Identifying Opposing Views in Online Discourse

by

Muuo Wambua

A research paper submitted in conformity with the requirements
for the degree of Master of Science
Graduate Department of Computer Science
University of Toronto

Abstract

Discourse in online media often takes the form of siloed discussions where unpopular views tend to get drowned out by the majority. This is especially true in platforms such as Reddit, where only the most popular opinions get visibility. In this work, we propose that this problem may be solved by building tools that surface opposing views and I advance a means for automatically identifying disagreeing views on Reddit.

Acknowledgements

I'd like to first of all acknowledge the computer science community for being so generous with its knowledge and resources over the years, allowing people such as myself to immerse themselves into this exciting field at very little cost.

I'd also like to acknowledge my advisors Graeme Hirst and Frank Rudzicz for being similarly liberal with their guidance. Graeme, for constantly nudging me to think critically about my ideas and teaching me to put careful thought into how I present them. Frank, for being an infinite source of great ideas and living evidence that you can get everything done.

Last, I'd like to thank everyone who made my time with the department enjoyable: Gagandeep Singh, Sean Robertson, Chloe Pouprom, Demestres Kostas, Arvie Frydenlund, Siavash Kazemian, Patricia Thaine and the rest of BA4202's occupants; all the interesting people I met during CSGSBS cookiebreaks; and my family and friends in Kenya who provided support from afar.

This research would not have been possible without the financial support of the Natural Sciences and Engineering Research Council of Canada and the University of Toronto.

1 Introduction

The Internet is a great resource, giving people access to information that would be otherwise inaccessible. While a good portion of this information is provided by experts who take care in articulating their views, the proliferation of social media and the reduced complexity of online content creation have allowed the ordinary person to voice his or her opinion on a broad spectrum of topics. While it can be helpful to know that many people agree on a topic, readers are often only presented with one side of the issue (Del Vicario et al., 2015). While some work has been done towards recognizing the absence of opposing arguments (Stab and Gurevych, 2016) and detecting controversial topics (Dori-Hacohen and Allan, 2015; Jang and Allan, 2016), we still do not have access to information retrieval systems that are aware of and inform us of the existence of views that are contradictory to whichever ones are most accessible. I believe that such tools will present users with balanced views on controversial topics and will allow them to make better informed decisions.

In this work, I describe a means of automatically identifying opposing views in online text. This system does this by labeling document pairs as either agreeing or disagreeing. We make an effort to focus on the semantic content of the documents rather than more obvious lexical cues such as explicit negation. This differs from other approaches to identifying stance in tweets (Mohammad et al., 2016), including those of Zarrella and Marsh (2016) and Augenstein et al. (2016). Specifically, our system is topic independent, and identifies stance with respect to individual documents rather than the documents' stance on a broad topic. Because, this work focuses on social platforms such as Reddit our documents take the form of comments.

While our intent in identifying opposing opinions is positive, it is important to note that the same methods have the potential for more malicious use. These methods can be used to suppress opinions that run contrary to states and corporations that have the ability to censor them (Thomet, 2018).

2 Background and Related Work

2.1 Inferring Agreement

Inference of agreement can be viewed as an extension of the task of recognizing textual entailment (Berant et al., 2011), the task of determining if a given premise entails a given hypothesis. That is, without additional context, a rational human would likely infer a hypothesis is true if the premise is true. While agreement can sometimes be implied by the absence of contradiction e.g. an instance where two people provide different reasons for why they dislike some object, it is simpler to focus on instances where agreeing arguments assert the same relationships or properties, and thus can be entailed by one another. Alternative approaches, such as those presented by Skeppstedt et al. (2016), instead infer agreement and disagreement by detecting the polarity of a document towards another. However, this is difficult to apply to online media where authors often fail to provide explicit citation to other work.

State-of-the-art approaches to the problem of recognizing textual entailment have managed remarkable performance using both statistical methods (Bowman et al., 2016) and logical inference (Angeli et al., 2016). However, the majority of the existing work has focused on recognizing textual entailment between sentence pairs. To infer entailment between arguments spanning the length of a document, we must find ways of reasoning over the entire document or find ways of reducing the documents to shorter segments of text that can be reasoned over using existing methodology.

Early attempts at solving the entailment problem, such as Bos and Markert (2005)'s solution, explored the use of formal logical inference. However, these methods were reliant on full semantic interpretation, that is, the conversion of statements in natural language to a formal meaning representation. Owing to the complex and dynamic nature of natural language, algorithms capable of accurately and consistently performing full semantic interpretation have remained elusive. Other authors have instead chosen to focus on the use of "shallow" measures of syntactic and semantic similarity (MacCartney et al., 2006; Hickl et al., 2006), which led to the works of MacCartney and Manning (2007, 2008, 2009) that propose a compromise between full semantic interpretation and shallow reasoning that they term as Natural Logic.

In the recent past, Bowman (2013) and Bowman et al. (2015b) have the explored the use of fixed-length representations of sentences to solve logical inference problems. They demonstrate through the use of tree-structured recurrent neural networks (Goller and Küchler, 1996) that the appropriate composition of these vector representations results in systems that are able to automatically infer textual

entailment with high accuracy. The supervised methods used to derive these vector representations are reliant on the existence of large corpora of human-annotated examples such as one produced by Bowman et al. (2015a) which consists of sentence pairs derived from image captions, labelled as entailing, contradictory, or neutral to each other.

2.2 Inferring Disagreement

Sadly, detecting contradiction is not as simple as finding cases of non-entailment (de Marneffe et al., 2008). A case of non-entailment could simply imply negation or neutrality. While mismatching information between sentences is often a good cue of non-entailment, it is not sufficient for contradiction detection which requires more precise comprehension of the consequences of sentences. De Marneffe et al. (2008) provide a formal definition of the problem, enumerate different types of contradictions and argue for the consideration of event coreference when assessing sentence pairs for contradiction. Ritter et al. (2008) extend their work by recognizing the existence of many seeming contradictions that are only unearthed through the use of background knowledge. Similar to early attempts at solving the textual entailment problem, their methodology is also reliant on the existence of accurate information extraction systems and human-annotated knowledge-bases.

While Bowman et al. (2015b) show that it is possible to use vector representations of sentences to infer contradiction between them, they limit themselves to evaluating contradiction at the sentence level. Li et al. (2017) argue that traditional context-based word embeddings are not powerful enough for contradiction detection because contrasting words share similar context and will therefore be mapped to representations that are closer together in vector space. They therefore develop a neural-network architecture tailored to learn contradiction-specific word embeddings by minimizing the semantic gap between entailing pairs and maximizing the gap between contradicting pairs of sentences.

2.3 Neural Natural Language Inference Models

The task of finding textual entailment and contradiction can also be termed as problems of natural language inference (NLI), in which the goal is to determine the core inferential relationship between two sentences. MacCartney (2009) defined NLI as the problem of determining whether a natural language hypothesis h can be reasonably be inferred from a natural language premise p.

A key feature that distinguishes NLI from other forms of logical inference is that

the problem inputs are expressed in natural language, whereas research in knowledge representation and methods for automated deduction typically assume that the inputs already take the form of some formal representation such as first-order logic. As such, a lot of the early work in NLI revolved around automatically translating sentences expressed in natural language to formal representations.

MacCartney (2009) presented a pipeline of tools that produce alignments between p and h and lexical relation information that the core logic requires to make its inference. This approach, and the work done to follow it up (Watanabe et al., 2012; Angeli et al., 2016) found mixed results, with models favoring precision at the expense of recall.

In an attempt to remedy that problem, recent work in NLI has placed an emphasis on using probabilistic methods (most employing some form of artificial neural networks) to both encode the meaning of sentences and infer the relationships between them. This paradigm has helped researchers achieve state-of-the-art results on large corpora such as the Stanford NLI corpus (Bowman et al., 2015a) and the Multi-Genre NLI corpus (Williams et al., 2017).

2.3.1 Vector Representations of Words and Sentences

In the same way that symbolic logical inference methods as a first step require sentences to be encoded using formal representations, the use of artificial neural networks (ANN) requires us to encode our sentences as vectors.

