were labeled as positively fitting within the respective class. For example, “Eating a cup of blueberries a day does not reduce your risk of cancer” was labeled a causal sentence. The reason for this labeling rule is sentences such as the foregoing one allow the reader to extract causation information from the sentence.

If a sentence contained both a causal and correlational relationship, the sentence was labeled as causal. One such sentence is, “And a study we looked at in 2012 suggested people who owned cats had a higher risk of suicide, as their pets could make them vulnerable to a Toxoplasma gondii (T. gondii) infection.” If a correlation solely related to time, a trend, it was not labeled as a correlation. An example of such a database sentence is, “By the late 1990s, countries such as Indonesia and Brazil were increasing their commercial production by about 10 percent a year.” We chose not to label trends as we sought to identify correlations that are time invariant. If a sentence contained both information and knowledge, it was labeled as fitting within the smaller class, knowledge.

4.3 Wikinews Big Data Set

Snorkel works well in increasing the classification accuracy based on a big data set of noisily labeled data together with a small gold labeled data set. As a big data set, we downloaded a recent version of the English Wikinews data from the web.¹ We filter out articles with titles that contain the prefixes “Template:”, “Portal:”, “Wikinews:”, “Category:” or “Comments:”, because we are not interested in those. “Templates” because they are empty placeholders. “Portal” indicates specific information on the Wikinews portal and its usage. Articles containing the “Wikinews” prefix often simply list short headlines of independent news which are too short for our use cases. “Categories” also indicate some templates that allocate space for news articles with a certain function such as future articles or single dates. Finally, “Comments” are articles that only indicate legacy comments that are now moved to a newly created comments namespace. We also filter out articles that start with a “Redirect” in their text bodies. From a total of 100,277 articles originally culled from Wikinews, 28,460 remained after the filtering-out process. Using NLTK’s sentence parser, we derived 327,366 sentences from these articles. For our experiments we use a subset of this data holding 10,000 sentences. We refer to this data set by “Wiki” data.

4.4 Wikinews Gold Data Set

From the Wiki data, we sampled 500 sentences that we annotated manually for causal and correlating relations. The task-specific, balanced data sets (Sec. 3.3) have sample sizes: 296 (causal versus none), 332 (causal-correlating versus neither), and 36 (causal versus correlating). We refer to this data set as the “WikiGold” data.

5 Supervised Classification

5.1 Classifier Set Up

Data Pre-processing: We perform sentence pre-processing by removing punctuation and lower casing the text. We also enrich the text with part-of-speech (POS) tags using the NLTK library version 3.5. We did not remove stop words because such words include key function words such as “because” which often indicate causal relationships.²

For the traditional classifiers, we compute the tf-idf matrix using TfidfVectorizer of Python’s scikit-learn module, and for the bag of words matrix, we use CountVectorizer (c.f., Tab. 2 showing classifiers used).

---

¹“enwikinews-latest-pages-articles_20201009.xml”
https://dumps.wikimedia.org/enwikinews/
²We performed experiments with lemmatizing and stemming the text, but this did not increase performance.
| topic     | #articles | #sentences | causal vs. corr | caus/corr vs. neither | causal vs. none |
|-----------|-----------|------------|----------------|------------------------|----------------|
| ai        | 4         | 390        | 19             | 3                      | 89             |
| animal    | 2         | 65         | 6              | 3                      | 18             |
| art protest | 3         | 147        | 13             | 3                      | 68             |
| blm       | 3         | 483        | 25             | 18                     | 115            |
| books     | 6         | 251        | 7              | 4                      | 54             |
| china     | 1         | 176        | 21             | 8                      | 74             |
| covid     | 19        | 1398       | 82             | 82                     | 398            |
| creativity | 8         | 617        | 41             | 61                     | 232            |
| culture   | 12        | 715        | 20             | 3                      | 107            |
| depression | 10        | 724        | 57             | 77                     | 319            |
| economy   | 12        | 868        | 34             | 37                     | 191            |
| education | 1         | 136        | 10             | 15                     | 61             |
| entertainment | 2        | 180        | 6              | 0                      | 29             |
| farming   | 1         | 42         | 5              | 2                      | 18             |
| food      | 8         | 416        | 17             | 13                     | 95             |
| gender    | 1         | 105        | 6              | 2                      | 23             |
| health    | 1         | 74         | 6              | 24                     | 45             |
| history   | 2         | 78         | 3              | 2                      | 20             |
| home      | 1         | 33         | 3              | 0                      | 8              |
| math      | 3         | 268        | 27             | 3                      | 79             |
| music     | 1         | 45         | 0              | 3                      | 10             |
| politics  | 4         | 237        | 8              | 2                      | 52             |
| poverty   | 10        | 240        | 9              | 62                     | 108            |
| psychology | 4         | 213        | 16             | 16                     | 109            |
| science   | 3         | 142        | 10             | 15                     | 108            |
| sports    | 1         | 15         | 1              | 0                      | 2              |
| technology | 3         | 268        | 11             | 5                      | 59             |
| total     | 126       | 8326       | 463            | 463                    | 2424           |

