Modeling the Fake News Challenge as a Cross-Level Stance Detection Task

Costanza Conforti
Language Technology Lab
University of Cambridge
cc918@cam.ac.uk

Mohammad Taher Pilehvar
Language Technology Lab
University of Cambridge
mp792@cam.ac.uk

Nigel Collier
Language Technology Lab
University of Cambridge
nhc30@cam.ac.uk

Abstract

The 2017 Fake News Challenge Stage 1, a shared task for stance detection of news articles and claims pairs, has received a lot of attention in recent years [ea18]. The provided dataset is highly unbalanced, with a skewed distribution towards *unrelated* samples - that is, randomly generated pairs of news and claims belonging to different topics. This imbalance favored systems which performed particularly well in classifying those noisy samples, something which does not require a deep semantic understanding.

In this paper, we propose a simple architecture based on conditional encoding, carefully designed to model the internal structure of a news article and its relations with a claim. We demonstrate that our model, which only leverages information from word embeddings, can outperform a system based on a large number of hand-engineered features, which replicates one of the winning systems at the Fake News Challenge [HASC17], in the stance detection of the *related* samples.

1 Introduction

Stance classification has been identified as a key subtask in rumor resolution [ZAB+18]. Recently, a similar approach has been proposed to address fake news detection: as a first step towards a comprehensive model for news veracity classification, a corpus of news articles, stance-annotated with respect to claims, has been released for the Fake News Challenge (FNC-1)[1].

Characteristics of the corpus - The FNC-1 corpus is based on the EMERGENT dataset [FV16], a collection of 300 claims and 2,595 articles discussing the claims. Each article is labeled with the stance it expresses toward the claim and summarized into a headline by accredited journalists, in the framework of a project for rumor debunking [Sil15].

For creating the FNC-1 corpus, the headlines and the articles were paired and labeled with the corresponding stance, distinguishing between *agreeing* (AGR), *disagreeing* (DSG) and *discussing* (DSC). Additional 266 labeled samples were added to avoid cheating [ea18]. Moreover, a number of *unrelated* (UNR) samples were obtained by randomly matching headlines with articles discussing a different claim. As shown in Table 1, the final class distribution was highly skewed in favor of the UNR class, which amounted to almost three quarters of the samples (Table 1).

Characteristics of the FNC-1 winning models - As a consequence of being randomly generated, classification of UNR samples is relatively easy. Moreover, given that the UNR samples constitute the large majority of the corpus, most competing systems were designed in order to perform well on this easy-to-discriminate class. In fact, the three FNC-1 winning teams proposed relatively standard architectures (mainly based on multilayer perceptrons, MLPs) leveraging a large number of classic, hand-engineered NLP features. While those systems performed very well on the UNR class - reaching a F_1 score higher than .99 - they were not as effective in the AGR, DSG and DSC classification [ea18].

FNC-1 as a Cross-Level Stance Detection task - As shown in Table 3, one specific characteristic of the FNC-1 corpus consists in the clear asymmetry in length between the headlines and the articles. While headlines consist of one sentence, the structure

[1] http://www.fakenewschallenge.org/

Table 1: Example of an agreeing headline from the FNC-1 training set, with its related document divided into paragraphs (doc 1880). Each paragraph may express a different stance with respect to the claim, as indicated in the first column.

	Headline. No, a spider (probably) didn't crawl through a man's body for several days
	Article.
AGR	"Fear not arachnophobes, the story of Bunbury's "spiderman" might not be all it seemed.
DSC	[...] scientists have cast doubt over claims that a spider burrowed into a man's body [...] The story went global [...]
DSG	Earlier this month, Dylan Thomas [...] sought medical help [...] he had a spider crawl underneath his skin.
DSG	Mr Thomas said a [...] dermatologist later used tweezers to remove what was believed to be a "tropical spider".
(noise)	[image via Shutterstock]
AGR	But it seems we may have all been caught in a web... of misinformation.
DSC/**AGR**	Arachnologist Dr Framenau said [...]it was "almost impossible" [...] to have been a spider [...]
(noise)	Dr Harvey said: "We hear about people going on holidays and having spiders lay eggs under the skin". [...]
(noise)	Something which is true, [...] is that certain arachnids do live on humans. We all have mites living on our faces [...]
(noise)	Dylan Thomas has been contacted for comment."

