FEVER: a large-scale dataset for Fact Extraction and VERification

James Thorne[1], **Andreas Vlachos**[1], **Christos Christodoulopoulos**[2], and **Arpit Mittal**[2]

[1]Department of Computer Science, University of Sheffield
[2]Amazon Research Cambridge
{j.thorne, a.vlachos}@sheffield.ac.uk
{chrchrs, mitarpit}@amazon.co.uk

Abstract

In this paper we introduce a new publicly available dataset for verification against textual sources, FEVER: Fact Extraction and VERification. It consists of 185,445 claims generated by altering sentences extracted from Wikipedia and subsequently verified without knowledge of the sentence they were derived from. The claims are classified as SUPPORTED, REFUTED or NOTENOUGHINFO by annotators achieving 0.6841 in Fleiss κ. For the first two classes, the annotators also recorded the sentence(s) forming the necessary evidence for their judgment. To characterize the challenge of the dataset presented, we develop a pipeline approach and compare it to suitably designed oracles. The best accuracy we achieve on labeling a claim accompanied by the correct evidence is 31.87%, while if we ignore the evidence we achieve 50.91%. Thus we believe that FEVER is a challenging testbed that will help stimulate progress on claim verification against textual sources.

1 Introduction

The ever-increasing amounts of textual information available combined with the ease in sharing it through the web has increased the demand for verification, also referred to as fact checking. While it has received a lot of attention in the context of journalism, verification is important for other domains, e.g. information in scientific publications, product reviews, etc.

In this paper we focus on verification of textual claims against textual sources. When compared to textual entailment (TE)/natural language inference (Dagan et al., 2009; Bowman et al., 2015), the key difference is that in these tasks the passage to verify each claim is given, and in recent years it typically consists a single sentence, while in verification systems it is retrieved from a large set of documents in order to form the evidence. Another related task is question answering (QA), for which approaches have recently been extended to handle large-scale resources such as Wikipedia (Chen et al., 2017). However, questions typically provide the information needed to identify the answer, while information missing from a claim can often be crucial in retrieving refuting evidence. For example, a claim stating “Fiji's largest island is Kauai.” can be refuted by retrieving “Kauai is the oldest Hawaiian Island.” as evidence.

Progress on the aforementioned tasks has benefited from the availability of large-scale datasets (Bowman et al., 2015; Rajpurkar et al., 2016). However, despite the rising interest in verification and fact checking among researchers, the datasets currently used for this task are limited to a few hundred claims. Indicatively, the recently conducted Fake News Challenge (Pomerleau and Rao, 2017) with 50 participating teams used a dataset consisting of 300 claims verified against 2,595 associated news articles which is orders of magnitude smaller than those used for TE and QA.

In this paper we present a new dataset for claim verification, FEVER: Fact Extraction and VERification. It consists of 185,445 claims manually verified against the introductory sections of Wikipedia pages and classified as SUPPORTED, REFUTED or NOTENOUGHINFO. For the first two classes, systems and annotators need to also return the combination of sentences forming the necessary evidence supporting or refuting the claim (see Figure 1). The claims were generated by human annotators extracting claims from Wikipedia and mutating them in a variety of ways, some of which were meaning-altering. The verification of each

Proceedings of NAACL-HLT 2018, pages 809–819
New Orleans, Louisiana, June 1 - 6, 2018. ©2018 Association for Computational Linguistics

claim was conducted in a separate annotation process by annotators who were aware of the page but not the sentence from which original claim was extracted and thus in 31.75% of the claims more than one sentence was considered appropriate evidence. Claims require composition of evidence from multiple sentences in 16.82% of cases. Furthermore, in 12.15% of the claims, this evidence was taken from multiple pages.

To ensure annotation consistency, we developed suitable guidelines and user interfaces, resulting in inter-annotator agreement of 0.6841 in Fleiss κ (Fleiss, 1971) in claim verification classification, and 95.42% precision and 72.36% recall in evidence retrieval.

To characterize the challenges posed by FEVER we develop a pipeline approach which, given a claim, first identifies relevant documents, then selects sentences forming the evidence from the documents and finally classifies the claim w.r.t. evidence. The best performing version achieves 31.87% accuracy in verification when requiring correct evidence to be retrieved for claims Supported or Refuted, and 50.91% if the correctness of the evidence is ignored, both indicating the difficulty but also the feasibility of the task. We also conducted oracle experiments in which components of the pipeline were replaced by the gold standard annotations, and observed that the most challenging part of the task is selecting the sentences containing the evidence. In addition to publishing the data via our website[1], we also publish the annotation interfaces[2] and the baseline system[3] to stimulate further research on verification.

Claim: The Rodney King riots took place in the most populous county in the USA.

[wiki/Los_Angeles_Riots]
The 1992 Los Angeles riots, also known as the Rodney King riots were a series of riots, lootings, arsons, and civil disturbances that occurred in Los Angeles County, California in April and May 1992.