Table 1: left: The Gold data’s topic-wise distribution by articles and sentences; right: topic-wise distribution by number of sentences for each of our three classification categories

BERT Model: We choose a BERT pre-trained model using the implementation for TensorFlow (bert-for-tf2). Specifically, we use the model known as “BERT-base”. For the hyper parameters, we choose the maximum token as 150, a batch size of 32, a learning rate of 1e-5, and a maximum of 5 epochs with a patience of 1. We utilize an Adam optimizer and a sparse categorical cross-entropy loss function. The BERT model’s sentence representation vector is the hidden state of the representation of the first token (CLS) in the final BERT layer. (Devlin et al., 2018)

Our model feeds this CLS token vector into a drop out layer (set to .8) to prevent overfitting. Then, the model utilizes a dense layer to output the probabilities for the binary classes.

5.2 Results & Discussion

For the BERT model, there was a strong positive relationship between the training data size and accuracy. The BERT model was the most successful model for all classification tasks except correlation versus causation. This is not surprising since this task had significantly less training data than the other three tasks. The BERT model on average did 5.9% better than the next most successful model for the three classification tasks where there was a significant amount of training data.

On average, the BERT models had a wider dispersion between the precision and recall scores (9.3%) than the best simple models for each classification task (2.5%). There was an even split for both the BERT and simple models on whether they did better on accuracy or recall.

In examining the true positives generated by the BERT causal versus none model, we saw that four of the ten sentences BERT was most confident about (with 99.9% confidence level) all contained the word “lead”. The remaining six sentences all contained verbs that indicate causation. For some sentences, BERT was confident but wrong when a word associated with causal relationships (such as worst) was used without introducing any cause-effect pair such as, “one of the worst behaviors sometimes exhibited by owl observers is feeding the bird”. BERT was most confident and correct in identifying sentences that were not causal when

---

3These are: affect, contribute, encourage, foster, produce, improve, and modify
the sentence had a proper name in it.

For the cause-correlation versus neither task, BERT was most confident and correct when a verb strongly associated with causation or correlation appeared in the sentence. The sentences which correctly classified neither causative nor correlative and for which the model had the highest confidence were ones which were either short or contained proper nouns.

For the correlation versus causal task, we found that the three sentences that Naive Bayes, the most successful algorithm, most confidently labeled as causal (91% confidence level) all contained “percent” in them. 30% of the top 10 sentences that the Naive Bayes algorithm correctly identified as causal and for which Naive Bayes was most confident contained function words associated with causation.6

For the knowledge versus information task, the 20 sentences the BERT model was most confident and correct classifying as knowledge mostly related to depression or other health issues and reflects the database’s subject matter. For the 20 sentences the model was most confident and correct in identifying as informational sentences, the sentences contained either a pronoun or a proper noun and a majority contained both of these parts of speech. 7

### 6 Weakly Supervised Classification

To determine the effectiveness of data augmentation, we now compare neural classification through supervised learning to a mixed approach of weak supervision and neural classification. The experiments are performed on our three main tasks.

