Identifying Quantifiably Verifiable Statements from Text

Pegah Jandaghi
University of Southern California
jandaghi@usc.edu

Jay Pujara
Information Sciences Institute
jpujara@isi.edu

Abstract

Humans often describe complex quantitative data using trend-based patterns. Trend-based patterns can be interpreted as higher order functions and relations over numerical data such as extreme values, rates of change, or cyclical repetition. One application where trends abound are descriptions of numerical tabular data. Therefore, the alignment of numerical tables and textual description of trends enables easier interpretations of tables. Most existing approaches can align quantities in text with tabular data but are unable to detect and align trend-based patterns about data. In this paper, we introduce the initial steps for aligning trend-based patterns about the data, i.e. the detection of textual description of trends and the alignment of trends with a relevant table. We introduce the problem of identifying *quantifiably verifiable statements (QVS)* in the text and aligning them with tables and datasets. We define the structure of these statements and implement a structured based detection. In our experiments, we demonstrate our method can detect and align these statements from several domains and compare favorably with traditional sequence labeling methods.

1 Introduction

There is a wealth of information locked in the numerical tables, spanning different domains and real world applications (e.g., financial reports). Since numerical tables can contain dense, high-dimensional quantitative data, they are often accompanied by textual descriptions that support easy interpretation. In many cases, these textual interpretations are used without inspecting the raw data in the numerical table. [1]

When humans generate textual descriptions of numerical data, they rarely refer to the individual quantitative points, but frequently use trend-based patterns in their statements. Trend-based patterns in numerical data are description of functions and patterns over one or more dataset points in the numerical dataset. In other words, trend-based patterns are created by quantitative analysis over numerical data. The ability to identify these statements and their underlying data source is a prerequisite for many tasks such as fact-checking, natural language understanding in specific domains (e.g stock market), question answering and etc.

To date, many existing works (Ciampaglia et al., 2015; Shi and Weninger, 2016; Pan et al., 2018) have focused on the extraction of subject, predicate, object triples from text. Triple representations readily align with factual data stored as triples in knowledge graphs. However, in a trend-based statement, the supporting data is generally a derived measure on dimensionally-aligned data (such as time series) which cannot readily be validated with triple-based data. Another major limitation of current extraction and alignment methods(Ibrahim et al., 2019; Madaan et al., 2016; Roy et al., 2015) is that they are limited to the statements with first-order trends and are unable to detect and match the second-order descriptions over quantities. In this work, we focus on statements containing higher order trends about numerical data. These higher order trends are created by quantitative analysis over data. Hence, their detection and alignment requires linguistic, symbolic and quantitative reasoning.

In this paper, we introduce a pipeline, for identification and alignment of quantifiably verifiable statement i.e., statements that contain trend-based patterns about data. In the first step of our pipeline, the quantifiably verifiable statements are identified. Then they are aligned with the relevant evidence from a pre-collected dataset.

We define a quantifiably verifiable statement (QVS) as a textual span that expresses a numerical relationship in a dataset and can be objectively validated using an authoritative data set. For example, the statement “US gas prices rose in

[1] NON-ARCHIVAL submission

Proceedings of the First Workshop on Matching From Unstructured and Structured Data (MATCHING 2023), pages 14–22
July 13, 2023 ©2023 Association for Computational Linguistics

2018." describes a change in value (rise) and can be objectively validated using a dataset of commodity pricing information collected by the World Bank. In the effort to align these statements, the detection component converts these statements into an *indicator* and a *trend* structure, formally defined in the next section. Intuitively, indicators allow a system to identify a specific dataset as reference dataset that is described the claim, while the trend expresses a particular data relationship that can be computationally checked on the data. The next step after identifying QVS is the alignment. The alignment step aims to find the relevant information that can be used in verification of the statement. In this paper, we presume the relevant information appears as datasets from which the QVS can be generated without any external source of information or reasoning step. For example, the statement " House prices continued their record-setting growth into May," can be generated using the US house price index dataset. As the initial step for finding the relevant information, the alignment component finds candidate datasets from a pool of pre-created datasets. The candidate selection is based on finding the datasets which are semantically similar to the indicator of the statement. e.g the indicator "Mortgage rate" is more likely to be related to the table "US house price" rather than "Cigarette sales". Our contributions are:

- We define the class of quantifiably verifiable facts and their structure
- We implement a method that detects and aligns quantifiably verifiable statement with a relevant dataset
- We create the first dataset containing real world news from public sources with parallel relevant tables.