of an article is better described as a sequence of paragraphs, where each paragraph plays a different role in telling a story. Single paragraphs usually expresses different views of a topic. Following the terminology introduced by [JPN14], we propose to call this variant of the classic Stance Detection task *Cross-Level Stance Detection*.

As shown in Table 1, an article consists in passages presenting a news, reporting about interviews, giving general background information and discussing similar events happened in the past. In contemporary news-writing prose, the most salient information is usually condensed in the very first paragraphs, following the *Inverted Pyramid* style. This allows the reader for rapid decision making [Sca00].

For these reasons, we believe that detecting the stance of an article with respect to a headline requires a deep understanding not only of the position taken in each paragraph with respect to the headline, but also of the the complex interactions within the article's paragraphs, as illustrated by the example in Table 3. On the contrary, compressing both the headline's and article's content into fixed-size vectors, as in the feature-based systems described in the previous paragraph, fails in detecting those fine-grained relationships and results in sub-optimal performance on the stance detection of AGR, DSG and DSC samples.

To test this assumption, we propose a simple architecture based on conditional encoding, which is designed in order to model the complex interactions between headlines and articles described above, and we compare it with one of the feature-based systems which won the FNC-1 [HASC17].

	instances	AGR	DSG	DSC	UNR
FNC-1	75,385	7.4%	2.0%	17.7%	72.8%
FNC-1-rel	20,491	27.2%	7.5%	65.2%	-

Table 2: Label distribution for the FNC-1 dataset, with and without the UNR samples.

In order to be able to assess the ability of the systems to model the complex headline-article interplay described above, we filter out the noisy UNR samples and consider only the *related* samples (AGR, DSG and DIS). Those samples were manually collected and labeled by professional journalists and require deep semantic understanding in order to be classified, constituting a difficult task even for humans. This is evident when looking at the inter-annotator agreement of human raters, which drops from Fleiss' $\kappa = .686$ to .218 when including or excluding the UNR samples, as reported in [ea18]. The final label distribution is reported in Table 1.

2 Models

2.1 Feature-based approach

We implemented the model proposed by the team *Athene*, which was ranked second at FNC-1[2]. The model consists of a 7-layer MLP with ReLU activation. On the top of the architecture, a softmax layer is used for prediction (Figure 1).

Input is given in the form of a large matrix of hand-engineered features. The considered set includes the concatenation of feature vectors which separately consider the headline and the article - like the presence of refuting or polarity words (taken from a hand-selected list of words as 'hoax' or 'debunk', and tf-idf weighted Bag of Words vectors - and features which combine the headline and the article (*joint features* in Figure 1) - like word/ngram overlap between the headline and the article, and cosine similarity of the embeddings of

[2]We used *Athene* as the baseline as the FNC-1 winning model was an ensemble [BSP17].

Table 3: Asymmetry in length in the FNC-1 corpus.

	headline	article	paragraph
avg #tokens	12.40	417.69	30.88
avg #par/article	-	11.97	-

nouns and verbs between the headline and the article. Moreover, topic-based features based on non-negative matrix factorization, latent Dirichlet allocation and latent semantic indexing were used. For a detailed description of the features, refer to [HASC17].

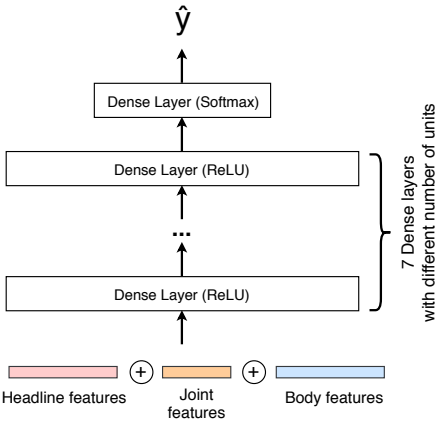

Figure 1: The feature-based model proposed by [HASC17].