[wiki/Los_Angeles_County]
Los Angeles County, officially the County of Los Angeles, is the most populous county in the USA.

Verdict: Supported

Figure 1: Manually verified claim requiring evidence from multiple Wikipedia pages.

2 Related Works

Vlachos and Riedel (2014) constructed a dataset for claim verification consisting of 106 claims, selecting data from fact-checking websites such as PolitiFact, taking advantage of the labelled claims available there. However, in order to develop claim verification components we typically require the justification for each verdict, including the sources used. While this information is usually available in justifications provided by the journalists, they are not in a machine-readable form. Thus, also considering the small number of claims, the task defined by the dataset proposed remains too challenging for the ML/NLP methods currently available. Wang (2017) extended this approach by including all 12.8K claims available by Politifact via its API, however the justification and the evidence contained in it was ignored in the experiments as it was not machine-readable. Instead, the claims were classified considering only the text and the metadata related to the person making the claim. While this rendered the task amenable to current NLP/ML methods, it does not allow for verification against any sources and no evidence needs to be returned to justify the verdicts.

The Fake News challenge (Pomerleau and Rao, 2017) modelled verification as stance classification: given a claim and an article, predict whether the article supports, refutes, observes (neutrally states the claim) or is irrelevant to the claim. It consists of 50K labelled claim-article pairs, combining 300 claims with 2,582 articles. The claims and the articles were curated and labeled by journalists in the context of the Emergent Project (Silverman, 2015), and the dataset was first proposed by Ferreira and Vlachos (2016), who only classified the claim w.r.t. the article headline instead of the whole article. Similar to recognizing textual entailment (RTE) (Dagan et al., 2009), the systems were provided with the sources to verify against, instead of having to retrieve them.

A differently motivated but closely related dataset is the one developed by Angeli and Manning (2014) to evaluate natural logic inference for common sense reasoning, as it evaluated sim-

[1] http://fever.ai

[2] https://github.com/awslabs/fever

[3] https://github.com/sheffieldnlp/fever-baselines

ple claims such as “not all birds can fly” against textual sources — including Wikipedia — which were processed with an Open Information Extraction system (Mausam et al., 2012). However, the claims were small in number (1,378) and limited in variety as they were derived from eight binary ConceptNet relations (Tandon et al., 2011).

Claim verification is also related to the multilingual Answer Validation Exercise (Rodrigo et al., 2009) conducted in the context of the TREC shared tasks. Apart from the difference in dataset size (1,000 instances per language), the key difference is that the claims being validated were answers returned to questions by QA systems. The questions and the QA systems themselves provide additional context to the claim, while in our task definition the claims are outside any particular context. In the same vein, Kobayashi et al. (2017) collected a dataset of 412 statements in context from high-school student exams that were validated against Wikipedia and history textbooks.

3 Fact extraction and verification dataset

The dataset was constructed in two stages[4] :

Claim Generation Extracting information from Wikipedia and generating claims from it.

Claim Labeling Classifying whether a claim is supported or refuted by Wikipedia and selecting the evidence for it, or deciding there’s not enough information to make a decision.

3.1 Task 1 - Claim Generation

The objective of this task was to generate claims from information extracted from Wikipedia. We used the June 2017 Wikipedia dump, processed it with Stanford CoreNLP (Manning et al., 2014), and sampled sentences from the introductory sections of approximately 50,000 popular pages.[5]

The annotators were given a sentence from the sample chosen at random, and were asked to generate a set of **claims** containing a single piece of information, focusing on the entity that its original Wikipedia page was about. We asked the annotators to generate claims about a single fact which could be arbitrarily complex and allowed for a variety of expressions for the entities.

If only the source sentences were used to generate claims then this would result in trivially verifiable claims, as the new claims would in essence be simplifications and paraphrases. At the other extreme, if we allowed world knowledge to be freely incorporated it would result in claims that would be hard to verify on Wikipedia alone. We address this issue by introducing a **dictionary**: a list of terms that were (hyper-)linked in the original sentence, along with the first sentence from their corresponding Wikipedia pages. Using this dictionary, we provide additional knowledge that can be used to increase the complexity of the generated claims in a controlled manner.

The annotators were also asked to generate **mutations** of the claims: altered versions of the original claims, which may or may not change whether they are supported by Wikipedia, or even if they can be verified against it. Inspired by the operators used in Natural Logic Inference (Angeli and Manning, 2014), we specified six types of mutation: paraphrasing, negation, substitution of an entity/relation with a similar/dissimilar one, and making the claim more general/specific.

During trials of the annotation task, we discovered that the majority of annotators had difficulty generating non-trivial negation mutations (e.g. mutations beyond adding “not” to the original). Besides providing numerous examples for each mutation, we also redesigned the annotation interface so that all mutation types were visible at once and highlighted mutations that contained “not” in order to discourage trivial negations. Finally, we provided the annotators with an ontology diagram to illustrate the different levels of entity similarity and class membership.