The task of encoding words as vectors is straightforward, where the simplest methods involve mapping each word (or word part) to a vector of some preselected length. Whenever a word is used as an input to an ANN, the corresponding vector is looked up and used to represent it. These vectors can be created using distributional methods (Pennington et al., 2014; Mikolov et al., 2013) or learned alongside other parameters during training (Bengio et al., 2003).

However, finding fixed length representations for sentences often proves more difficult owing to the combinatorial explosion of the number of possible sentences given a certain vocabulary. Most methods therefore focus on efficiently finding these representations on the fly, rather than using a giant lookup-table.

Continuous Bag of Words The simplest of such techniques involves looking up vector representations of each of the words in the sentences, and using their sum or average as a vector representation of the sentence. This strategy, referred to as Continuous Bag of Words (CBOW) has been shown to be effective for a number of tasks (Wieting et al., 2015; Adi et al., 2016), but fails to recover any information

about word order outside of any latent properties of language possibly embedded in the individual word vector representations.

Long Short-Term Memory Networks Other methods instead include a special learned neural network component that takes in a sequence of words and outputs a fixed-length vector. A popular family of such models is recurrent neural networks (RNNs) (Sutskever, 2013), including long short-term memory (LSTM) networks (Hochreiter and Schmidhuber, 1997) where the input is fed into the network in sequential order (from left to right, or the reverse). Bowman et al. (2015a) demonstrated that LSTM-encoded representations could outperform unigram and bigram-based features on NLI tasks. Liu et al. (2016) and Tai et al. (2015) went on to show that bidirectional LSTMs, and tree-structured LSTMs could offer a better way of generating sentence encodings for semantic tasks.

Decomposable Attention Models In recent work, Parikh et al. (2016) bypass sentence encoding entirely, and instead build representations of the entire sentence pair without ever constructing representations for the constituent sentences. They do this by using soft attention (Bahdanau et al., 2014) to align words in p and h, and aggregate the result of comparing the phrases to their soft-alignments.

Chen et al. (2016) extend their work by using various LSTM networks to generate vector representations of the phrases' constituent words that take into account word order.

3 Datasets

This section describes the two datasets used in this work. The first, a large standard NLI corpus, was used to provide a benchmark for our methodology. The second consisted of arguments collected from Reddit, and was used to train a model capable of inferring agreement and disagreement between comments posted to the platform.

3.1 Stanford Natural Language Inference Corpus

The SNLI corpus (Bowman et al., 2015a) is a collection of 570,000 human-written English sentence pairs manually labelled for balanced classification with the labels entailment, contradiction, and neutral, supporting the task of NLI. At the time of its creation, it was two orders of magnitude larger than any similar corpus, and allowed neural network-based methods to perform well on natural language inference tasks for the first time.

Being a large and high-quality corpus whose sentences and labels were written by humans in a grounded, naturalistic context, this corpus provides us with a standard way of testing the addition of syntactic information to decomposable attention models described in chapter 2.

It is worth noting that Williams et al. (2017) have since created a similar, multi-genre corpus to complement the SNLI corpus. However, the modified decomposable attention models were not evaluated against it in this work.

3.2 Political Arguments from Reddit

Reddit is a social news aggregation and discussion website where members submit content such as links, text posts, and images, which are then voted up or down by other members. Posts are organized by subject into user-created areas of interest called "subreddits", which cover a variety of topics including news, science, movies, video games, music, books, fitness, food, and image-sharing. Submissions with more up-votes appear towards the top of their subreddit and, if they receive enough votes, ultimately on the site's front page.

The experiments described in this work used user comments posted on the `r/politics` subreddit between December 2005 and August 2017[1]. There were a number of reasons for using the `r/politics` subreddit:

- `r/politics` is a subreddit for the discussion of current and explicitly political U.S. news. The community attempts to cater for the needs and interests of

[1]Retrieved from Jason Baumgartner's `pushshift.io`

the full spectrum of political leanings. This results in discussions that include the views of left-, centre-, and right-leaning members. This yields plenty of instances of polar arguments when commenting users have differing opinions on various political issues.

- As a result of strict contribution guidelines and perhaps a more mature community, the discussions on `r/politics` tend to more civil and coherent compared to a lot of the communities on Reddit. This reduces instances of deliberately unintelligible and nonsensical comments.

- Politics attracts interest from a larger segment of the population compared to more esoteric topics. The `r/politics` subreddit has approximately 3.6 million subscribers, and nearly perpetual active discussion on the most current news stories, resulting in a total of 61 million comments over the course of 12 years.

3.2.1 Extracting Comment Pairs

The pairs of agreeing and disagreeing comments were selected to meet the following criteria: i) inferring the comments' meaning should require as little context as possible; ii) the pairs should express similar or polar ideas; iii) they should, however, not express explicit approval or disapproval.

Each selected pair consisted of a 'parent' comment and another made in response to it. We selected only pairs with top-level comments, that is, where the parent comment was made in response to the main post. This was made under the assumption that the context required to understand a response increases with each level of hierarchy. An example of the comment hierarchy is provided in figure 1.

The *agreement* and *disagreement* of the comments was ensured by selecting pairs where the response started with an expression of explicit agreement or disagreement. Abbott et al. (2011) show that discourse markers such as "Right" and "No" can be strong indicators of agreement and disagreement respectively. Examples of such response prefixes are provided in table 1. Note that while contrasting conjunctions appended to agreeing prefixes (such as "I agree, but") can be good indicators of disagreement, many users used them as a segue into a related but neutral argument. As such, these prefixes were not considered.

We ignore discourse markers that occur mid-comment in an effort to isolate comments that, without their prefixes, do not explicitly express agreement/disagreement. Thus, we can arrive at our final comment pair by removing the prefixes from each response.