2 Problem Definition

In this section, we formally define the problem of identification and alignment of quantifiably verifiable statements(QVS). Let T be a textual corpus consisting of assertions $A \in T$ where each assertion is a natural language statement that can be represented as a sequence of tokens. A quantifiably verifiable assertion makes a claim about a value or set of values in a single or multiple datasets. In this paper, we focus on a subset of verifiable assertions that make a claim about a single dataset. We assume all such claims can be represented by a function $f(A, D_A)$, that, given a claim (A) and a dataset (D_A) as input, designates the claim as true ($\top$) or false ($\bot$). Let V be the set of all claims in verifiable assertions. Formally, for each $A \in V$ if $\exists D_A, f_A$ s.t. $f_A(A, D_A) \in \{\top, \bot\}$ where f_A is a function that can verify A by analyzing the values of D_A. Table 1 contains examples of QVS. In the following subsections we define the substasks of QVS identification and alignment.

2.1 Identification of QVS

A QVS is structured as a sentence which contains an indicator i and a trend t linked to that indicator. Trend and Indicator are each a sequence of tokens. An *indicator* is defined as a concept that can be quantitatively measured either directly or using a commonly agreed upon proxy and its value can vary in time. Therefore, there exists a corresponding time series for each indicator. Indicators are either expressed in text as noun phrases, e.g., "Africa's GDP", "the price of crude oil in Nigeria", or they are expressed in multiple noun phrases in a statement, e.g "sales for durable goods" in the sentence 'Sales increased for durable goods in US'. In this paper, we limit the domain of indicators to the single noun phrases. Indicators provide a reference of the dataset which the statement is describing. More specifically for a claim A, an indicator can be used when looking for D_A i.e reference dataset. In other words, indicators are text spans in the statement referring to a dataset (D_A). They are either name of a currently available dataset or a potential dataset. *Trends* are sequences of words in the sentences and can have several different forms, ranging from a statement about a specific data point or points ("San Francisco's temperatures in January were an outlier"), a pattern spanning several values ("overnight rainfall will increase"), a reference to an aggregate measure ("low temperatures for Sunday"), a comparison against another dataset("compared to last year's snowfall") or a recurring pattern. Table 1 contains sample statements for each trend form. This definition of trends includes higher order descriptions i.e they do not directly express the quantities in dataset and are describing a function over data points. e.g in the statement "The world's population continues to grow" the trend is referring to the continuous increase in the value and does not mention the exact value of the world's population. For assured alignment of these statements to numerical data, the method should be able to detect the increasing patterns in this dataset. In other words, alignment of these statements requires more in-depth reasoning over data which we call functional reasoning. In considering verifiable assertions, we define a quantifiably verifiable assertion (trend-indicator verifiable assertion) to be a subset $V_{ti} \in V$ where each assertion $A \in V_{ti}$ can be expressed in the form of $\langle t, i \rangle$. For example, the statement "The Netherlands trade surplus narrowed to EUR 4.05 billion" will be expressed as ⟨The Netherlands trade surplus, narrowed to EUR 4.05 billion⟩. The challenges in identification of QVS include:

Variability: A trend-based pattern can be described in numerous ways. For example the phrases "the sharp upward trend began" and "demands has been rising since" are both describing the same increasing pattern in the data. Therefore, there is a high linguistic variability on QVS.

Domain Dependency: Trend-based patterns are interpreted differently depending on their domains i.e the terminology used to describe a trend-based pattern varies between domains . For example, the cyclic pattern is interpreted as "measles annual wave" in the epidemiology domain while it is interpreted as "cycles of glacial advance and retreat" in the environment domain.