#### 6.1 Snorkel Library

We use the recently released programming library Snorkel (Ratner et al., 2020), which was developed to improve utilizing handcrafted rules to build training data. It works by detecting unidentified correlations and other dependencies among these rules (called “labeling functions”). Snorkel’s learning approach is twofold: First, the LabelModel learns the parameters for accuracy and correlation structure based on weak supervision represented by the labeling functions. The LabelModel generates noisy labels for the training data by weighting the labeling functions. In addition, a final classification model (an LSTM in this work) is applied, which generalizes the information learned from the weighted labeling functions by using the noisy labeled data as training data (Ratner et al., 2020).

| task                      | train | test | classifier | vector | r   | p   | F1  | a   |
|---------------------------|-------|------|------------|--------|-----|-----|-----|-----|
| causal vs none            | 3136  | 785  | Bayes G    | tf-idf | .60 | .62 | .61 | .61 |
|                           | 3136  | 785  | Bayes G    | bow    | .63 | .61 | .62 | .62 |
|                           | 3136  | 785  | Log Reg    | tf-idf | .71 | .68 | .70 | .70 |
|                           | 3136  | 785  | Log Reg    | bow    | .69 | .73 | .71 | .71 |
|                           | 3136  | 785  | Naive B    | tf-idf | .73 | .62 | .67 | .65 |
|                           | 3136  | 785  | Naive B    | bow    | .73 | .60 | .66 | .64 |
|                           | 3136  | 785  | RF         | tf-idf | .68 | .69 | .68 | .68 |
|                           | 3136  | 785  | RF         | bow    | .65 | .74 | .69 | .68 |
|                           | 3374  | 844  | BERT       | context | .75 | .79 | .77 | .78 |
| causal vs correlating     | 3877  | 970  | Bayes G    | tf-idf | .58 | .60 | .59 | .60 |
|                           | 3877  | 970  | Bayes G    | bow    | .74 | .63 | .68 | .60 |
|                           | 3877  | 970  | Log Reg    | tf-idf | .77 | .68 | .72 | .70 |
|                           | 3877  | 970  | Log Reg    | bow    | .72 | .72 | .72 | .72 |
|                           | 3877  | 970  | Naive B    | tf-idf | .86 | .61 | .71 | .66 |
|                           | 3877  | 970  | Naive B    | bow    | .86 | .60 | .70 | .64 |
|                           | 3877  | 970  | RF         | tf-idf | .68 | .70 | .69 | .70 |
|                           | 3877  | 970  | RF         | bow    | .62 | .75 | .68 | .71 |
|                           | 3877  | 970  | BERT       | context | .78 | .83 | .79 | .79 |
| knowledge vs info         | 740   | 186  | Bayes G    | tf-idf | .76 | .71 | .73 | .73 |
|                           | 740   | 186  | Bayes G    | bow    | .75 | .68 | .72 | .70 |
|                           | 740   | 186  | Log Reg    | tf-idf | .70 | .78 | .74 | .75 |
|                           | 740   | 186  | Log Reg    | bow    | .66 | .74 | .70 | .72 |
|                           | 740   | 186  | Naive B    | tf-idf | .82 | .77 | .79 | .78 |
|                           | 740   | 186  | Naive B    | bow    | .79 | .73 | .76 | .75 |
|                           | 740   | 186  | RF         | tf-idf | .67 | .74 | .70 | .72 |
|                           | 740   | 186  | RF         | bow    | .59 | .86 | .70 | .75 |
|                           | 740   | 186  | BERT       | context | .71 | .51 | .62 | .52 |
| knowledge vs info         | 4251  | 1063 | Bayes G    | tf-idf | .79 | .63 | .70 | .67 |
|                           | 4251  | 1063 | Bayes G    | bow    | .78 | .67 | .72 | .70 |
|                           | 4251  | 1063 | Log Reg    | tf-idf | .79 | .78 | .79 | .79 |
|                           | 4251  | 1063 | Log Reg    | bow    | .76 | .79 | .77 | .78 |
|                           | 4251  | 1063 | Naive B    | tf-idf | .86 | .74 | .80 | .78 |
|                           | 4251  | 1063 | Naive B    | bow    | .82 | .74 | .78 | .78 |
|                           | 4251  | 1063 | RF         | tf-idf | .75 | .73 | .74 | .74 |
|                           | 4251  | 1063 | RF         | bow    | .72 | .76 | .74 | .75 |
|                           | 4251  | 1063 | BERT       | context | .86 | .80 | .82 | .83 |

Table 2: Results of comparing experiment showing (pre)cision, (rec)all, (f)1-score, and (a)ccuracy; Experiments run using (Bayes G)aussian, BERT, (Log)istic (Reg)ression, (Na¨ıve B)ayes and (R)andom (F)orest classifiers and model; Classes are causal, (co)relating and none/neither, and knowledge and (info)rmation; For BERT, training data set sizes include validation data (80-20-split)

---

5 An example sentence is, “four in ten hispanics are members of the working class compared to 28 percent of blacks 25 percent of whites but just 16 percent of Asians.”