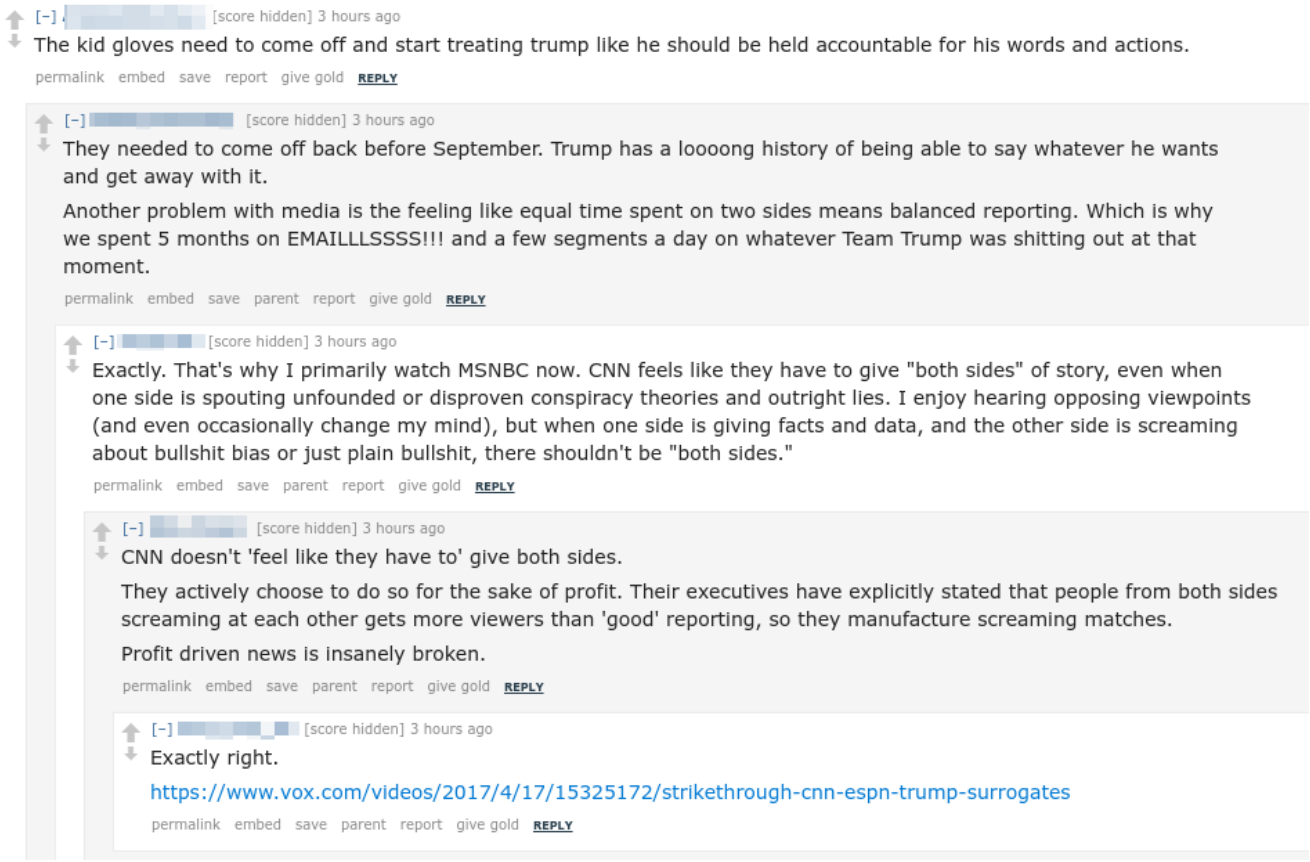


Figure 1: Example of a discussion on Reddit

Table 1: Examples of explicit expression of agreement/disagreement

	Prefixes
agreement	I agree, Agreed, I concur, I feel you, Absolutely, Exactly, ...
disagreement	No, Well no, Wrong, Incorrect, False, You're wrong, ...

Table 2: Examples of comments collected from `r/politics`

Parent Comment:	Maybe a draft is a good thing. Nothing would wake up Americans more than receiving orders to report to the nearest recruiting depot.
Child Comment:	I think that the draft is actually one of the greatest catalysts of the strength of the Sixties peace movements; Look at the usual Sixties activist: middle-class, educated, white. Honestly, I think that it is exactly for this reason that the United States is wary of enacting a draft, because it would suddenly snap all the young adults out of their MTV-induced trances in a bout of self-entitled indignation.
Polarity:	Agree
Parent Comment:	Also note: the number of hospital closures and layoffs in Red States.,Those layoffs are hurting the patients - as the GOP refuse to accept ACA.
Child Comment:	The ACA creates demand for health services. Denying the medicaid expansion just kills that.
Polarity:	Agree
Parent Comment:	These are the same polls that had Hilary in the lead by a lot....
Child Comment:	those polls were from before the election. That was nearly a year ago now. Did you just awake from a coma or something?
Polarity:	Disagree
Parent Comment:	"Unbend that knee! Play some football! Thats what they're paying you for. Stop politicizing everything that the Alt-Left wants. Just play some football. The fans have spoken and their paychecks will suffer if they keep it up.
Child Comment:	You can't tell them what to do and neither can Trump.
Polarity:	Disagree

3.2.2 Evaluating Corpus Quality

Given that the corpus was constructed automatically, it was deemed necessary to have a human annotator evaluate the quality of the extracted pairs. We therefore randomly selected 100 sentence pairs from the corpus, and had an annotator indicate whether the sentences agreed or disagreed. We found that the annotator agreed on the labels for **88%** of the sampled data with a Cohen's kappa score of **0.76**.

The most common reasons for disagreement between the human and automatic annotators are listed below:

- During comment extraction, care was taken to ignore comment pairs where the initial comment was a question, and the response began with 'No'. However, this filtering failed to account for comments that ended with assertions that did not warrant an explicit 'Yes/No' response but posed a question in one of the preceding sentences.

  This led to instances of comment pairs where the respondent would begin their comment with 'No' (as a response to the initial question), and follow up with comments that agreed with the parent's assertions.

- Because the comments in the chosen subreddit were mostly made in response to 'then-current' news-articles, the commenters often had access to *contextual information* that the human annotator did not have access to. This led to a number of cases where the annotator labelled the pairs incorrectly either because they chose randomly or made false assumptions.

- A number of comments were *sarcastic*. This led the comment-pair extractor into assuming that the expressions explicit agreement and disagreement in the child comments were indicative of their actual position, while in fact they were not.

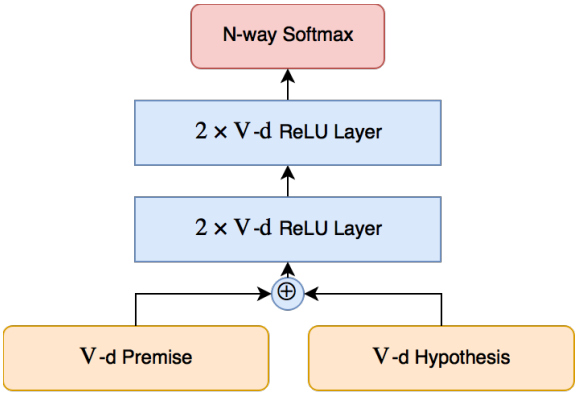


Figure 2: Neural Network's final feed-forward stage H.

4 Experiments

In this chapter I describe my attempts to find models capable of predicting the relationships between the premise p and hypothesis h for each of the datasets in chapter 3. Because of its relative simplicity, I used a feed-forward neural network with a CBOW sentence encoding as a baseline model. I then compared this baseline to architectures that used LSTMs and decomposable attention models to construct their sentence encodings (as described in section 2.3 and 4.1).

4.1 Model Architectures

All of the models described in this section produce two *V-dimensional* vectors v_1 and v_2, each being a fixed-length vector representation of the premise and hypothesis sentences respectively. After generating these representations, the two vectors were concatenated and fed into a two-layer feed-forward neural network with ReLU activations (Glorot et al., 2011). A final softmax layer outputs a score for each of the N possible relationships between the premise and hypothesis. This final stage, illustrated in figure 2 can be described by the function H, where:

$$\hat{y} = H([v_1, v_2]) \tag{1}$$

All of the models described also used 300 dimensional GloVe embeddings (Pennington et al., 2014) to represent each of the words. Each embedding vector was normalized to have an ℓ_2 norm of 1 and projected down to V dimensions. In a manner similar to Parikh et al. (2016), out-of-vocabulary (OOV) words were each assigned random embeddings each initialized to mean 0 and standard deviation 1. All

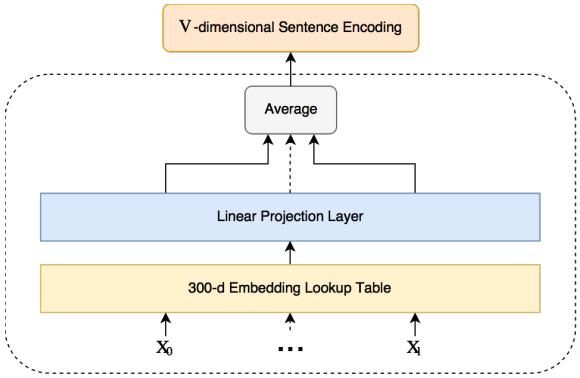


Figure 3: Sentence encoding using CBOW.

embeddings remained fixed during training, but the projection matrix was trained. All other parameter weights (hidden layers etc.) were initialized using a Xavier intialization (Glorot and Bengio, 2010).

During training, we used multi-class cross-entropy loss with dropout regularization, and hyperparameter search was done to find the best performing values of projected vector size V, learning rate, and dropout probability.

4.1.1 Continuous Bag of Words (CBOW)

This sentence encoding stage, illustrated in figure 3 was straightforward to implement. Each sentence's projected word-embeddings were averaged to produce the vectors v_1 and v_2 that were fed into the final stage H.

4.1.2 Long Short-Term Memory (LSTM)

This sentence encoder uses a sequence of LSTM cells (illustrated in figure 4) to generate a sentence encoding that takes into account word order and word count. Each LSTM cell takes, as an input, the embedding of the current word x_t, and the output h_{t-1} and the hidden state C_{t-1} corresponding to the previous word/cell.