2.2 Alignment of quantifiably verifiable statements to datasets

With the extracted statement $A = \langle t, i \rangle$, we now define the task of finding the relevant dataset D_A. In this work, D_A is a time series stored in a table. Let D be the set of all time series indicators. The alignment of A is the task of finding $D_A \in D$ such that the values in D_A are necessary and sufficient for the verification of A and every $A' = \langle t', i \rangle$ which has the same indicator as A. For example the quantifiably verifiable statement "In 2012, non-metro child poverty increased to 26.7", expressed as ⟨ non-metro child poverty, increased to 26.7⟩ is aligned with a dataset called "child poverty rate in non metropolitan areas". The alignment problem is similar to the entity linking problem (Shen et al., 2015) and has similar challenges i.e name variation and ambiguity. Name variation addresses the challenge that dataset may be referred to with different names in texts e.g "Senior citizen Population" and "The Population 65 Years and Older" are referring to the same indicator. The ambiguity addresses the challenge that the indicator in the sentence might be referring to more than a single dataset and in order to align it to the dataset correctly more information is required. e.g the indicator "growth" in the statement "Many developing countries, like India and China are experiencing robust growth" can be referring to "economic growth in China" or "Chinas growth in production" or etc. In addition to the mentioned challenges, indicators can be highly correlated or be subset of each others which causes the ambiguity in the alignment e.g the indicator "Midwest gasoline price" is the subset of "US gasoline price". Another challenge is the appearance of operations over indicators. e.g "average sea temparature", "Total operating expenses".

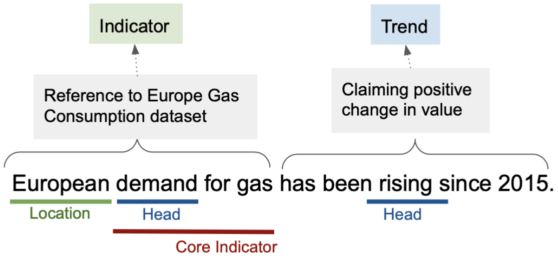

Indicator
Trend
Reference to Europe Gas Consumption dataset
Claiming positive change in value
European demand for gas has been rising since 2015.
Location
Head
Head
Core Indicator

Figure 1: An example of a QVS identified by trend and indicator.

3 Method

In this section, we describe the process of identification and alignment of QVS. Given a set of documents as input, the identification method determines if they are QVS based on their structure. It also provides the trend and indicator representation $\langle t, i \rangle$ of the detected statements which will be used in the alignment method. The alignment method finds D_A for a given quantifiably verifiable statement $A = \langle t, i \rangle$ by using the indicator structure. Ideally the alignment method is provided with a list of all of the datasets (i.e. time series indicators in tables) and finds the dataset (time series) which the statement is making claim about. The dataset alignment is based on the detected i in the statement. For example in the statement "European gas demand has been rising", where "European gas demand" is identified as the indicator by detection component, the alignment selects the "Europe gas consumption" as the output. We describe the method for each task in the following subsections.

3.1 Identification of QVS

A QVS is structured as a sentence containing an indicator i and a trend t linked to that indicator where

Trend Type	Representative Words	Statement
Value at an instant	recorded, been	The official poverty rate in 2019 was 10.5 percent
Statistical function	minimum, average	Volatility peaked at 52% on Tuesday in Brent
Changes over interval	drop, fall	TSLA stock has plummeted 15% in the past three months
Second order effect	accelerated, rebound	South Africa Private Credit Growth Accelerates
Comparison	higher, relative	Prices are up 5.2 percent from the same quarter last year
Recurrent pattern	cycle, seasonality	The sun exhibits a slight brightening and dimming on 11-year cycle

Table 1: Sample quantifiably verifiable statements

trend and indicator are each a sequence of tokens in that statement. Our method is a pipeline, containing a module to detect candidate sequences for trends followed by a module that detects candidate indicator sequences.

We describe each module in the following subsections:

3.1.1 Trend Candidate Detection

As defined previously, a trend expresses a particular data relationship in the text and can have several different forms. The sequence of tokens in a trend are structured to have a head term. Head term is often a trend independent of the context. The trend candidate detection module identifies the keywords in the sentence that are highly likely to be a trend head. e.g., in the statement "About 4 million children did not have any health insurance coverage in 2018, an increase of 425k from the previous year." the detection module selects $increase$ as a trend candidate. Our approach for trend detection is based on the similarity between the trends i.e., the words that appear as trends in statements are likely to have high semantic or context similarity. For example in the statements "Egg prices are skyrocketing" and "TSLA stock has plummeted 15 percent", "skyrocket" and "plummet" have high semantic similarity. Given the prior knowledge that "plummet" is a trend in that statement, we can infer that "skyrocket" is likely a trend as well. With the similarity assumption, we created a trend lexicon and a binary classifier to determine whether a word is a trend candidate. We explain each of these components in the following subsections.