6 These functions words were: because, by and so.

7 These sentences contained a pronoun and referred to a specific individual’s experience, thus contained information. One of the informational sentences the BERT model correctly and most confidently labeled was, “when she was 17 in 2013 mary climbed out of her bedroom window and ran across a field”.

8 https://github.com/snorkel-team/snorkel
6.2 Train-Test Data Splits
We use the Wiki data (Sec. 4.3) for training Snorkel’s LabelModel\(^8\) and the Gold data (comprising our Gold (Sec. 4.1) and WikiGold (Sec. 4.4) data) for testing it. The neural model then uses the predicted labels from the LabelModel as training data, and, again, the Gold data for testing. Note that we combine 80% of our Gold/WikiGold data with the noisy labeled data to train the neural classifier model and 20% of the Gold/WikiGold data for testing. Fig. 1 illustrates this procedure.

6.3 Baseline on Gold Data
First, we run a baseline experiment using GloVe embeddings (Pennington et al., 2014).\(^9\) Our LSTM uses GloVe embeddings from the 100-dimensions file as features for the neural network. We choose 1000 as the vocabulary length, 100 as output dimension and 100 as input length. Following an Embedding layer, we add the LSTM layer (followed by a 256-dimensions Dense layer with RELU Activation and 0.5 Dropout). For the output, another 1-dimensional Dense layer with Sigmoid Activation is added. While fitting the model, we choose a batch size of 128, 10 epochs and a min\_delta of 1e-4.\(^11\)

6.4 Train on Wiki Data using Function Rules
The LSTM (same parameters as above) is trained using noisily labeled Wiki data plus 80% of Gold/WikiGold, and tested on 20% of the Gold/WikiGold (see Tab. 4).

Our labeling functions are generated by a script that uses keywords typically appearing in causal and correlating sentences. For causal sentences we collected keywords affect, alter, ameliorate, aggravate, because, cause, consequence, contribute, decrease, due, effect, exacerbate, help, impact, improve, increase, lead, reason, reduce, relate, result, so, spur, trigger and worsen. As keywords for correlating sentences we use accompany, associate, correlate, indicate, link, predict, and tend. We collected these terms while preparing the Gold data.

This approach is comparable with work by (Girju, 2003) who uses specific verbs placed between two noun phrases to indicate causal relations. Depending on the task, we label a sentence 0 when none of the causal or causal and correlating words appear in the sentences. For causal versus correlating, we label -1 when neither appear.

6.5 Part-of-Speech Features on Top
We also add labeling functions generated from POS tags.\(^12\) Tab. 5 lists the POS tags used. Specifically, the rules applied are: (1) if either of the following POS tags is contained in the sentence’s POS tag sequence: IN, WDT, RB, MD, then the sentence fits in the causal category; and (2) if the CC tag is found within the sentence, then the sentence fits in the correlating sentences category.

After experimenting with a simple entropy measure to find POS tag signals, we expand the POS tags list for identifying causal rules. We add: PRP, RP, and TO.

6.6 Results & Discussion
For the baseline experiment, the accuracy, recall, precision and f1-score values are listed in Tab. 3. The cause versus correlation task performs significantly worse than the other two tasks. However, the training data we have available for this task is only about one fifth of the training data available for the other two tasks. Comparing these results to those in Tab. 2, one finds that the baseline accuracy scores align with the BERT accuracy scores. This means the baseline models and the BERT model perform best on the task with the most data; the cause-correlation versus neither task, and the poorest on the one with the least data: the cause versus correlation task.