The cell's state C_t and output h_t are then updated using the expressions provided in equation 2. W_f, b_f, W_i, b_i, W_c, b_c, W_o, and b_o are parameters learned during

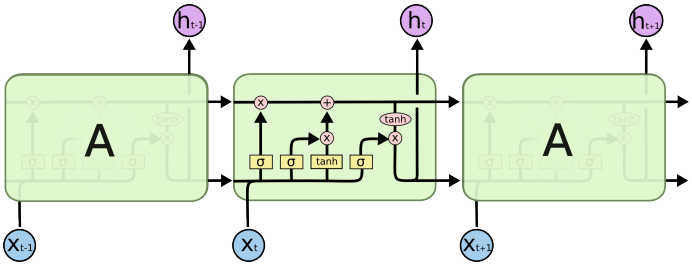


Figure 4: Illustration of a chain of LSTM cells.[2]

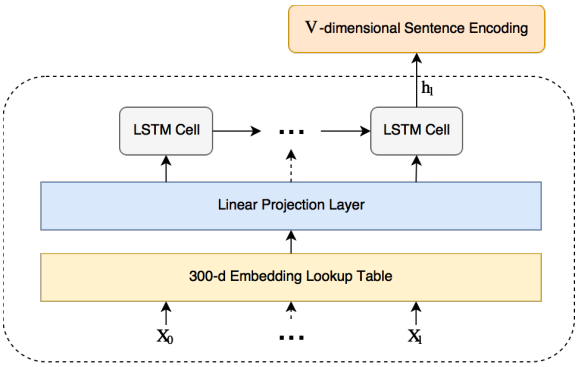


Figure 5: Sentence encoding using LSTM network.

training, and the $\circ$ operator denotes an elementwise multiplication.

$$
\begin{aligned}
f_t &= \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \\
i_t &= \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \\
C &= tanh(W_C \cdot [h_{t-1}, x_t] + b_c) \\
C &= f_t \circ C_{t-1} + i_t \circ C \\
o_t &= \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \\
h_t &= o_t \circ tanh(C_t)
\end{aligned}
\tag{2}
$$

As illustrated in figure 5, embeddings $x_0, x_1, ..., x_l$ each corresponding to a word in a sentence of length l are fed into a chain of LSTM cells, and the output h_l of the final cell is used as the sentence encoding v_i.

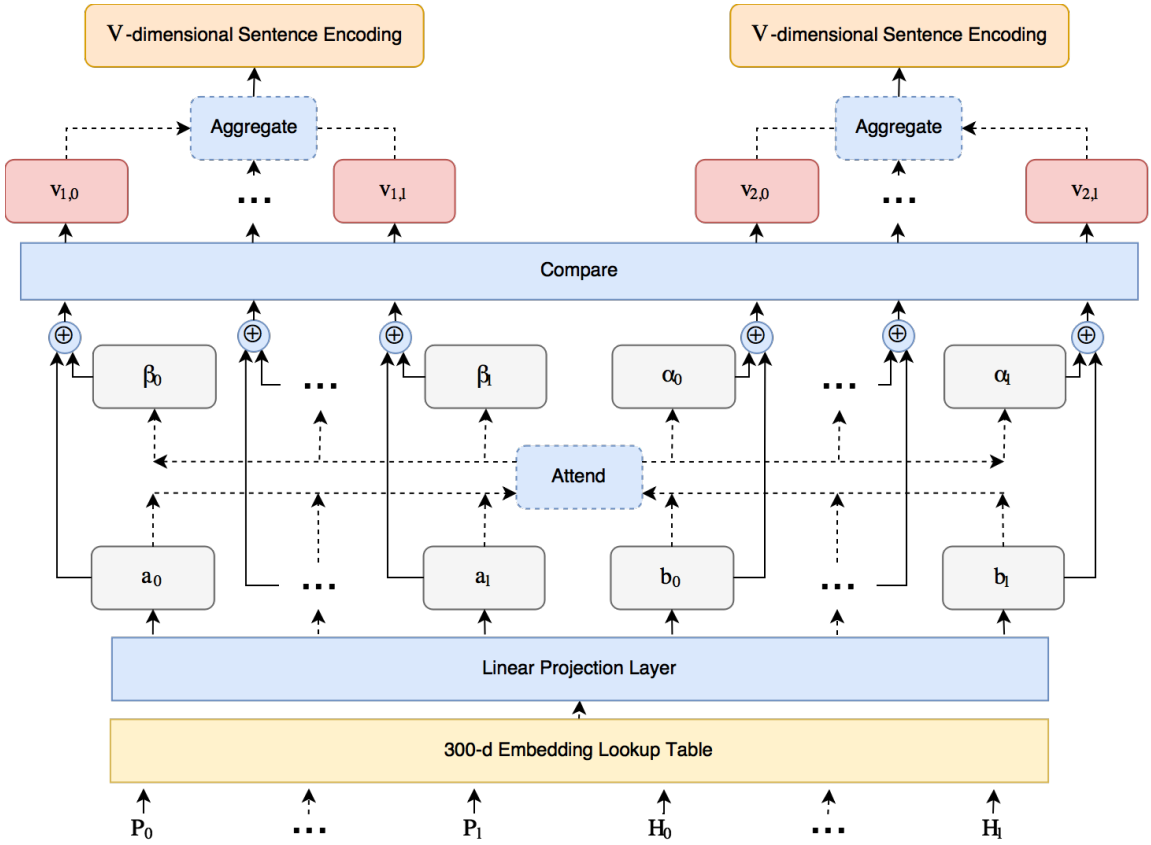


Figure 6: Sentence encoding using Parikh et al. (2016)'s decomposable attention model.

4.1.3 Decomposable Attention

Decomposable attention models use attention to break natural language inference problems into subproblems that can be solved separately, making the problem trivially parallelizable. It's worth noting that this model assumes that in NLI it it suffices to simply align bits of local text substructure and then aggregate this information. The model therefore loses a lot of information about the individual sentences' local substructure, but when Parikh et al. (2016) first published their work they were able to achieve state-of-the-art accuracy on the SNLI corpus with an order of magnitude fewer parameters than comparable models.

Attend The model's first step is to use Attention to create a soft alignment between the embedded representation of the premise and hypothesis, $\mathbf{a} = \{a_0, a_1, ...a_{\ell_a}\}$ and $\mathbf{b} = \{b_0, b_1, ...b_{\ell_b}\}$ respectively. ℓ_a and ℓ_b denote the number of tokens in each of the premise and hypothesis.

This step involves first computing unnormalized attention weights e_{ij} using a function $F^{'}$ which is decomposed to an expression involving F, a feed-forward neural network with ReLU activations. This is expressed in equation 3.

$$e_{ij} = F^{'}(\mathbf{a}, \mathbf{b}) = F(\mathbf{a})^T F(\mathbf{b}) \tag{3}$$

The attention weights are then normalized as shown in equation 4 to produce β_i (the subphrase in $\mathbf{a}$ that is softly aligned to b_i) and α_i (the subphrase in $\mathbf{b}$ that is softly aligned to a_i).

$$\begin{aligned} \beta_i &= \sum_{j=1}^{\ell_b} \frac{\exp(e^{ij})}{\sum_{k=1}^{\ell_b} \exp(e^{ik})} \mathbf{b_j} \\ \alpha_i &= \sum_{i=1}^{\ell_a} \frac{\exp(e^{ij})}{\sum_{k=1}^{\ell_a} \exp(e^{kj})} \mathbf{a_i} \end{aligned} \tag{4}$$

Compare The aligned phrases are then compared by feeding a concatentation of each pair through a function G which is also a feed-forward neural network with ReLU activations:

$$\begin{aligned} \mathbf{v}_{1,i} &= G([\mathbf{a_i}, \beta_i]) \\ \mathbf{v}_{2,i} &= G([\mathbf{b_i}, \alpha_i]) \end{aligned} \tag{5}$$

[2]Illustration copied from Christopher Olah (`http://colah.github.io/posts/2015-08-Understanding-LSTMs/`)

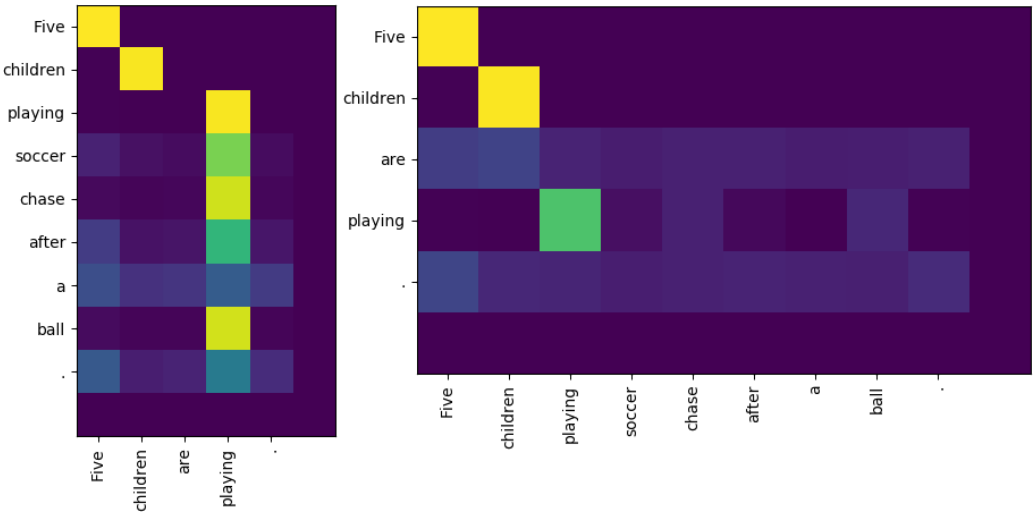


Figure 7: Alignments α and β between the sentences "Five children playing soccer chase after a ball" and "Five children are playing.". Darker values indicate values closer to zero.