This process resulted in claims (both extracted and mutated) with a mean length of 9.4 tokens which is comparable to the average hypothesis length of 8.3 tokens in Bowman et al. (2015).

3.2 Task 2 - Claim Labeling

The annotators were asked to label each individual claim generated during Task 1 as SUPPORTED, REFUTED or NOTENOUGHINFO. For the first two cases, the annotators were asked to find the evidence from any page that supports or refutes the claim (see Figure 2 for a screenshot of the interface). In order to encourage inter-annotator consistency, we gave the following general guideline:

[4] The annotation guidelines for both stages are provided in the supplementary materials

[5] These consisted of 5,000 from a Wikipedia ‘most accessed pages’ list and the pages hyperlinked from them.

> If I was given only the selected sentences, do I have strong reason to believe the claim is true (supported) or stronger reason to believe the claim is false (refuted). If I'm not certain, what additional information (dictionary) do I have to add to reach this conclusion.

In the annotation interface, all sentences from the introductory section of the page for the main entity of the claim and of every linked entity in those sentences were provided as a default source of evidence (left-hand side in Fig. 2). Using this interface the annotators recorded the sentences necessary to justify their classification decisions. In order to allow exploration beyond the main and linked pages, we also allowed annotators to add an arbitrary Wikipedia page by providing its URL and the system would add its introductory section as additional sentences that could be then selected as evidence (right-hand side in Fig. 2). The title of the page could also be used as evidence to resolve co-reference, but this decision was not explicitly recorded. We did not set a hard time limit for the task, but the annotators were advised not to spend more than 2-3 minutes per claim. The label NOTENOUGHINFO was used if the claim could not be supported or refuted by any amount of information in Wikipedia (either because it is too general, or too specific).

3.3 Annotators

The annotation team consisted of a total of 50 members, 25 of which were involved only in the first task. All annotators were native US English speakers and were trained either directly by the authors, or by experienced annotators. The interface for both tasks was developed by the authors in collaboration with an initial team of two annotators. Their notes and suggestions were incorporated into the annotation guidelines.

The majority of the feedback received from the annotators was very positive: they found the task engaging and challenging, and after the initial stages of annotation they had developed an understanding of the needs of the task which let them discuss solutions about edge cases as a group.

3.4 Data Validation

Given the complexity of the second task (claim labeling), we conducted three forms of data validation: 5-way inter-annotator agreement, agreement against *super-annotators* (defined in Section 3.4.2), and manual validation by the authors. The validation for claim generation was done implicitly during claim labeling. As a result 1.01% of claims were skipped, 2.11% contained typos and 6.63% of the generated claims were flagged as too vague/ambiguous and were excluded e.g. or "Sons of Anarchy premiered.".

3.4.1 5-way Agreement

We randomly selected 4% ($n = 7506$) of claims which were not skipped to be annotated by 5 annotators. We calculated the Fleiss κ score (Fleiss, 1971) to be 0.6841 which we consider encouraging given the complexity of the task. In comparison Bowman et al. (2015) reported a κ of 0.7 for a simpler task, since the annotators were given the premise/evidence to verify a hypothesis against without the additional task of finding it.

3.4.2 Agreement against *Super-Annotators*

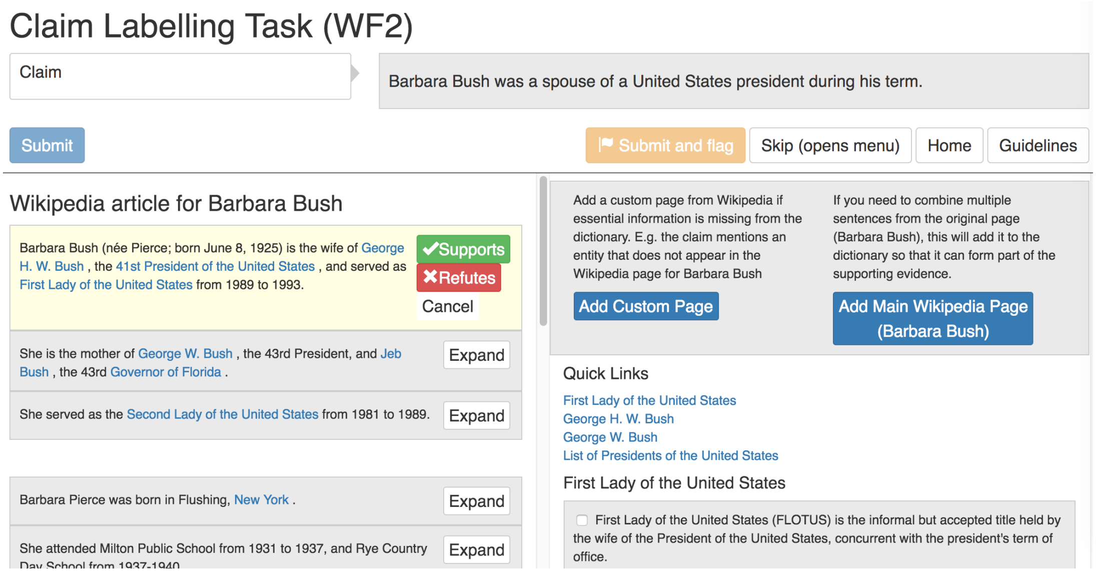
We randomly selected 1% of the data to be annotated by *super-annotators*: expert annotators with no suggested time restrictions. The purpose of this exercise was to provide as much coverage of evidence as possible. We instructed the *super-annotators* to search over the whole Wikipedia for every possible sentence that could be used as evidence. We compared the regular annotations against this set of evidence and the precision/recall was 95.42% and 72.36% respectively.