Utilizing POS tags and keywords for creating the labeling rules, we reach the baseline in the causal-correlation versus neither task (see Tab. 4). The recall score is better than the baseline. Thus, the training data generated from the POS tags must provide the model with more diverse data; consequently, the classification model is able to recognize more nuanced patterns indicative of causation and correlation. This ability might help the model to perform well in finding true positives (causal and correlative sentences). However, many false positives also show that the rules (key words and

\(^8\)Snorkel’s LabelModel class (Ratner et al., 2019) uses the covariances of the junction tree representation of the dependency tree built by applying the learning functions to all data points. After matrix completion, the resulting conditional probabilities of the functions serve as parameters to re-weight and combine the label output.

\(^9\)Using GloVe gives us a comprehensive view on the problem. Being more straightforward, it better applies for testing against weak supervision and keeps computational costs low.

\(^10\)Presenting a comparable problem, we follow https://www.kaggle.com/kredy10/simple-lstm-for-text-classification

\(^11\)The Wiki and WikiGold data were POS-tagged using the NLTK library version 3.4.5. The Gold data was POS-tagged using NLTK 3.5.
Figure 1: Snorkel workflow: Training and testing data splits in Snorkel’s LabelModel and the neural model

| task                     | train (validate) on          | test on         | r  | p  | f1 | a  | loss |
|--------------------------|------------------------------|-----------------|----|----|----|----|------|
| ca vs none               | WikiGold+Gold               | WikiGold+Gold   | .93| .57| .71| .62| .70  |
| sample size              | 2699 (675)                  | 844             |    |    |    |    |      |
| ca/co vs neither         | WikiGold+Gold               | WikiGold+Gold   | .45| .77| .57| .64| .66  |
| sample size              | 3315 (829)                  | 1036            |    |    |    |    |      |
| ca vs co                 | WikiGold+Gold               | WikiGold+Gold   | .75| .53| .62| .55| .74  |
| sample size              | 616 (154)                   | 192             |    |    |    |    |      |

Table 3: Results of baseline experiment

| task                     | train on          | test on          | r  | p  | f1 | a  | loss |
|--------------------------|-------------------|------------------|----|----|----|----|------|
| ca vs none               | Wiki             | WikiGold+Gold    | .87| .60| .71| .64|      |
| sample size              | 9929             | 4218             |    |    |    |    |      |
| ca/co vs neither         | Wiki             | WikiGold+Gold    | .85| .59| .70| .63|      |
| sample size              | 9929             | 5180             |    |    |    |    |      |
| ca vs co                 | Wiki             | WikiGold+Gold    | .00| 1.0| .00| .50|      |
| sample size              | 9929             | 962              |    |    |    |    |      |

Table 4: Results of mixed approach: Validation data on top of training data. Tables show (r)ecall, (p)recision, (f1)-score, and (a)ccuracy

POS tags are not specific enough, since they also appear in “neither” sentences.

The causal versus none task performs comparable to the baseline, showing a slightly lower recall and f1-score.

The results of the LabelModel show an accuracy of 50% for the causal versus correlation task. Only having one POS feature to identify correlating sentences causes this dismal performance. The LabelModel appears to ignore the correlative POS rule, and instead predicts almost only causal sentences. Since the majority of model rules points to each sentence falling within the causal category, the model only reaches 50% accuracy. Moreover, in the causal versus correlation task, because we arbitrarily assigned the correlative sentences to the positive polarity and the causal sentences to the negative polarity, the model created with the POS feature rules leads to almost no true positives.

However, the LSTM reaches a much higher precision (at the cost of low recall) and higher accuracy compared to the baseline. Looking at samples, one can conclude that the “CC” pattern is a weak one as most true negatives (causal sentence votes) do contain the “CC” pattern too. One also finds that many of false negatives (predicted causal (0), actual correlating (1)—which negatively impacts the recall—do also contain a “CC” pattern. These samples also tend to be relatively short, which often is indicative for being a causal sentence rather than a correlating one.

6.7 Fisher Score for POS Tag Features

Inspired by (Cheng et al., 2007), we perform a Fischer Score test (Duda et al., 1973; Fisher, 1934) to verify suitable POS features. The Fisher Score is defined as follows:

\[ FS = \frac{\sum_{i=1}^{c} n_i (\mu_i - \mu)^2}{\sum_{i=1}^{c} n_i \sigma_i^2} \]  

where \( c \) is the number of classes (2 in our case), \( n_i \) is the number of data samples in class \( i \), \( \mu_i \) and \( \mu \) are the average appearance of the given feature in class \( i \) and the whole dataset respectively. \( \sigma_i \) is the standard deviation of the feature’s appearance.
in class $i$. Our features have a boolean character, which means that they either appear in a sample or they don’t. Hence, our $\mu$’s cannot be greater than 1. The higher the FS, the more important is the given feature.