Trend Lexicon Now we describe how we collected a set of keywords that are frequently used to express trends. We collected a corpus of 76 web articles from different domains, including financial and economic reports, environmental science articles, and health and medical writing. Across these different domains, we identified six general classes of trends that were used in time series trend analysis tools (Lloyd et al., 2014; Streibel et al., 2013) which are: values at an instant, statistical functions over a series, changes over an interval, recurrent patterns, second-order effects, and comparisons to baselines or other data. Table 1 contains examples from these classes of trends. To ensure having adequate samples from every trend class, for each trend class, we manually curated a sample set of statements containing a trend from that class i.e., a sample statement set for statistical functions. Then, for each trend sequence in the sample sets, we specified a representative word as trend keyword. e.g., for the statistical function trend type with the sample statements "Inflation Rate in the United States averaged 3.27 percent", "The year 1969 marked a peak in population growth", the words "average", "peak" where selected as representative words for this trend type. These words are representative for trend classes and are used as the initial lexicon. This lexicon contains 60 trend keywords a subset of them is in Table 1. The words which are highly similar to this lexicon are potential trend candidates since words that appear as trends tend to have high semantic or contextual similarity.

Trend Candidate Classifier Given an input document and a set of lexicon, this component classifies the tokens of the document as trend candidates or not based on their similarity to the trend lexicon. As mentioned previously, the trend lexicon contains representative words from all trend types and high similarity of a word with members of this set is an indicator of potential trend. Contextualized word embedding (ELMo) (Peters et al., 2018) have been shown to capture semantic and context of the words. ELMo embeddings capture both the context dependent and context independent features of words. By using ELMo internal states, we can asses the similarity of the words at different levels. Therefore, we used ELMo embeddings to assess the syntactic, semantic and contextual similarity of the words in our task i.e we assumed that trends from a same trend type have close ELMo representation. More specifically, we assumed that any trend candidate will have similar ELMo representation with a member in the collected lexicon. With this assumption, we created a binary probabilistic classifier (logistic regression) to decide whether a word is a trend candidate based on its similarity to the members in the lexicon. We created a feature vector for each input token in the input document by computing the similarity of the token with elements in the trend lexicon. i.e each entry in the similarity vector of a token w, is a semantic similarity score of w and a member from the lexicon. The similarity score is the cosine similarity between ELMO embeddings of the tokens. We use the created similarity vector of each token as the feature vector of that token for the classifier. To reduce the effort of labeling data and creating a training set for this classifier, we used bootstrapping (Yarowsky, 1995) in the training process. The process started with a subset of labeled trends randomly selected from economic news articles[2]. We expanded the initial labeled data iteratively. In each iteration, a set of unlabeled words were sampled and a human annotator labeled them as trend and non-trend. The samples were selected by uncertainty sampling to improve the classifier recall. With uncertainty sampling, we selected a subset of unlabeled tokens that the classifier was not confident about their label i.e the probability of being trend and not trend were close. Then, the new annotated samples were added to the training data. At each iteration, after adding the new labeled samples, we retrained the classifier and evaluated its performance on a development set. We continued the process of expanding the labeled set and retraining until the classifier achieved high accuracy on the development set.

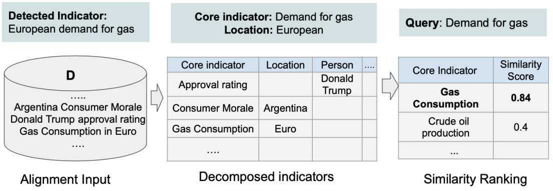
3.2 Indicator Candidates Detection

We defined indicators as text spans in the statement that refer to a dataset. An indicators is a name of an existing dataset, a proxy to an existing dataset or a measurable concept that we can create a dataset by collecting its values over time. In this paper, we are interested in detecting indicators that trends are making claim about. Therefore, our method should capture the dependency between trend and indicators while detecting QVS. To incorporate the dependency of indicators to the trends, our indicator detection utilizes the notion of triggers. (Lin et al., 2020) introduced "entity triggers" as group of tokens in a sentence explaining why humans recognize named entities. Similar to the named entity triggers, we consider trends as triggers for indicators i.e. explanations for why human recognize indicators in the sentences. The indicator detection module training phase includes the trends in the QVS labeled as explanation.