3.4.3 Validation by the Authors

As a final quality control step, we chose 227 examples and annotated them for accuracy of the labels and the evidence provided. We found that 91.2% of the examples were annotated correctly. 3% of the claims were mistakes in claim generation that had not been flagged during labeling. We found a similar number of these claims which did not meet the guidelines during a manual error analysis of the baseline system (Section 5.8).

3.4.4 Findings

When compared against the *super-annotators*, all except two annotators achieved $>$ 90% precision and all but 9 achieved recall $>$ 70% in evidence retrieval. The majority of the low-recall cases are for claims such as "Akshay Kumar is an actor." where the *super-annotator* added 34 sentences as evidence, most of them being filmography listings (e.g. "In 2000, he starred in the Priyadarshan-directed comedy Hera Pheri").


Claim Labelling Task (WF2)
Claim
Submit
Submit and flag
Skip (opens menu)
Home
Guidelines
Wikipedia article for Barbara Bush
✔Supports
✖Refutes
Cancel
Expand
Expand
Expand
Expand
Add Custom Page
Add Main Wikipedia Page (Barbara Bush)
Quick Links
First Lady of the United States
George H. W. Bush
George W. Bush
List of Presidents of the United States
First Lady of the United States

Barbara Bush was a spouse of a United States president during his term.

Barbara Bush (née Pierce; born June 8, 1925) is the wife of George H. W. Bush , the 41st President of the United States , and served as First Lady of the United States from 1989 to 1993.

She is the mother of George W. Bush , the 43rd President, and Jeb Bush , the 43rd Governor of Florida .

She served as the Second Lady of the United States from 1981 to 1989.

Barbara Pierce was born in Flushing, New York .

She attended Milton Public School from 1931 to 1937, and Rye Country Day School from 1937-1940

Add a custom page from Wikipedia if essential information is missing from the dictionary. E.g. the claim mentions an entity that does not appear in the Wikipedia page for Barbara Bush

If you need to combine multiple sentences from the original page (Barbara Bush), this will add it to the dictionary so that it can form part of the supporting evidence.

□ First Lady of the United States (FLOTUS) is the informal but accepted title held by the wife of the President of the United States, concurrent with the president's term of office.

Figure 2: Screenshot of Task 2 - Claim Labeling

During the validation by the authors, we found that most of the examples that were annotated incorrectly were cases where the label was correct, but the evidence selected was not sufficient (only 4 out of 227 examples were labeled incorrectly according to the guidelines).

We tried to resolve this issue by asking our annotators to err on the side of caution. For example, while the claim “Shakira is Canadian” could be labeled as REFUTED by the sentence “Shakira is a Colombian singer, songwriter, dancer, and record producer”, we advocated that unless more explicit evidence is provided (e.g. “She was denied Canadian citizenship”), the claim should be labeled as NOTENOUGHINFO, since dual citizenships are permitted, and the annotators’ world knowledge should not be factored in.

A related issue is entity resolution. For a claim like “David Beckham was with United.”, it might be trivial for an annotator to accept “David Beckham made his European League debut playing for Manchester United.” as supporting evidence. This implicitly assumes that “United” refers to “Manchester United”, however there are many Uniteds in Wikipedia and not just football clubs, e.g. United Airlines. The annotators knew the page of the main entity and thus it was relatively easy to resolve ambiguous entities. While we provide this information as part of the dataset, we argue that it should only be used for system training/development.

4 Baseline System Description

We construct a simple pipelined system comprising three components: document retrieval, sentence-level evidence selection and textual entailment. Each component is evaluated in isolation through oracle evaluations on the development set and we report the final accuracies on the test set.

Document Retrieval We use the document retrieval component from the DrQA system (Chen et al., 2017) which returns the k nearest documents for a query using cosine similarity between binned unigram and bigram Term Frequency-Inverse Document Frequency (TF-IDF) vectors.