2.2 Conditional approach

In order to model the headline-article interactions described in Section 1, we adapt the bidirectional conditional encoding architecture first proposed by [ARVB16] for stance detection of tweets.

First, the article is split into n paragraphs. Both the headline and the paragraphs are converted into their embedding representations. The headline is then processed by a Bi-LSTM$_h$ (Eq 1). Each paragraph is then encoded by a further Bi-LSTM$_{S1}$ (Eq 2), whose initial cell states are initialized with the last states of the respectively forward and backward LSTMs which compose Bi-LSTM$_h$ (see Figure 2 for a representation of the architecture's forward part). As pointed out in [ARVB16], this allows Bi-LSTM$_{S1}$ to read the paragraph in a headline-specific manner.

$$H_h = \text{Bi-LSTM}_h(E_h) \tag{1}$$

$$H_{s_i} = \text{Bi-LSTM}_{S1}(E_{s_i}) \qquad \forall i \in \{1, ..., n\} \tag{2}$$

where $E_h \in R^{e \times H}$ and $E_{si} \in R^{e \times S_i}$ are respectively the embedding matrix of the headline and of the i_{th} paragraph, H and S_i are respectively the headline and the i_{th} paragraph length, e is the embedding size, l is the hidden size, $H_h \in R^{l \times H}$ and $H_{s_1} \in R^{l \times S_i}$.

Then, each paragraph representation, conditionally encoded on the headline, is processed by another Bi-LSTM$_{S2}$, conditioned on the previous paragraph. We start the paragraph-conditioned reading of the article from the bottom, as we assume the most salient

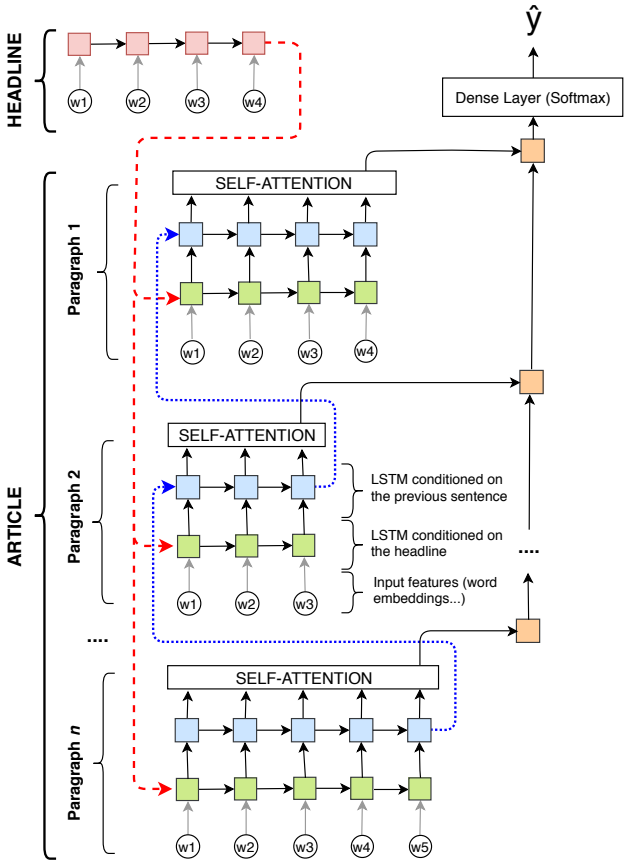

Figure 2: Model based on conditional encoding (best seen in color). Networks represented with the same color share the weights. Dotted arrows represent conditional encoding. Due to lack of space, we represent only the forward part of the encoder. However, headline and paragraph encoders (the red, green and blue networks in the figure) are Bi-LSTM.