Aggregate The two sets of comparison vectors are then aggregated by summation resulting in vectors $\mathbf{v_1}$ and $\mathbf{v_2}$ that encode the meaning shared between each sentence and its companion:

$$\begin{aligned} \mathbf{v}_1 &= \sum_{i=1}^{\ell_a} \mathbf{v}_{1,i} \\ \mathbf{v}_2 &= \sum_{i=1}^{\ell_b} \mathbf{v}_{2,i} \end{aligned} \tag{6}$$

An added advantage of using attention mechanisms to solve NLI problems is that values of the soft alignments provide insight into the model's ability to correctly align subphrases in the premise and hypothesis that may contribute to their semantic relationship. An example of such an alignment is illustrated in figure 7.

4.2 Using Grammatical Information in Neural Natural Language Inference Models

While the methods described above do a reasonably good job at encoding sentence meaning, they all fail to factor in what we know about the syntactic and semantic

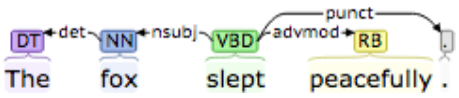


Figure 8: A dependency parse of "The fox slept peacefully.".

structure of language. Two particularly useful indicators of grammatical structure are parts of speech and typed dependencies.

While parts of speech indicate the grammatical function of individual words, typed dependencies provide a simple indication of the grammatical relationship between words in a sentence. Most forms of dependency relationships are provided in a format similar to those recommended by de Marneffe and Manning (2008) and give a more intuitive way of thinking about the relationships between words in a sentence. For instance, figure 8 indicates a dependency parse of the sentence, *The fox slept peacefully.*, where we see that *fox* is the subject of *slept*.

Tai et al. (2015) show us how this structure can be used to design LSTM networks where the direction of dependency relationships rather than linear order is used to determine how hidden states are propagated along the chain. However because doing this results in additional model complexity, it was desireable to see if including this grammatical information in simpler models would yield improved results.

To test this, all the sentence pairs in the SNLI corpus were parsed with Parsey-McParseface, a pretrained instance of Andor et al. (2016)'s transition-based neural dependency parser. Using these parses, we were able to represent each token as a triple consisting of: i) its pretrained embedding, ii) a unique (randomly initialized) d-dimensional embedding corresponding to the type of its dependency relationship and iii) the pretrained embedding corresponding to the head of its dependency relationship.

These triples were then concatenated and fed into each of the three models described in this section, with only the randomly initialized dependency-type embeddings updated during training. The word embeddings were instead fed through a projection layer similar to the one described in the introduction to this section.

For comparison, we ran the same experiments with the tokens represented by i) the token's embedding and the dependency-type embedding, ii) the token's embedding and a POS-tag embedding, and iii) just the token's embedding.

Results and Discussion

The results of these experiments are provided in table 3. The reader will note that our accuracy on plain tokens differs from 86.3, the value reported by Parikh et al.

Table 3: Accuracy of decomposable attention Models trained on SNLI dataset with and without grammatical information.

	Accuracy
Tokens	85.99%
Tokens+POS	**86.23%**
Tokens+Dep+Head	85.46%

(2016). This is likely due to slight differences in model construction and training. For instance, due to constraints in access to computational resources, we evaluated each of the models with a batch-size of 1024, rather than one of 4, which was the size chosen by Parikh et al. (2016) but would have led to a much slower training time. However, because these experiments were chiefly carried out to gauge the impact of including grammatical information, we believe it's more important to measure the change in model accuracy.

As evidenced by the table, including grammatical information does not seem to improve model accuracy by a considerable amount. Adding information about a word's dependencies decreases accuracy, while adding part-of-speech only increases accuracy by 0.28%.

There may be a number of reasons for this result. It's possible that because many of the tokens shared the same head, including an embedding corresponding to the head token for each of them would create a lot of redundant information that would act as noise to the network during training. Tai et al. (2015) get around this problem by composing a single vector representation for each head, that is a function of all its dependents. It is also possible that the dependency parser is not able to accurately parse all of the sentences, leading to the propagation of errors into the NLI model.

The increment in accuracy as a result of including part-of-speech information may be because the training set is large enough to latently encode sufficient information about how the syntactic properties of words affect the semantic relationships of the sentences they are contained in.

4.3 Agreement and Disagreement on Reddit

To evaluate our models' ability to infer agreement and disagreement in online discourse, we used the dataset described in section 3 to create a training set and test set consisting of 0.8 and 0.2 of the 115,680 labeled comment pairs.

The comments were then tokenized and parsed using a transition-based dependency parser, and all comments were truncated to a maximum of 100 tokens. Each

Table 4: Accuracy of the various models trained on Reddit dataset with and without POS information

	CBOW	**LSTM**	**Decomposable Attention**
Tokens	0.6794	**0.7737**	0.7180
Tokens+POS	0.6858	**0.7747**	0.7138

of the tokens was then converted to its embedded representation using the methodology described in section 4.1. We also evaluated the effect of including grammatical information in the form of POS-tags. Information from dependency relationships was excluded because of their poor performance in the experiments described in section 4.2.

The results of these experiments are provided in table 4. The accuracy reported for each combination of token and sentence encodings was obtained after searching for the best combination of hyperparameters (learning rate, size of hidden state and dropout rates).

It is somewhat remarkable that the LSTM does a better job at generating sentence encodings when compared to the decomposable attention model, which has in the past (Parikh et al., 2016) outperformed LSTM encoders on the SNLI corpus. A possible explanation for this discrepancy is the fact that the SNLI corpus was constructed from image captions. Because image captions describe a well-defined scene, they tend to be shorter and more concise, and will tend to use more specific vocabulary. On the other hand, social discourse (more so on political topics) is more likely to carry lexical ambiguity, and will require the reader to factor in sentence structure to infer the meanings of individual words. It is therefore insufficient to simply align tokens in the premise and hypothesis without factoring in the context in which they occur.

Given that the decomposable attention model produces alignments α and β between the premise and hypothesis, it also made sense to inspect the alignment weights to get insight into how the model carries out inference. Sample values of α are plotted out in figures 9 and 10. Each of these plots reveals the unfortunate fact that despite our best efforts to remove obvious indicators of negation, a good number of them made their way into the dataset. This number was significant enough to cause the model to place a large emphasis on the presence of these indicators (sometimes to its success and others to its detriment).