3.3 Dataset Alignment

In this component, with the identified $A = \langle t, i \rangle$ and a set of dataset indicators D, our method finds the most relevant indicator $D_A \in D$ such that the values in D make it possible to verify A. In other words, A is a valid assertion created by reasoning over values in D_A. The alignment component utilizes the structure of the detected indicator i. For each indicator, we have defined a structure consisting of a *core indicator*, *head term*, and *dimensions*. The core indicator is defined as a subtree of the phrase dependency tree that is both necessary and sufficient to identify the corresponding dataset. Specifically, this corresponds to the smallest subtree that is conceptually meaningful and can be measured and adding additional contextual phrases will not affect the identity of measured quantity. The root of the core indicator subtree is identified as the head term and corresponds to the general concept class of the indicator. Finally, the dimensions specify the particular subset of the core indicator measurements that are relevant. As a concrete example, for “the price of crude oil in Nigeria”, the core indicator is “price of crude oil,” the head term is “price” and the dimension is “Nigeria” (location of measurements). Figure 1 shows a sample indicator with its components. To find an aligned dataset with i, i is decomposed to dimensions using spaCy(Honnibal et al., 2020) name entity recognizer. The decomposition reduces the task of indicator alignment to core indicator alignment i.e our goal is find elements in D with similar core indicator to is core indicator. We used SentenceTransformer(Reimers and Gurevych, 2019) for computing the semantic similarity between different core indicators. Using semantic similarity enables us to overcome dataset name variation e.g “new loans” indicator in “Since 1988, Sub Saharan Africa is getting very little in terms of new loans” is considered similar to “Foreign Direct Investment”

[2] https://data.world/crowdflower/us-economic-performance



Core indicator	Location	Person	....
Approval rating		Donald Trump	
Consumer Morale	Argentina		
Gas Consumption	Euro		
....			

Core Indicator	Similarity Score
Gas Consumption	**0.84**
Crude oil production	0.4
...	

Figure 2: Steps in dataset alignment

dataset despite the textual difference. Our method selects the element with the most similar core indicator to i as the candidate for aligned indicator. In the presence of dimensions (e.g. location) in the detected indicator i, we further use the dimensions for more accurate alignment. More specifically, we select all indicators from D which have the same core indicator as i, then we select the indicator with same dimensions from this indicator pool as the aligned indicator. We also use temporal information in the statement to guarantee the existence of the trend in the aligned indicator values i.e., We extract temporal information about the detected trend t from the statement and check availability of values in the detected time span in the aligned indicator. Our method finally outputs the indicator from D which has the closest core indicator to i core indicator and its other dimensions matches those dimensions of i. Figure 2 shows the alignment process for an example indicator.

4 Experiments

We designed and conducted experiments to show the effectiveness of QVS detection and alignment.

4.1 Detection Experiment

In this experiment, we evaluated the performance of QVS detection. Our detection method relies on extracting trend and indicators from the statements i.e it assigns a tag from $\{trend, indicator, none\}$ to every token in the statement and classifies a sentence as quantifiably verifiable if $trend$ and $indicator$ tags appear in the statement. We compare the detection method with sequence tagging and claim detection methods as baselines:

ClaimBuster(Hassan et al., 2017) is an automated Fact-checking system that assigns a checkworthiness score to claims. Since a QVS is a valid claim about a dataset, any claim detection method should identify it. We used ClaimBuster as a baseline, ran claim detection and selected the claims with scores higher than 0.5 as QVS.

LSTM(Lample et al., 2016) has shown great performance for sequence tagging tasks e.g. Named Entity Recognition. We used LSTM with ELMo embeddings of the tokens as inputs. We trained the model using the data we used for training trend candidate detection and classified a QVS if both $trend$ and $indicator$ tags appeared in the statement.

For this experiment, we created 3 dataset, manually labeled them using brat (Stenetorp et al., 2011) annotation tool. The datasets are:

TE: We collected 100 articles from Trading Economic[3] containing news about economic indicators. There are 375 sentences from which 341 are QVS. The content of these articles follow the same structure but vary in terminology.

WSJ: We believed that articles published in wall street journal frequently contain QVS. We collected 100 articles from this source and sampled a statements from each article. The final dataset contains 45 QVS. The articles in this dataset have similar context however the statements demonstrate a high variability in terms of trends descriptions.

Covid: We sampled 1000 news headlines [4] during the coronavirus pandemic in 2020. Our sample consists of 1159 sentences from which 152 are QVS. These articles are from different domains and sources, making this dataset challenging for the detection task.