The higher the FS, the more important is the given feature.

| Feature | part-of-speech | ca vs n | ca vs co | ca/co vs n |
|---------|----------------|---------|----------|------------|
| IN      | preposition    | .03     | .00      | .04        |
| WDT     | Wh-determiner  | .02     | .02      | .01        |
| RB      | adverb         | .01     | .00      | .02        |
| MD      | modal          | .04     | .03      | .03        |
| PRP     | personal pronoun | .01   | .03      | .00        |
| RP      | particle       | .00     | .00      | .00        |
| TO      | to             | .04     | .02      | .03        |
| CC      | coordinating conjunct. | .02 | .00      | .02        |

Table 5: Fisher Score for our POS-tag features calculated on our Gold dataset

The POS-tags IN, MD, and TO are outstanding for each of the three tasks, all of which we use to find causal sentences. It confirms that modals and subordinating conjunctions play a special role for classifying causal sentences.

6.8 Summary

Summarizing, Snorkel works moderately well in the sentence classifying epistemic tasks. The causal versus none task works comparable to the baseline, the causal versus correlating task reaches a higher f1-score with lower accuracy, and the causal-correlating versus neither task reaches a 15%-higher f1-score while accuracy is on par with the baseline. The overall aim to improve classification accuracy using a big corpus of noisily labeled data worked partly for our problem. Note that even though we ran some basic clean-up steps on our Wiki data, we are aware that special characters, prefixes or re-codings of them can corrupt a significant portion of the data set’s sentences, hence leading to affecting noise.

7 Conclusion

To classify sentences based upon their epistemic functions, we created a new data set containing 8,327 sentences that we annotated for whether a sentence communicates a causal or correlating relationship, and whether it expresses information or knowledge. Utilizing BERT, the three epistemic function models achieved sufficient accuracy that future work should include determining the models’ abilities to solve real world problems. In that vein, we plan to utilize our models for a range of downstream tasks such as event detection, question answering, and discourse parsing. Our approach may be able to improve upon these tasks by providing the underlying model additional features through which to comprehend the text in a more nuanced fashion. In addition, as the Snorkel approach appears promising, we plan in future work to determine if there is a systematic method, which can be utilized to determine the underlying functions needed to gather the weakly-supervised data. If this can be accomplished, then a lesser amount of gold data may be needed. As gold data is generally the biggest bottleneck that NLP practitioners face, a systematic approach to Snorkel may allow for a significant time reduction for building future NLP models.

References

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. In Proceedings of the 27th international conference on computational linguistics, pages 1638–1649.

Nabiha Asghar. 2016. Automatic extraction of causal relations from natural language texts: a comprehensive survey. arXiv preprint arXiv:1605.07895.

France Bouthillier and Kathleen Shearer. 2002. Understanding knowledge management and information management: the need for an empirical perspective. Information research, 8(1):8–1.

Tommaso Caselli and Piek Vossen. 2017. The event storyline corpus: A new benchmark for causal and temporal relation extraction. In Proceedings of the Events and Stories in the News Workshop, pages 77–86.

Hong Cheng, Xifeng Yan, Jiawei Han, and Chih-Wei Hsu. 2007. Discriminative frequent pattern analysis for effective classification. In 2007 IEEE 23rd international conference on data engineering, pages 716–725. IEEE.

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

Richard O Duda, Peter E Hart, et al. 1973. Pattern classification and scene analysis, volume 3. Wiley New York.

Ronald Aylmer Fisher. 1934. Two new properties of mathematical likelihood. Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character, 144(852):285–307.
Roxana Girju. 2003. Automatic detection of causal relations for question answering. In Proceedings of the ACL 2003 workshop on Multilingual summarization and question answering, pages 76–83.