information to be concentrated in the beginning (see Section 1).

$$H_{s_i} = \text{Bi-LSTM}_{S2}(H_{s_i}) \qquad \forall i \in \{1, ..., n\} \tag{3}$$

resulting in a matrix $H_{s_i} \in R^{l \times S_i}$. We employ a similar self-attention mechanism as in [ea16] in order to soft-select the most relevant elements of the sentence. Given the sequence of vectors $\{h_1, ..., h_S\}$ which compose H_{S_i}, the final representation of the i_{th} paragraph s_i is obtained as follows:

$$u_{it} = \tanh(W_s h_{it} + b_s) \tag{4}$$

$$\alpha_{it} = \exp \frac{u_{it}^\top u_s}{\sum_t u_{it}^\top u_s} \tag{5}$$

$$s_i = \sum_t \alpha_t h_{it} \tag{6}$$

Table 4: Macro-averaged precision, recall and F_1 scores on the development and test set

	dev set			test set		
	P_m	R_m	F_{1m}	P_m	R_m	F_{1m}
Feat-based model	.359	.350	.350	.388	.361	.367
Cond model	.685	.716	.699	.505	.503	.486

where the hidden representation of the word at position t, u_{it}, is obtained though a one-layer MLP (Eq 4). The normalized attention matrix α_t is then obtained though a softmax operation (Eq 5). Finally, s_i is computed by a weighted sum of all hidden states h_t with the weight matrix α_t (Eq 6). The sentence representations $\{s_1, ..., s_n\}$ are aggregated using a backward LSTM, as in Figure 2. The final prediction $\hat{y}$ is obtained with a softmax operation over the tagset.

3 Experiments

3.1 (Hyper-)Parameters

For the feature-based model, we downloaded the feature matrices used by [HASC17] for their FNC-1 best submission[3] and selected the columns corresponding to the related samples. For the conditional model, we initialized the embedding matrix with word2vec embeddings[4]. Only words which occurred more than 7 times were included in the embedding matrix. Words not included in word2vec were zero-initialized. In order to avoid overfitting, we did not fine-tune the embeddings during training. The main structures of the models were implemented in keras, using Tensorflow for implementing customized layers. Refer to Appendix **??** for the complete list of hyperparameters used to train both architectures.

3.2 Evaluation Metrics

In the FNC-1 context, a so-called FNC score was proposed for evaluation: this hierarchical evaluation metric gives 0.25 points for a correct REL/UNR classifications, which is incremented of 0.75 points in case of a correct AGR/DSA/DSC classification[5]. This was motivated by the high imbalance in favor of UNR class.

However, as in our experiments we are only considering REL samples, the FNC score does not constitute a useful evaluation metric. Following [ea18], we use macro-averaged precision, recall and F_1 score, which is less affected by the high class imbalance (Table 1).

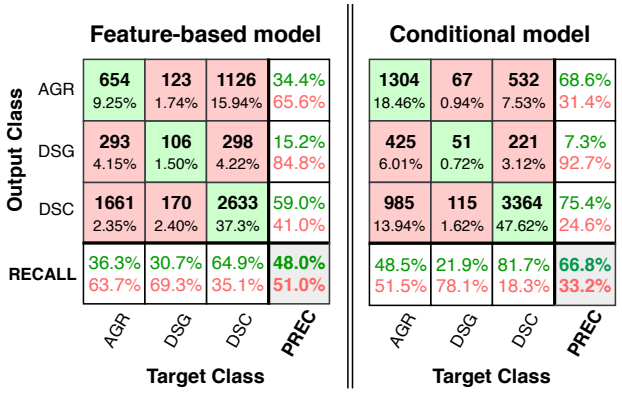

Figure 3: Confusion Matrices of the predictions of both the feature-based and the conditional model on the test set.