Aside from overemphasis of obvious negation, it was also noted that in a number of instances the model attended on the final punctuation mark in the parent comment, and the first token in the child comment. Examples of this are given in

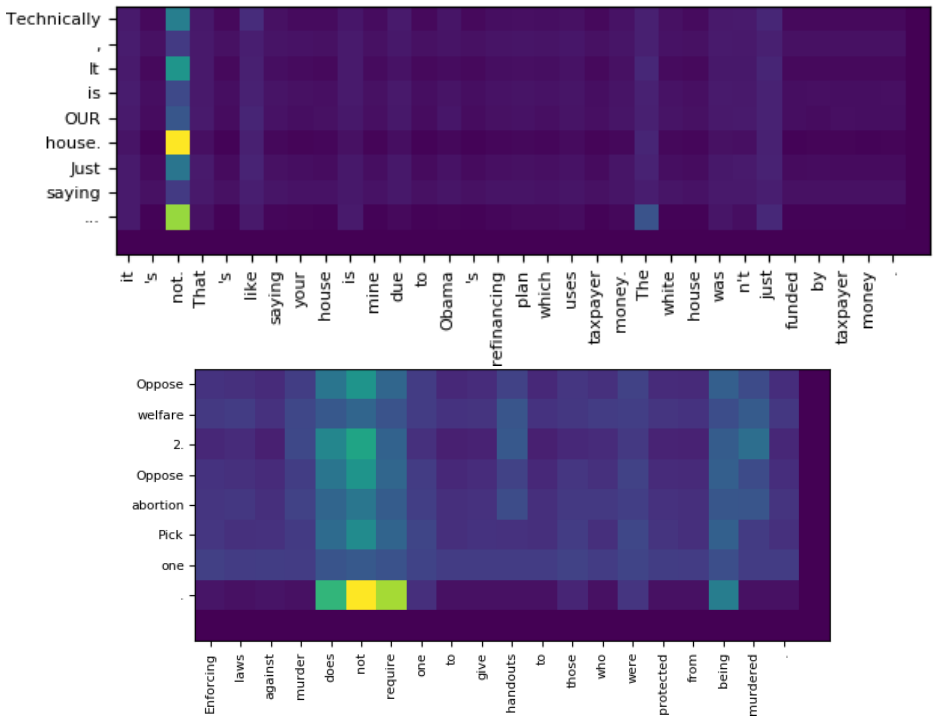


Figure 9: Plots of α alignments for comment pairs where the decomposable attention model's inference was correct.

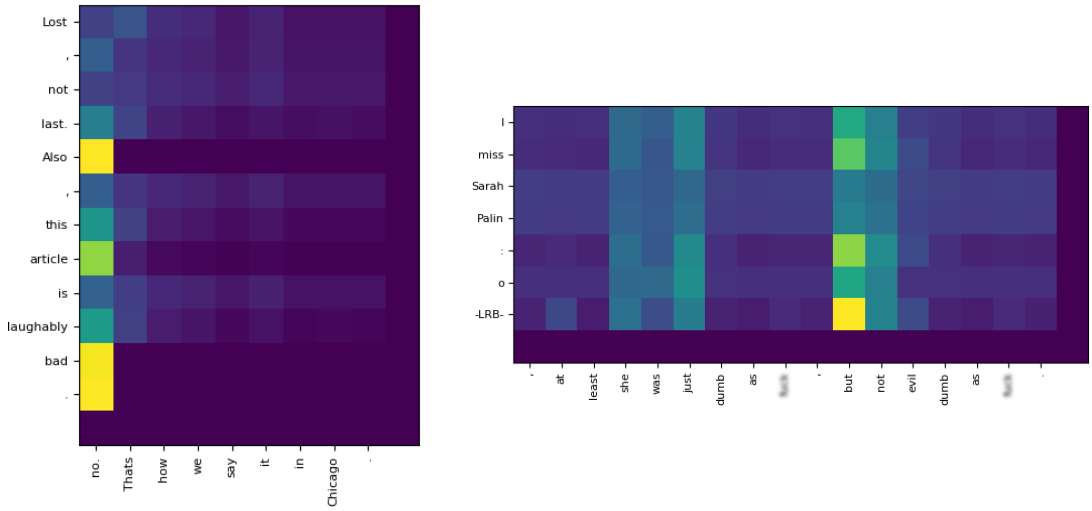
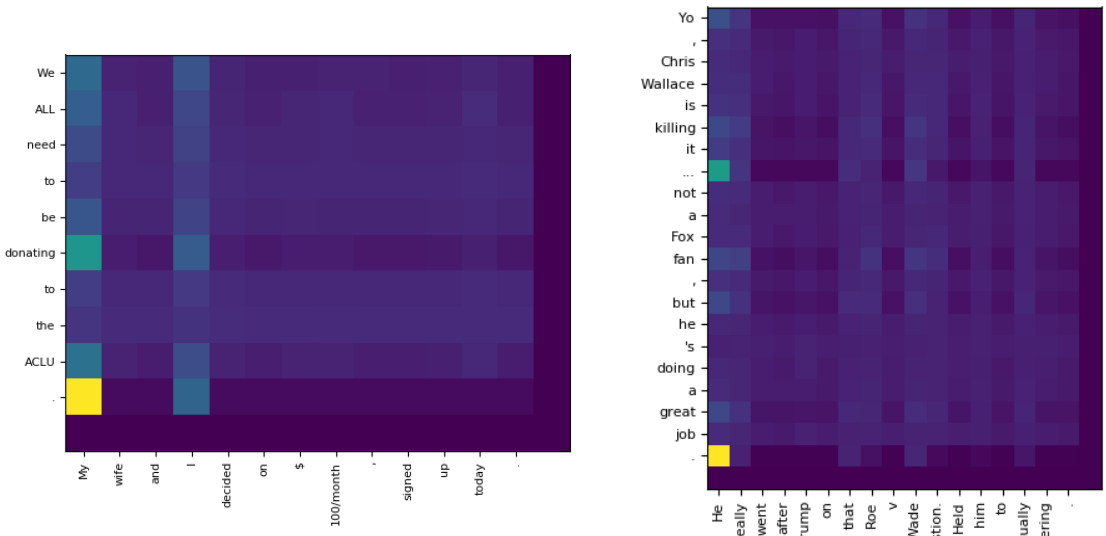


Figure 10: Plots of α alignments for comment pairs where the decomposable attention model's inference was incorrect.

figure 11. This may be an indicator that the first token in the child token is a strong indicator of the nature of the agreeing or disagreeing prefix that was stripped from the original comment. It also demonstrates that the model is more reliant on the lexical properties of the child comment than the alignments of related sub-phrases from the parent and child comments. This is includes, for example, ('donating', 'signed up today') in figure 11a and ('doing a great job', 'Held him to actually answering') in figure 11b.

It is interesting that the decomposable attention model that was so apt at aligning related sentence substructure in the SNLI corpus would fail at doing so for the Reddit comments. This may be explained by differences in the way disagreement and agreement manifest themselves in the two corpora. The SNLI corpus was constructed by asking annotators to provide sentences that were entailing, contradictory, or neutral to a supplied premise. And though corpus' authors tried to enforce rules that promoted sentence quality, the provided hypotheses often resembled paraphrased versions of their corresponding premises. On the other hand, comments in Reddit expressed agreement and disagreement in a style unique to the authors. In addition to this, comments on Reddit tend to use a far larger vocabulary, resulting in a larger number of words that fall out of the GloVe embeddings' vocabulary.

Seeing how useful the alignments of the decomposable model were in finding the flaws in our model architecture and data collection, we saw it fit to find a way of interpreting the results of the model that used an LSTM for sentence encoding.



(a) (b)

Figure 11: Plots of α alignments for comment pairs where the decomposable attention model placed undue emphasis on punctuation and the beginning of the child comment's sentences.

We therefore stacked the decomposable attention model's attention mechanism on top of an LSTM, resulting in a model that learns to align the hidden states of the LSTM cells corresponding to each of the sentences' tokens, rather than their raw embeddings.