Roxana Girju, Brandon Beamer, Alla Rozovskaya, Andrew Fister, and Suma Bhat. 2010. A knowledge-rich approach to identifying semantic relations between nominals. Information processing & management, 46(5):589–610.

Roxana Girju, Dan I Moldovan, et al. 2002. Text mining for causal relations. In FLAIRS conference, pages 360–364.

Kenichi Iwatsuki, Florian Boudin, and Akiko Aizawa. 2020. An evaluation dataset for identifying communicative functions of sentences in english scholarly papers. In Proceedings of the 12th Language Resources and Evaluation Conference, pages 1712–1720.

Christopher SG Khoo, Jaklin Kornfilt, Robert N Oddy, and Sung Hyon Myaeng. 1998. Automatic extraction of cause-effect information from newspaper text without knowledge-based inferencing. Literary and Linguistic Computing, 13(4):177–186.

Jihyeok Kim, Reinald Kim Amplayo, Kyungjae Lee, Sua Sung, Minji Seo, and Seung-won Hwang. 2019. Categorical metadata representation for personalized text classification. Transactions of the Association for Computational Linguistics, 7:201–215.

Tomáš Kočiský, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. 2018. The narrativeqa reading comprehension challenge. Transactions of the Association for Computational Linguistics, 6:317–328.

Zhaoning Li, Qi Li, Xiaotian Zou, and Jiannao Ren. 2021. Causality extraction based on self-attentive bilSTM-CRF with transferred embeddings. Neurocomputing, 423:207–219.

Joseph P. Magliano and Bradford H. Pillow. 2021. Learning - causal reasoning - events, event, students, and effect. https://education.stateuniversity.com/pages/2163/Learning-CAUSAL-REASONING.html#ixzz6vc5QYKKa. Accessed: May 2021.

Francisco Javier Garcia Marco and Miguel Angel Esteban Navarro. 1993. On some contributions of the cognitive sciences and epistemology to a theory of classification. KO KNOWLEDGE ORGANIZATION, 20(3):126–132.

Egene J. Meehan. 1988. The Thinking Game: A Guide to Effective Study. Chatham House Publishers, Inc.

Paramita Mirza and Sara Tonelli. 2016. Catena: Causal and temporal relation extraction from natural language texts. In Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pages 64–75.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pages 1532–1543.

Volha Petukhova and Harry Bunt. 2011. Incremental dialogue act understanding. In Proceedings of the Ninth International Conference on Computational Semantics (IWCS 2011).

Alexander Ratner, Stephen H Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Ré. 2020. Snorkel: Rapid training data creation with weak supervision. The VLDB Journal, 29:709–730.

Alexander Ratner, Braden Hancock, Jared Dumnmon, Frederic Sala, Shreyash Pandey, and Christopher Ré. 2019. Training complex models with multi-task weak supervision. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 33, pages 4763–4771.

Joseph Ratner et al. 1939. Intelligence in the modern world. john dewey’s philosophy. Journal of Philosophy, 36(21).

Bryan Rink, Cosmin Adrian Bejan, and Sandra Harabagiu. 2010. Learning textual graph patterns to detect causal event relations. In Twenty-Third International FLAIRS Conference.

Jordi Vallverdú i Segura. 2009. Computational epistemology and e-science: A new way of thinking. Minds and Machines, 19(4):557–567.

Avirup Sil, Fei Huang, and Alexander Yates. 2010. Extracting action and event semantics from web text. In AAAI Fall Symposium: Commonsense Knowledge. Citeseer.

Larry Spence. 2005. Knowledge versus information. Unpublished, on file with authors.

Yun Zhang, David Lo, Xin Xia, and Jian-Ling Sun. 2015. Multi-factor duplicate question detection in stack overflow. Journal of Computer Science and Technology, 30(5):981–997.

Sendong Zhao, Ting Liu, Sicheng Zhao, Yiheng Chen, and Jian-Yun Nie. 2016. Event causality extraction based on connectives analysis. Neurocomputing, 173:1943–1950.

Xinyu Zuo, Pengfei Cao, Yubo Chen, Kang Liu, Jun Zhao, Weihua Peng, and Yuguang Chen. 2021. Learnda: Learnable knowledge-guided data augmentation for event causality identification. arXiv preprint arXiv:2106.01649.
A Appendix: Sources of Web Articles