Sentence Selection Our simple sentence selection method ranks sentences by TF-IDF similarity to the claim. We sort the most-similar sentences first and tune a cut-off using validation accuracy on the development set. We evaluate both DrQA and a simple unigram TF-IDF implementation to rank the sentences for selection. We further evaluate impact of sentence selection on the RTE module by predicting entailment given the original documents without sentence selection.

Recognizing Textual Entailment We compare two models for recognizing textual entailment. For a simple well-performing baseline, we selected Riedel et al. (2017)’s submission from the 2017 Fake News Challenge. It is a multi-layer perceptron (**MLP**) with a single hidden layer which uses term frequencies and TF-IDF cosine similarity between the claim and evidence as features.

Evaluating the state-of-the-art in RTE, we used a decomposable attention (**DA**) model between the claim and the evidence passage (Parikh et al., 2016). We selected it because at the time of development this model was the highest scoring system for the Stanford Natural Language Inference task (Bowman et al., 2015) with publicly available code that did not require the input text to be parsed syntactically, nor was an ensemble.

The RTE component must correctly classify a claim as NOTENOUGHINFO when the evidence retrieved is not relevant or informative. However, the instances labeled as NOTENOUGHINFO have no evidence annotated, thus cannot be used to train RTE for this class. To overcome this issue, we simulate training instances for the NOTENOUGHINFO through two methods: sampling a sentence from the nearest page (NEARESTP) to the claim as evidence using our document retrieval component and sampling a sentence from Wikipedia uniformly at random (RANDOMS).

5 Experiments

5.1 Dataset Statistics

We partitioned the annotated claims into training, development and test sets. We ensured that each Wikipedia page used to generate claims occurs in exactly one set. We reserved a further 19,998 examples for use as a test set for a shared task.

Split	SUPPORTED	REFUTED	NEI
Training	80,035	29,775	35,639
Dev	3,333	3,333	3,333
Test	3,333	3,333	3,333
Reserved	6,666	6,666	6,666

Table 1: Dataset split sizes for SUPPORTED, REFUTED and NOTENOUGHINFO (NEI) classes

5.2 Evaluation

Predicting whether a claim is SUPPORTED, REFUTED or NOTENOUGHINFO is a 3-way classification task that we evaluate using accuracy. In the case of the first two classes, appropriate evidence must be provided, at a sentence-level, to justify the classification. We consider an answer returned correct for the first two classes only if correct evidence is returned. Given that the development and test datasets have balanced class distributions, a random baseline will have $\sim 33\%$ accuracy if one ignores the requirement for evidence for SUPPORTED and REFUTED.

We evaluate the correctness of the evidence retrieved by computing the F_1-score of all the predicted sentences in comparison to the human-annotated sentences for those claims requiring evidence on our complete pipeline system (Section 5.7). As in Fig. 1, some claims require multi-hop inference involving sentences from more than one document to be correctly supported as SUPPORTED/REFUTED. In this case all sentences must be selected for the evidence to be marked as correct. We report this as the proportion of *fully supported claims*. Some claims may be equally supported by different pieces of evidence; in this case one complete set of sentences should be predicted.

Systems that select information that the annotators did not will be penalized in terms of precision. We recognize that it is not feasible to ensure that the evidence selection annotations are complete, nevertheless we argue that they are useful for automatic evaluation during system development. For a more reliable evaluation we advocate crowd-sourcing annotations of false-positive predictions at a later date in a similar manner to the TAC KBP Slot Filler Validation (Ellis et al., 2016).

5.3 Document Retrieval

The document retrieval component of the baseline system returns the k nearest documents to the claim using the DrQA (Chen et al., 2017) TF-IDF implementation to return the k-nearest documents. In the scenario where evidence from multiple documents is required, k must be greater than this figure. We simulate the upper bound in accuracy using an oracle 3-way RTE classifier that predicts SUPPORTED/REFUTED ones correctly only if the documents containing the supporting/refuting evidence are returned by document retrieval and always predicts NOTENOUGHINFO instances correctly independently of the evidence. Results are shown in Table 2.

5.4 Sentence Selection

Mirroring document retrieval, we extract the top l-most similar sentences from the k-most relevant documents using TF-IDF vector similarity. We modified document retrieval component of DrQA (Chen et al., 2017) to select sentences using bigram TF-IDF with binning and compared this to a simple unigram TF-IDF implementation using NLTK (Loper and Bird, 2002). Using the param-

k	Fully Supported (%)	Oracle Accuracy (%)
1	25.31	50.21
5	55.30	70.20
10	65.86	77.24
25	75.92	83.95
50	82.49	90.13
100	86.59	91.06

Table 2: Dev. set document retrieval evaluation.

eters $k = 5$ documents and $l = 5$ sentences, 55.30% of claims (excluding NOTENOUGHINFO) can be fully supported or refuted by the retrieved documents before sentence selection (see Table 2). After applying the sentence selection component, 44.22% of claims can be fully supported using the extracted sentences with DrQA and only 34.03% with NLTK. This would yield oracle accuracies of 62.81% and 56.02% respectively.