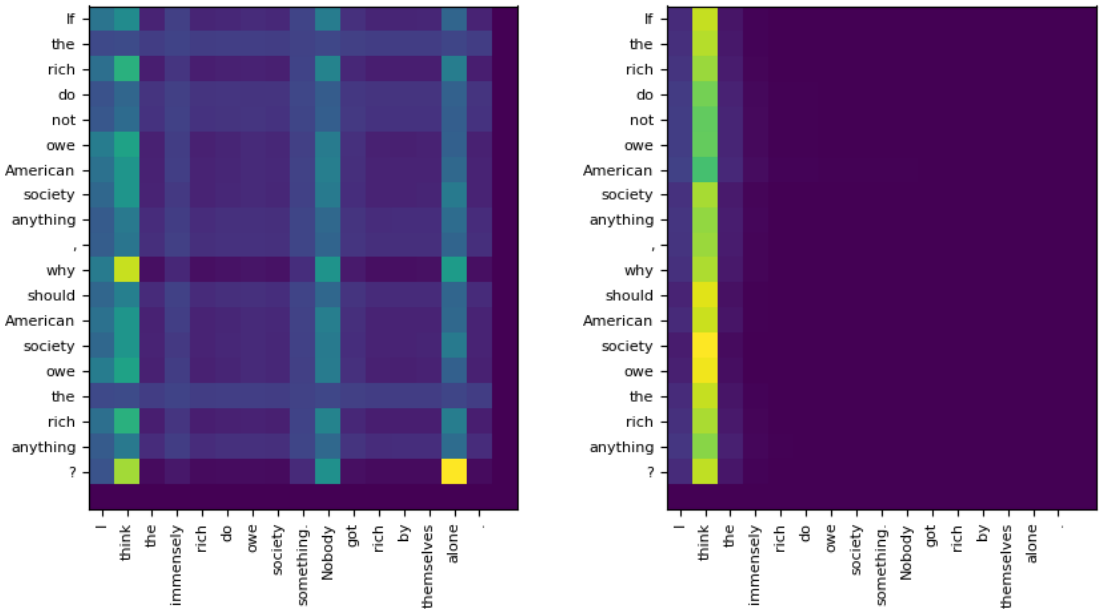
Because a unidirectional LSTM propagates information in one direction, the alignments resulting from this architecture appeared as smears starting from the token of interest and spreading to cells whose hidden states are a function of those of their preceding neighbors. This made the alignment plots harder to interpret. We therefore evaluated the use of a Bidirectional LSTM (BiLSTM) to generate hidden states for each token. As its name implies, a BiLSTM propagates hidden states, in both directions ensuring that all of a token's neighbors are taken into consideration when encoding its fixed-length representation.

Through hyperparameter search, we were able to train a model that achieved accuracy comparable to the plain LSTM models (**77.48%**). It is possible that gains to accuracy may have been made by making further refinements, however, this was sufficient to observe how an LSTM-based model's inferences differed from that of a plain decomposable attention model.

We found that while a lot of the weaknesses of decomposable attention remained evident, the LSTM-based model was able to identify more subtle indicators of disagreement. For example, by placing emphasis on phrases such as 'I think' and 'If you think' it was able to correctly infer relationships of disagreement that the plain model got wrong. Figure 12 illustrates an example of this.

These results show that models designed to solve NLI tasks can be used to infer the agreement and disagreement between comments gathered from Reddit. The model performs significantly better than random guessing, and there is evidence that using complex models that factor in word order and inter-sentence alignment can provide a significant performance boost to a simple CBOW baseline.

However, by using sentence alignments to gain a better understanding of the models' inference process, we have also observed that greater care must be taken in automatically constructing a corpus of agreeing and disagreeing pairs. Allowing the bias used to automatically select comment-pairs to seep into the final dataset causes models to overemphasize features that prevent generalization. Possible remedies to these problems, and extensions to this work are discussed in the conclusion.



(a) (b)

Figure 12: Plots of α alignments of the same comment pair generated using a) a decomposable attention model and b) a decomposable attention model stacked on a BiLSTM.

5 Conclusion

5.1 Summary of Contributions

In this work, I have explored methods of identifying opposing views in online discourse. I have a presented a means for the unsupervised extraction of a large corpus of agreeing and disagreeing comment pairs, whilst ensuring that a majority of the examples are free of more obvious indicators of agreement and disagreement. I have explored the use of grammatical information, and a variety of sentence encodings in building classifiers capable of inferring the semantic relationships between these comment pairs. Last, I have explored the use attention mechanisms to provide insight into the decision making process of NLI models.

To my knowledge, this is the first work that attempts to analyse comments made in reaction to news events on Reddit. While a lot has been done in identifying stance on Twitter (Mohammad et al., 2016), it presents a different challenge because its users have to express their views in 140 characters, and will try and include topical context (in the form of hashtags) to improve the visibility of their tweets. On the other hand, Reddit allows up to 40,000 characters and because comments are made in threads dedicated to specific topics, writers can safely assume that they already share a lot of context with the reader.

To achieve this task, I have provided methodology for going about the unsupervised labelling of agreeing and disagreeing pairs. I hope that this work can act as a starting point for future work that requires the extraction of large corpora of agreeing and disagreeing document pairs from online discourse.

In addition, this work takes an approach that stratifies models according to the way that they generate sentence/document encodings. Coincidentally, this paradigm was the theme of the RepEval 2017 shared task (Nangia et al., 2017) in which participants were challenged to come up with robust ways of generating fixed-length document encodings suited to solving a multi-genre NLI task. The tasks described in this document may therefore serve as an additional proving ground for future work in improving document encoding.

I have also demonstrated that grammatical information extracted from syntactic parses does not provide significant gains to model accuracy when used as features. As discussed in section 4.2, this information is made most useful by factoring the syntactic structure of the sentences when constructing their fixed-length encodings.

Last, I believe that I have made a contribution in showing how attention mechanisms can be used to used to improve the interpretability of NLI models. Lipton (2016) states that the demand for interpretability most commonly arises when there

is a mismatch between the formal objectives of supervised learning and the real world costs in a deployment settings. Sure enough, in this work, setence alignments drawn from the attention mechanisms revealed a number of instances where models were able to make correct inferences by focusing on irrelevant features. Probing the models' decision making also revealed that their misbehavior was often times as a result of flaws in the dataset rather than the models' architecture. The alignments also revealed how the different methods of constructing sentence encodings impact a model's ability to factor in the sentences' constituent words.

5.2 Limitations and Future Work

The results presented in section 4.3 indicate that the quality of the data extracted from Reddit had a negative impact on their significance. Some of the issues encountered trying to use this data include:

1. The methods used to strip obvious indicators of agreement and disagreement were not as effective as anticipated. A lot more work can be done towards ensuring that a smaller fraction of the training set includes these markers.

2. A number of the comment pairs lacked sufficient contextual information to enable the models or even a human annotator to decide their semantic relationship. Future work should take into account the news article, topic of discussion and current events before making an inference about agreement/disagreement. This will require efficient ways of storing and referencing information from a larger pool of documents during inference.

3. Outside of using the human ranking provided by Reddit, it is hard to automatically gauge the quality of discourse. As a result, our dataset contains comments that were either sarcastic or made in jest but also happened to be well-received by their readers. Future work will therefore need to incorporate methods capable of detecting and excluding comments that exhibit sarcasm and humour.

In addition to improving the data collection process, it would also be useful to evaluate the use of models that factor in the syntactic structure of sentences when constructing sentence encodings, such as the methods proposed by Tai et al. (2015) and Chen et al. (2016).

Despite having made progress towards accurate pairwise comparison of comments made in online discourse, we have not explored ways of rapidly retrieving documents that are in agreement/disagreement with a query document. In the future, we might

be able to achieve this by instead focusing on generating fixed-length sentence encodings such that the level of agreement and disagreement is correlated with some distance measure between the vector representations. This would allow us to retrieve agreeing/disagreeing comments either through clustering or locality sensitive hashing (Andoni et al., 2015).

It is the author's hope that this work will contribute to the eventual construction of a system capable of retrieving opposing views on the Internet, and grant people access to a broader spectrum of ideas and opinions.

References

Rob Abbott, Marilyn Walker, Pranav Anand, Jean E Fox Tree, Robeson Bowmani, and Joseph King. 2011. How can you say such things?!?: Recognizing disagreement in informal political argument. In *Proceedings of the Workshop on Languages in Social Media*, pages 2–11. Association for Computational Linguistics.

Y. Adi, E. Kermany, Y. Belinkov, O. Lavi, and Y. Goldberg. 2016. Fine-grained Analysis of Sentence Embeddings Using Auxiliary Prediction Tasks. *ArXiv e-prints arXiv:1608.04207*.

A. Andoni, P. Indyk, T. Laarhoven, I. Razenshteyn, and L. Schmidt. 2015. Practical and Optimal LSH for Angular Distance. *ArXiv e-prints arXiv:1509.02897*.

Daniel Andor, Chris Alberti, David Weiss, Aliaksei Severyn, Alessandro Presta, Kuzman Ganchev, Slav Petrov, and Michael Collins. 2016. Globally normalized transition-based neural networks. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2442–2452, Berlin, Germany. Association for Computational Linguistics.