Table 2 shows the results of this experiment. As expected, ClaimBuster has a high recall and low precision since it detects a wide range of claims. We also observe that our methods achieves the highest accuracy in all datasets and outperforms LSTM model. For the TE dataset, since the majority of the articles are QVS, the recall is the important criteria. Though our method does not have the highest F-1 scores, the recall of our method is as high as ClaimBuster. Which indicates our method ability to overcome context dependency challenge in the TE and detect QVS. For the WSJ and Covid dataset, our method outperforms in terms of F-1 i.e. it achieves higher recall and precision.

4.2 Indicator Alignment Experiment

In this experiment, we evaluated the performance of dataset alignment. We created 3 dataset:

TE: This dataset contains a list of 234 indicators from Trading Economics as D. For a subset of 40 of these indicators, we collected sentences about that indicator from TE and ran the alignment for

[3] https://tradingeconomics.com

[4] https://www.kaggle.com/sagunsh/coronavirus-news-headline

Dataset	TE			WSJ			Covid		
Method	Pr	Re	F-1	Pr	Re	F-1	Pr	Re	F-1
LSTM	.81	.89.	.84	.57	.33	.41	.12	.72	.21
ClaimBuster	.91	.99	**.99**	.45	.95	.61	.19	.94	.32
Our method	.92	.99	.95	.71	.89	**.79**	.5	.57	**.53**

Table 2: Results of experiment 1: quantifiably verifiable statements detection

Method	Our method		String matching		GloVe matching	
Dataset	Pr	Re@3	Pr	Re@3	Pr	Re@3
Gov	**.96**	.98	.78	.88	.51	.67
TE	**.66**	.76	.43	.56	.17	.23
Covid	.91	.93	**.93**	.94	.4	.48

Table 3: Results for experiment 2: dataset alignment. The Pr columns shows the precision and Re@3 is the recall at 3.

the detected indicator in those statements. The indicators in TE dataset are classified to tables based on their topic which includes: Government, Labour, Markets and etc. This classification alludes that each topic include semantically similar indicators. For example the indicators "Corporate tax rate" and "Sales tax rate" are under the "Tax rate" topic.

Gov: We extracted 52 tables and sentences about an indicator in those tables from US government domains. We extracted indicators in those tables which resulted in 52 reference indicators. In this dataset, name variation is low i.e the majority of the indicators appear exactly as they are in the table

Covid: We sampled 234 news headlines about Covid. The sample statements were about Covid infection, death and recovered indicators. We used the indicator list from the TE and Gov as the reference set and evaluated the accuracy of aligning the Covid related indicators. Although the number of indicators in the headlines are limited, the ambiguity is high in this dataset. For example, the statement "UK coronavirus toll passes 19,000" could be aligned with the "covid confirmed cases" and "covid death cases".

We compare our alignment method with baseline methods: string matching and GloVe(Pennington et al., 2014) similarity. For each method, we choose the closest dataset indicator as the aligned dataset. We report the precision of the aligned dataset. For a more thorough evaluation, we also selected the top 3 matched datasets from each method, and reported whether the correct dataset is withing those choices (Recall@3). The results of the experiment are in Table 3. As shown in the table, the baseline methods have good performance in the Gov dataset. This is due to the low ambiguity and name variation in this dataset. The GloVe matchings poor performance in the TE and Covid dataset is rooted in the prevalence of domain specific words(OOV) in these datasets. However our method is robust in those cases. The string matching method has its lowest perforamnce in the TE dataset since the matching fails to achieve high performance in the datasets with high name variation. We observe that for the TE dataset, the difference of Recall@3 and precision are higher compared to the other datasets. This is caused by the presence of indicators which are semantically similar. Overall we observe that our method achieves a reasonable accuracy in all datasets. While it has a slightly lower accuracy in covid dataset where the indicators in the statements are similar to the reference set, it outperforms other methods in more challenging datasets.

4.3 Conclusion and Future Work

We introduced a novel problem of identifying QVS in text and aligning them with tables. We designed a system that extracts and aligns QVS using natural language processing toolkits and semantic features. In our ongoing work, we are working to create more specific alignment of QVS and tables i.e. finding the underlying datapoints and the relation between them. We hope to extend the application of our method and assemble an end-to-end solution for verification of QVS that includes identifying indicators in documents, finding relevant datapoints for verification, and trend analysis systems to compare assertions with data.