5.5 Recognizing Textual Entailment

The RTE component is trained on labeled claims paired with sentence-level. Where multiple sentences are required as evidence, the strings are concatenated. As discussed in Section 4, such data is not annotated for claims labeled NOTENOUGHINFO, thus we compare random sampling-based and similarity-based strategies for generating it. We evaluate classification accuracy on the development set in an oracle evaluation, assuming correct evidence sentences are selected (Table 3). Additionally, for the DA model, we predict entailment given evidence, using the AllenNLP (Gardner et al., 2017) pre-trained Stanford Natural Language Inference (SNLI) model for comparison.

Model	Accuracy (%)		
	NEARESTP	RANDOMS	SNLI
MLP	65.13	73.81	-
DA	80.82	88.00	38.54

Table 3: Oracle classification on claims in the development set using gold sentences as evidence

The random sampling (RANDOMS) approach (where a sentence is sampled at random from Wikipedia in place of evidence for claims labeled as NOTENOUGHINFO) yielded sentences that were not only semantically different to the claim, but also unrelated. While the the accuracy of models trained with sampling approach is higher in oracle evaluation setting, this may not yield a better system in the pipeline setting. In contrast, the nearest page (NEARESTP) method samples a sentence from the highest-ranked page returned by our document retrieval module. This simulates finding related information that may not be sufficient to support or refute a claim. We will evaluate both RANDOMS and NEARESTP in the full pipeline setting, but we will not pursue the SNLI-trained model further as it performed substantially worse.

5.6 Full Pipeline

The complete pipeline consists of the DrQA document retrieval module (Section 5.3), DrQA-based sentence retrieval module (Section 5.4), and the decomposable attention RTE model (Section 5.5). The two parameters: k, describing the number documents and l, describing the number sentences to return were found using grid-search optimizing the RTE accuracy with the DA model. For the pipeline, we set $k = 5$ and $l = 5$ and report the development set accuracy, both with and without the requirement to provide correct evidence for the SUPPORTED/REFUTED predictions (marked as **ScoreEv** and **NoScoreEv** respectively).

Model	Accuracy (%)	
	NoScoreEv	ScoreEv
MLP / NP	41.86	19.04
MLP / RS	40.63	19.42
DA / NP	52.09	**32.57**
DA / RS	50.37	23.53

Table 4: Full pipeline results on development set

The decomposable attention model trained with NEARESTP is the most accurate when evidence is considered. Inspection of the confusion matrices shows that the RANDOMS strategy harms recall for the NOTENOUGHINFO class. This is due to the difference between the sampled pages in the training set and the ones retrieved in the development set causing related but uninformative evidence to be misclassified as SUPPORTED and REFUTED.

Ablation of the sentence selection module We evaluate the impact of the sentence selection module on both the RTE accuracy by removing it.

While the sentence selection module may improve accuracy in the RTE component, it is discarding sentences that are required as evidence to support claims, harming performance (see Section 5.4). We assess the accuracies in both oracle setting (similar to Section 5.5) (see Table 5) as well as in the full pipeline (see Table 6).

In the oracle setting, the decomposable attention models are worst affected by removal of the sentence selection module: exhibiting an substantial decrease in accuracy. The NEARESTP training regime exhibits a 17% decrease and the RANDOMS accuracy decreases by 19%, despite near-perfect recall of the NOTENOUGHINFO class.

Model	**Oracle Accuracy (%)**	
	NEARESTP	RANDOMS
MLP	57.16	73.36
DA	63.68	69.05

Table 5: Oracle accuracy on claims in the dev. set using gold documents as evidence (c.f. Table 3).

In the pipeline setting, we run the RTE component without sentence selection using $k = 5$ most similar predicted documents. The removal of the sentence selection component decreased the accuracy (NOSCOREEV) approximately 10% for both decomposable attention models.

Model	**Accuracy (%)**	
	NEARESTP	RANDOMS
MLP	38.85	40.45
DA	41.57	40.62

Table 6: Pipeline accuracy on the dev. set without the sentence selection module (c.f. Table 4).

5.7 Evaluating Full Pipeline on Test Set

We evaluate our pipeline approach on the test set based on the results observed in Section 5.6. First, we use DrQA to select select 5 documents nearest to the claim. Then, we select 5 sentences using our DrQA-based sentence retrieval component and concatenate them. Finally, we predict entailment using the Decomposable Attention model trained with the NEARESTP strategy. The classification accuracy is 31.87%. Ignoring the requirement for correct evidence (**NoScoreEv**) the accuracy is 50.91%, which highlights that while the systems were predicting the correct label, the evidence selected was different to that which the human annotators chose. The recall of the document and sentence retrieval modules for claims which required evidence on the test set was 45.89% (considering complete groups of evidence) and the precision 10.79%. The resulting F_1 score is 17.47%.