Gabor Angeli, Neha Nayak, and Christopher D. Manning. 2016. Combining natural logic and shallow reasoning for question answering. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 442–452, Berlin, Germany. Association for Computational Linguistics.

Isabelle Augenstein, Tim Rocktäschel, Andreas Vlachos, and Kalina Bontcheva. 2016. Stance detection with bidirectional conditional encoding. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 876–885, Austin, Texas. Association for Computational Linguistics.

D. Bahdanau, K. Cho, and Y. Bengio. 2014. Neural machine translation by jointly learning to align and translate. *ArXiv e-prints arXiv:1409.0473*.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A neural probabilistic language model. *Journal of Machine Learning Research*, 3(Feb):1137–1155.

Jonathan Berant, Ido Dagan, and Jacob Goldberger. 2011. Global learning of typed entailment rules. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 610–619, Stroudsburg, PA, USA. Association for Computational Linguistics.

Johan Bos and Katja Markert. 2005. Recognising textual entailment with logical inference. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, pages 628–635, Stroudsburg, PA, USA. Association for Computational Linguistics.

S. R. Bowman. 2013. Can recursive neural tensor networks learn logical reasoning? *ArXiv e-prints arXiv:1312.6192*.

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

Samuel R. Bowman, Jon Gauthier, Abhinav Rastogi, Raghav Gupta, Christopher D. Manning, and Christopher Potts. 2016. A fast unified model for parsing and sentence understanding. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1466–1477, Berlin, Germany. Association for Computational Linguistics.

Samuel R. Bowman, Christopher Potts, and Christopher D. Manning. 2015b. Recursive neural networks can learn logical semantics. In *Proceedings of the 3rd Workshop on Continuous Vector Space Models and their Compositionality*, pages 12–21, Beijing, China. Association for Computational Linguistics.

Q. Chen, X. Zhu, Z. Ling, S. Wei, H. Jiang, and D. Inkpen. 2016. Enhanced LSTM for natural language inference. *ArXiv e-prints arXiv:1609.06038*.

M. Del Vicario, A. Bessi, F. Zollo, F. Petroni, A. Scala, G. Caldarelli, H. E. Stanley, and W. Quattrociocchi. 2015. Echo chambers in the age of misinformation. *ArXiv e-prints arXiv:1509.00189*.

Shiri Dori-Hacohen and James Allan. 2015. *Automated Controversy Detection on the Web*. Springer International Publishing, Cham.

Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *In Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS'10). Society for Artificial Intelligence and Statistics*.

Xavier Glorot, Antoine Bordes, and Yoshua Bengio. 2011. Deep sparse rectifier neural networks. In *Proceedings of the Fourteenth International Conference on Artificial*

Intelligence and Statistics, volume 15 of *Proceedings of Machine Learning Research*, pages 315–323, Fort Lauderdale, FL, USA. PMLR.

Christoph Goller and Andreas Küchler. 1996. Learning task-dependent distributed representations by backpropagation through structure. In *Proceedings of the ICNN-96*, pages 347–352. IEEE.

Andrew Hickl, Kirk Roberts, Bryan Rink, and Ying Shi. 2006. Recognizing textual entailment with lcc's groundhog system. In *Proceedings of the Second PASCAL Challenges Workshop*, volume 18.

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

Myungha Jang and James Allan. 2016. Improving automated controversy detection on the web. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '16, pages 865–868, New York, NY, USA. ACM.

Luyang Li, Bing Qin, and Ting Liu. 2017. Contradiction detection with contradiction-specific word embedding. *Algorithms*, 10(2).

Z. C. Lipton. 2016. The Mythos of Model Interpretability. *ArXiv e-prints arXiv:1606.03490*.

Y. Liu, C. Sun, L. Lin, and X. Wang. 2016. Learning natural language inference using bidirectional LSTM model and inner-attention. *ArXiv e-prints arXiv:1605.09090*.

Bill MacCartney. 2009. *Natural language inference*. Ph.D. thesis, Stanford University, Department of Computer Science.

Bill MacCartney, Trond Grenager, Marie-Catherine de Marneffe, Daniel Cer, and Christopher D. Manning. 2006. Learning to recognize features of valid textual entailments. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 41–48, New York City, USA. Association for Computational Linguistics.

Bill MacCartney and Christopher D. Manning. 2007. Natural logic for textual inference. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, RTE '07, pages 193–200, Stroudsburg, PA, USA. Association for Computational Linguistics.

Bill MacCartney and Christopher D. Manning. 2008. Modeling semantic containment and exclusion in natural language inference. In *Proceedings of the 22nd International Conference on Computational Linguistics - Volume 1*, COLING '08, pages 521–528, Stroudsburg, PA, USA. Association for Computational Linguistics.

Bill MacCartney and Christopher D. Manning. 2009. An extended model of natural logic. In *Proceedings of the Eighth International Conference on Computational Semantics*, IWCS-8 '09, pages 140–156, Stroudsburg, PA, USA. Association for Computational Linguistics.

Marie-Catherine de Marneffe and Christopher D. Manning. 2008. The Stanford typed dependencies representation. In *COLING Workshop on Cross-framework and Cross-domain Parser Evaluation*.

Marie-Catherine de Marneffe, Anna N. Rafferty, and Christopher D. Manning. 2008. Finding contradictions in text. In *Proceedings of Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, HLT '08, pages 1039–1047, Columbus, Ohio. Association for Computational Linguistics.

T. Mikolov, K. Chen, G. Corrado, and J. Dean. 2013. Efficient Estimation of Word Representations in Vector Space. *ArXiv e-prints arXiv:1301.3781*.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. Semeval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 31–41, San Diego, California. Association for Computational Linguistics.

Nikita Nangia, Adina Williams, Angeliki Lazaridou, and Samuel Bowman. 2017. The repeval 2017 shared task: Multi-genre natural language inference with sentence representations. In *Proceedings of the 2nd Workshop on Evaluating Vector Space Representations for NLP*, pages 1–10. Association for Computational Linguistics.

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*, pages 2249–2255, Austin, Texas. Association for Computational Linguistics.

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.

Alan Ritter, Doug Downey, Stephen Soderland, and Oren Etzioni. 2008. It's a contradiction—no, it's not: A case study using functional relations. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 11–20, Stroudsburg, PA, USA. Association for Computational Linguistics.

Maria Skeppstedt, Magnus Sahlgren, Carita Paradis, and Andreas Kerren. 2016. Unshared task : (dis)agreement in online debates. In *The 54th Annual Meeting of the Association for Computational Linguistics*, pages 154–159. Association for Computational Linguistics.

Christian Stab and Iryna Gurevych. 2016. Recognizing the absence of opposing arguments in persuasive essays. In *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*, pages 113–118. Association for Computational Linguistics.

Ilya Sutskever. 2013. *Training recurrent neural networks*. Ph.D. thesis, University of Toronto, Department of Computer Science.

Kai Sheng Tai, Richard Socher, and Christopher D. Manning. 2015. Improved semantic representations from tree-structured long short-term memory networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1556–1566, Beijing, China. Association for Computational Linguistics.

Laurent Thomet. 2018. China drowns out critics of lifetime Xi Jinping presidency, as 'disagree' among words censored online. *Hong Kong Free Press*. Accessed: 2018-02-27.

Yotaro Watanabe, Junta Mizuno, Eric Nichols, Naoaki Okazaki, and Kentaro Inui. 2012. A latent discriminative model for compositional entailment relation recognition using natural logic. In *Proceedings of COLING 2012*, pages 2805–2820. The COLING 2012 Organizing Committee.

J. Wieting, M. Bansal, K. Gimpel, and K. Livescu. 2015. Towards universal paraphrastic sentence embeddings. *ArXiv e-prints arXiv:1511.08198*.

A. Williams, N. Nangia, and S. R. Bowman. 2017. A broad-coverage challenge corpus for sentence understanding through inference. *ArXiv e-prints arXiv:1704.05426*.

Guido Zarrella and Amy Marsh. 2016. Mitre at semeval-2016 task 6: Transfer learning for stance detection. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 458–463, San Diego, California. Association for Computational Linguistics.