4.4 Related Work

The problem of finding alignment between text and tables has been studied for the non-numerical tables (Bhagavatula et al., 2015). (Chen et al., 2021; Cheng et al., 2021) created datasets containing text and numerical tables aligned with them which are used for question answering with quatitative reasoning. The general problem of validating facts in textual data has largely been studied from the perspective of verifying specific triplified knowledge with an explicit set of relationships (Ciampaglia et al., 2015; Shi and Weninger, 2016; Pan et al., 2018). There have been recent studies on verifying statement about tabular and semi-structured data (Wenhu Chen and Wang, 2020; Schlichtkrull et al., 2021; Gupta et al., 2020). These approaches are can decide whether a statement is entailed from tables. There have been several studies on identifying check-worthy claims in text recently (Hassan et al., 2017, 2015; Jaradat et al., 2018). These approaches assign a check-worthy score to each sentence in a document. However, they lack a formal definition for check-worthy claims and do not support quantifiably verifying these claims. The approach in (Konstantinovskiy et al., 2018) has a very general definition for check-worthy claims and it is not possible to check the verifiability of most of them using any data set. (Thorne and Vlachos, 2017) checks the veracity of claims containing temporal numerical information associated with named entities. Information extraction approaches for relations have been intensely studied in both open-world (Etzioni et al., 2008) and ontology-based settings (Wimalasuriya and Dou, 2010). A subfield of extraction approaches that is closely related to our task is that of identifying cause-effect relationships in text (Asghar, 2016). In this subfield, common approaches include bootstrapping from a known set of keywords (Marcu and Echihabi, 2002), using NLP feature sets and semantic features (Rink and Harabagiu, 2010), analysis of graph relationships (Rink et al., 2010) and more recently, neural-network based approaches (de Silva et al., 2017). Identifying and summarizing trends in natural language, the inverse of the problem we tackle, has been notably studied in approaches such as the Automated Statistician (Lloyd et al., 2014; Hwang et al., 2016) and subsequent papers. A relevant research area is the quantification of cognitive expectations for specific increase and decrease trends using crowdsourced studies (Sharp et al., 2018).

References

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

Chandra Bhagavatula, Thanapon Noraset, and Doug Downey. 2015. Tabel: Entity linking in web tables. In *SEMWEB*.

Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matthew I. Beane, Ting-Hao Kenneth Huang, Bryan R. Routledge, and William Yang Wang. 2021. Finqa: A dataset of numerical reasoning over financial data. *ArXiv*, abs/2109.00122.

Zhoujun Cheng, Haoyu Dong, Zhiruo Wang, Ran Jia, Jiaqi Guo, Yan Gao, Shi Han, Jian-Guang Lou, and Dongmei Zhang. 2021. Hitab: A hierarchical table dataset for question answering and natural language generation.

Giovanni Luca Ciampaglia, Prashant Shiralkar, Luis M Rocha, Johan Bollen, Filippo Menczer, and Alessandro Flammini. 2015. Computational fact checking from knowledge networks. *PloS one*, 10(6):e0128193.

Tharini N de Silva, Xiao Zhibo, Zhao Rui, and Mao Kezhi. 2017. Causal relation identification using convolutional neural networks and knowledge based features. *World Academy of Science, Engineering and Technology, International Journal of Computer, Electrical, Automation, Control and Information Engineering*, 11(6):697–702.

Oren Etzioni, Michele Banko, Stephen Soderland, and Daniel S Weld. 2008. Open information extraction from the web. *Communications of the ACM*, 51(12):68–74.

Vivek Gupta, Maitrey Mehta, Pegah Nokhiz, and Vivek Srikumar. 2020. INFOTABS: Inference on tables as semi-structured data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2309–2324, Online. Association for Computational Linguistics.

Naeemul Hassan, Chengkai Li, and Mark Tremayne. 2015. Detecting check-worthy factual claims in presidential debates. *Proceedings of the 24th ACM International Conference on Information and Knowledge Management (CIKM)*, pages 1835–1838.

Naeemul Hassan, Anil Nayak, Vikas Sable, Chengkai Li, Mark Tremayne, Gensheng Zhang, Fatma Arslan, Josue Caraballo, Damian Jimenez, Siddhant Gawsane, Shohedul Hasan, Minumol Joseph, and Aaditya Kulkarni. 2017. Claimbuster: the first-ever end-to-end fact-checking system. *Proceedings of the VLDB Endowment*, 10:1945–1948.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Yunseong Hwang, Anh Tong, and Jaesik Choi. 2016. Automatic construction of nonparametric relational regression models for multiple time series. In *International Conference on Machine Learning*, pages 3030–3039.