5.8 Manual Error Analysis

Using the predictions on the test set, we sampled 961 of the predictions with an incorrect label or incorrect evidence and performed a manual analysis. Of these, 28.51% ($n = 274$) had the correct predicted label but did not satisfy the requirements for evidence. The information retrieval component of the pipeline failed to identify any correct evidence in 58.27% ($n = 560$) of cases which accounted for the large disparity between accuracy of the system when evidence was and was not considered. Where suitable evidence was found, the RTE component incorrectly classified 13.84% ($n = 133$) of claims.

The pipeline retrieved new evidence that had not been identified by annotators in 21.85% ($n = 210$) of claims. This was in-line with our expectation given the measured recall rate of annotators (see Section 3.4.2), who achieved recall of 72.36% of evidence identified by the super-annotators.

We found that 4.05% ($n = 41$) of claims did not meet our guidelines. Of these, there were 11 claims which could be checked without evidence as these either tautologous or self-contradictory. Some correct claims appeared ungrammatical due to the mis-parsing of named entities (e.g. *Exotic Birds* is the name of a band but could be parsed as a type of animal). Annotator errors (where the wrong label was applied) were present in 1.35% ($n = 13$) of incorrectly classified claims.

Interestingly, our system found new evidence that contradicted the gold evidence in 0.52% ($n = 5$) of cases. This was caused either by entity resolution errors or by inconsistent information present in Wikipedia pages (e.g. Pakistan was described as having both the 41st and 42nd largest GDP in two different pages).

5.9 Ablation of Training Data

To evaluate whether the size of the dataset is suitable for training the RTE component of the pipeline, we plot the learning curves for the DA and MLP models (Fig. 3). For each model, we trained 5 models with different random initializations using the NEARESTP method (see Section 5.5). We selected the highest performing model when evaluated on development set and report the oracle RTE accuracy on the test set. We observe that with fewer than 6000 training instances, the accuracy of DA is unstable. However, with more data, its accuracy increases with respect to the log of the number of training instances and exceeds that of MLP. While both learning curves exhibit the typical diminishing return trends, they indicate that the dataset is large enough to demonstrate the differences of models with different learning capabilities.

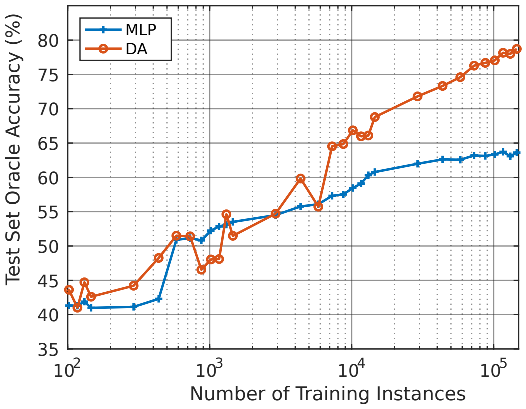


Figure 3: Learning curves for the RTE models.

6 Discussion

The pipeline presented and evaluated in the previous section is one possible approach to the task proposed in our dataset, but we envisage different ones to be equally valid and possibly better performing. For instance, it would be interesting to test how approaches similar to natural logic inference (Angeli and Manning, 2014) can be applied, where a knowledge base/graph is constructed by reading the textual sources and then a reasoning process over the claim is applied, possibly using recent advances in neural theorem proving (Rocktäschel and Riedel, 2017). A different approach could be to consider a combination of question generation (Heilman and Smith, 2010) followed by a question answering model such as BiDAF (Seo et al., 2016), possibly requiring modification as they are designed to select a single span of text from a document rather than return one or more sentences as per our scoring criteria. The sentence-level evidence annotation in our dataset will help develop models selecting and attending to the relevant information from multiple documents and non-contiguous passages. Not only will this enhance the interpretability of predictions, but also facilitate the development of new methods for reading comprehension.

Another use case for the FEVER dataset is claim extraction: generating short concise textual facts from longer encyclopedic texts. For sources like Wikipedia or news articles, the sentences can contain multiple individual claims, making them not only difficult to parse, but also hard to evaluate against evidence. During the construction on the FEVER dataset, we allowed for an extension of the task where simple claims can be extracted from multiple complex sentences.

Finally, we would like to note that while we chose Wikipedia as our textual source, we do not consider it to be the only source of information worth considering in verification, hence not using TRUE or FALSE in our classification scheme. We expect systems developed on the dataset presented to be portable to different textual sources.

7 Conclusions

In this paper we have introduced FEVER, a publicly available dataset for fact extraction and verification against textual sources. We discussed the data collection and annotation methods and shared some of the insights obtained during the annotation process that we hope will be useful to other large-scale annotation efforts.