3.3 Results and Discussion

Results of experiments are reported in Table 4. The proposed conditional model clearly outperforms the feature-based baseline for all considered metrics, despite having a considerably minor number of trainable parameters. Interestingly, the feature-based model seems to offer a better generalization over the test set, while the gap between development and test set performance in the conditional model seems to indicate overfitting.

Detailed performance on single classes is shown in Figure 3. Thanks to the presence of features specifically designed to target the presence of refuting words, the baseline model is able to reach a Precision of 15.2% in classifying the very infrequent DSG class (7.5% of occurrences). The conditional model, which did not receive any explicit signal of the presence of negation, suffers more from this data imbalance, and reaches a Precision of 7.3% on DSG samples. On the other hand, by flattening the entire article into a fixed-size vector, the feature-based system looses the nuances in the argumentative structure of the news story. As a consequence, this system struggles to distinguish between AGR and DSC samples and tends to favor the most frequent DSC class, which receives the highest Precision and Recall scores. On the contrary, the conditional model is able to spot the subtle differences between AGR and DSC samples, reaching high Precision and satisfactory Recall in both classes despite the large class imbalance - 27.7% AGR vs. 65.2% DSC samples.

[3] https://drive.google.com/open?id=0B0-muIdcdTp7UWVyU0duSDRUd3c

[4] https://code.google.com/archive/p/word2vec/

[5] https://github.com/FakeNewsChallenge/fnc-1-baseline/blob/master/utils/score.py

4 Conclusions

Given the results discussed in the previous Section, we believe the strategy of modeling the FNC-1 as an Asymmetric Stance Detection problem is promising. In future work, we will carry on a detailed qualitative analysis to test the extent to which our conditional model is able to model the narrative structures of articles and their interactions with the headlines. The generalizability of such architecture to other domains can be tested on other publicly available corpora, as the recently released ARC dataset by [ea18].

Acknowledgments

The first author (CC) would like to thank the Siemens Machine Intelligence Group (CT RDA BAM MIC-DE, Munich) and the NERC DREAM CDT (grant no. 1945246) for partially funding this work. The third author (NC) is grateful for support from the UK EPSRC (grant no. EP/MOO5089/1).

References

[ARVB16] Isabelle Augenstein, Tim Rocktäschel, Andreas Vlachos, and Kalina Bontcheva. Stance detection with bidirectional conditional encoding. In *Proceedings of EMNLP 2016*, pages 876–885, 2016.

[BSP17] Sean Baird, Doug Sibley, and Yuxi Pan. Talos targets disinformation with fake news challenge victory. https://blog.talosintelligence.com/2017/06/, 2017.

[ea16] Zichao Yang et al. Hierarchical attention networks for document classification. In *Proceedings of NAACL-HLT 2016*, pages 1480–1489, 2016.

[ea18] Andreas Hanselowski et al. A retrospective analysis of the fake news challenge stance-detection task. In *Proceedings of COLING 2018*, pages 1859–1874, 2018.

[FV16] William Ferreira and Andreas Vlachos. Emergent: a novel data-set for stance classification. In *Proceedings of NAACL-HLT 2016*, pages 1163–1168, 2016.

[HASC17] Andreas Hanselowski, PVS Avinesh, Benjamin Schiller, and Felix Caspelherr. Description of the system developed by team athene in the fnc-1. Technical report, 2017.

[JPN14] David Jurgens, Mohammad Taher Pilehvar, and Roberto Navigli. Semeval-2014 task 3: Cross-level semantic similarity. In *Proceedings of SemEval 2014*, pages 17–26, 2014.

[Sca00] Christopher Scanlan. *Reporting and writing: Basics for the 21st century*. Harcourt College Publishers, 2000.

[Sil15] Craig Silverman. *Lies, damn lies and viral content*. Columbia University, 2015.

[ZAB+18] Arkaitz Zubiaga, Ahmet Aker, Kalina Bontcheva, Maria Liakata, and Rob Procter. Detection and resolution of rumours in social media: A survey. *ACM Computing Surveys (CSUR)*, 51(2):32, 2018.