Yusra Ibrahim, Mirek Riedewald, Gerhard Weikum, and Demetrios Zeinalipour-Yazti. 2019. Bridging quantities in tables and text. In *2019 IEEE 35th International Conference on Data Engineering (ICDE)*, pages 1010–1021.

Israa Jaradat, Pepa Gencheva, Alberto Barrón-Cedeño, Lluís Màrquez, and Preslav Nakov. 2018. Claimrank: Detecting check-worthy claims in arabic and english. pages 26–30.

Lev Konstantinovskiy, Oliver Price, Mevan Babakar, and Arkaitz Zubiaga. 2018. Towards automated factchecking: Developing an annotation schema and benchmark for consistent automated claim detection.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *NAACL*.

Bill Yuchen Lin, Dong-Ho Lee, Ming Shen, Ryan Moreno, Xiao Huang, Prashant Shiralkar, and Xiang Ren. 2020. Triggerner: Learning with entity triggers as explanations for named entity recognition.

James Robert Lloyd, David K Duvenaud, Roger B Grosse, Joshua B Tenenbaum, and Zoubin Ghahramani. 2014. Automatic construction and natural-language description of nonparametric regression models. In *AAAI*, pages 1242–1250.

Aman Madaan, Ashish Mittal, Mausam, Ganesh Ramakrishnan, and Sunita Sarawagi. 2016. Numerical relation extraction with minimal supervision. In *AAAI*.

Daniel Marcu and Abdessamad Echihabi. 2002. An unsupervised approach to recognizing discourse relations. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 368–375. Association for Computational Linguistics.

Jeff Z Pan, Siyana Pavlova, Chenxi Li, Ningxi Li, Yangmei Li, and Jinshuo Liu. 2018. Content based fake news detection using knowledge graphs. In *International Semantic Web Conference*, pages 669–683. Springer.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proc. of NAACL*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Bryan Rink, Cosmin Adrian Bejan, and Sanda M Harabagiu. 2010. Learning textual graph patterns to detect causal event relations. In *FLAIRS Conference*.

Bryan Rink and Sanda Harabagiu. 2010. Utd: Classifying semantic relations by combining lexical and semantic resources. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 256–259. Association for Computational Linguistics.

Subhro Roy, Tim Vieira, and Dan Roth. 2015. Reasoning about quantities in natural language. *Transactions of the Association for Computational Linguistics*, 3:1–13.

Michael Sejr Schlichtkrull, Vladimir Karpukhin, Barlas Oguz, Mike Lewis, Wen-tau Yih, and Sebastian Riedel. 2021. Joint verification and reranking for open fact checking over tables. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers).*

Rebecca Sharp, Mithun Paul, Ajay Nagesh, Dane Bell, and Mihai Surdeanu. 2018. Grounding gradable adjectives through crowdsourcing. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018).*

Wei Shen, Jianyong Wang, and Jiawei Han. 2015. Entity linking with a knowledge base: Issues, techniques, and solutions. *IEEE Transactions on Knowledge and Data Engineering*, 27:443–460.

Baoxu Shi and Tim Weninger. 2016. Discriminative predicate path mining for fact checking in knowledge graphs. *Knowledge-Based Systems*, 104:123–133.

Pontus Stenetorp, Goran Topić, Sampo Pyysalo, Tomoko Ohta, Jin-Dong Kim, and Jun'ichi Tsujii. 2011. Bionlp shared task 2011: Supporting resources. In *Proceedings of BioNLP Shared Task 2011 Workshop*, pages 112–120, Portland, Oregon, USA. Association for Computational Linguistics.

Olga Streibel, Lars Wissler, Robert Tolksdorf, and Danilo Montesi. 2013. Trend template: Mining trends with a semi-formal trend model. volume 1088.

James Thorne and Andreas Vlachos. 2017. An extensible framework for verification of numerical claims.

Jianshu Chen Yunkai Zhang Hong Wang Shiyang Li Xiyou Zhou Wenhu Chen, Hongmin Wang and William Yang Wang. 2020. Tabfact : A large-scale dataset for table-based fact verification. In *International Conference on Learning Representations (ICLR)*, Addis Ababa, Ethiopia.

Daya C Wimalasuriya and Dejing Dou. 2010. Ontology-based information extraction: An introduction and a survey of current approaches.

David Yarowsky. 1995. Unsupervised word sense disambiguation rivaling supervised methods. In *ACL*.