In order to evaluate the challenge this dataset presents, we developed a pipeline approach that comprises information retrieval and textual entailment components. We showed that the task is challenging yet feasible, with the best performing system achieving an accuracy of 31.87%.

We also discussed other uses for the FEVER dataset and presented some further extensions that we would like to work on in the future. We believe that FEVER will provide a stimulating challenge for claim extraction and verification systems.

Acknowledgments

The work reported was partly conducted while James Thorne was at Amazon Research Cambridge. Andreas Vlachos is supported by the EU H2020 SUMMA project (grant agreement number 688139). The authors would like to thank the team of annotators involved in preparing this dataset.

References

Gabor Angeli and Christopher D. Manning. 2014. NaturalLI: Natural logic inference for common sense reasoning. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*. pages 534–545.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, pages 1870–1879. https://doi.org/10.18653/v1/P17-1171.

Ido Dagan, Bill Dolan, Bernardo Magnini, and Dan Roth. 2009. Recognizing textual entailment: Rational, evaluation and approaches. *Natural Language Engineering* 15(4):i–xvii. https://doi.org/10.1017/S1351324909990209.

Joe Ellis, Jeremy Getman, Dana Fore, Neil Kuster, Zhiyi Song, Ann Bies, and Stephanie Strassel. 2016. Overview of Linguistic Resources for the TAC KBP 2016 Evaluations : Methodologies and Results. *Proceedings of TAC KBP 2016 Workshop, National Institute of Standards and Technology, Maryland, USA* (Ldc).

William Ferreira and Andreas Vlachos. 2016. Emergent: a novel data-set for stance classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. San Diego, California, pages 1163–1168.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin* 76(5):378.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. 2017. AllenNLP: A Deep Semantic Natural Language Processing Platform .

Michael Heilman and Noah A. Smith. 2010. Good Question! statistical ranking for question generation. In *Proceedings of the 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. pages 609–617.

Mio Kobayashi, Ai Ishii, Chikara Hoshino, Hiroshi Miyashita, and Takuya Matsuzaki. 2017. Automated historical fact-checking by passage retrieval, word statistics, and virtual question-answering. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. volume 1, pages 967–975.

Edward Loper and Steven Bird. 2002. Nltk: The natural language toolkit. In *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics - Volume 1*. Association for Computational Linguistics, Stroudsburg, PA, USA, ETMTNLP '02, pages 63–70. https://doi.org/10.3115/1118108.1118117.

Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *ACL (System Demonstrations)*. pages 55–60.

Mausam, Michael Schmitz, Robert Bart, Stephen Soderland, and Oren Etzioni. 2012. Open language learning for information extraction. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. pages 523–534.

Ankur Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. 2016. A decomposable attention model for natural language inference. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Austin, Texas, pages 2249–2255. https://aclweb.org/anthology/D16-1244.

Dean Pomerleau and Delip Rao. 2017. Fake news challenge. `http://fakenewschallenge.org/`.

P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Benjamin Riedel, Isabelle Augenstein, George Spithourakis, and Sebastian Riedel. 2017. A simple but tough-to-beat baseline for the Fake News Challenge stance detection task. *CoRR* abs/1707.03264. http://arxiv.org/abs/1707.03264.

Tim Rocktäschel and Sebastian Riedel. 2017. End-to-end differentiable proving. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, California, United States*. volume abs/1705.11040. http://arxiv.org/abs/1705.11040.

Álvaro Rodrigo, Anselmo Peñas, and Felisa Verdejo. 2009. Overview of the answer validation exercise 2008. In Carol Peters, Thomas Deselaers, Nicola Ferro, Julio Gonzalo, Gareth J. F. Jones, Mikko Kurimo, Thomas Mandl, Anselmo Peñas, and Vivien Petras, editors, *Evaluating Systems for Multilingual and Multimodal Information Access: 9th Workshop of the Cross-Language Evaluation Forum*. pages 296–313.

Min Joon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2016. Bidirectional attention flow for machine comprehension. *CoRR* abs/1611.01603. http://arxiv.org/abs/1611.01603.

Craig Silverman. 2015. Lies, Damn Lies and Viral Content. http://towcenter.org/research/lies-damn-lies-and-viral-content/.

Niket Tandon, Gerard de Melo, and Gerhard Weikum. 2011. Deriving a Web-scale common sense fact database. In *Proceedings of the 25th AAAI Conference on Artificial Intelligence (AAAI 2011)*. AAAI Press, Palo Alto, CA, USA, pages 152–157.

Andreas Vlachos and Sebastian Riedel. 2014. Fact checking: Task definition and dataset construction. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*. ACL. http://www.aclweb.org/anthology/W14-2508.

William Yang Wang. 2017. "Liar, Liar Pants on Fire": A new benchmark dataset for fake news detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. http://aclweb.org/anthology/P17-2067.