Emergent: a novel data-set for stance classification

William Ferreira
Department of Computer Science
University College London, UK

Andreas Vlachos
Department of Computer Science
University of Sheffield, UK

Abstract

We present Emergent, a novel data-set derived from a digital journalism project for rumour debunking. The data-set contains 300 rumoured claims and 2,595 associated news articles, collected and labelled by journalists with an estimation of their veracity (*true*, *false* or *unverified*). Each associated article is summarized into a headline and labelled to indicate whether its stance is *for*, *against*, or *observing* the claim, where *observing* indicates that the article merely repeats the claim. Thus, Emergent provides a real-world data source for a variety of natural language processing tasks in the context of fact-checking. Further to presenting the dataset, we address the task of determining the article headline stance with respect to the claim. For this purpose we use a logistic regression classifier and develop features that examine the headline and its agreement with the claim. The accuracy achieved was 73% which is 26% higher than the one achieved by the Excitement Open Platform (Magnini et al., 2014).

1 Introduction

The advent of *New Media*, such as Twitter, Facebook, etc., enables news stories and rumours to be published in real-time to a global audience, bypassing the usual verification procedures used by more traditional *Old Media* news outlets. However, the line between Old and New Media is becoming blurred as news aggregators lift stories from social media and re-publish them without fact-checking.

This issue could be helped by developing methods for automated fact-checking of news stories, part of the *reporter's black box* envisioned in Cohen et al. (2011) and one of the main objectives in computational journalism. While this task is related to a variety of natural language processing tasks such as textual entailment and machine comprehension, it poses additional challenges due to its open-domain, real-world nature. Previous work by Vlachos and Riedel (2014) proposed using data from fact-checking websites such as Politifact[1], but the labelling provided by the journalists is only the degree of truthfulness of the claims, without any machine-readable verdicts to supervise the various steps in deciding it. Thus, the task defined by the dataset proposed remains too challenging for the NLP methods currently available.

In this paper we propose to use data from the Emergent Project (Silverman, 2015), a rumour debunking project carried out in collaboration with the Tow Center for Digital Journalism at Columbia Journalism School[2]. Consisting of 300 claims and 2,595 associated news articles, the Emergent project contains a rich source of labelled data that can be used in a variety of NLP tasks, created by journalists as part of their normal workflow, thus real-world and at no annotation cost.

We leverage the Emergent dataset to investigate the task of classifying the stance of a news article headline with respect to its associated claim, i.e. for each article headline we assign a stance label which is one of *for*, *against*, or *observing*, indicating whether the article is supporting, refuting, or just reporting the claim, respectively. The large number

[1] http://www.politifact.com/
[2] http://towcenter.org/

Proceedings of NAACL-HLT 2016, pages 1163–1168,
San Diego, California, June 12-17, 2016.

of claims in the dataset allows us to assess the generalization of the method evaluation to new claims more reliably than in previous work that either used a small number of claims (e. g. seven in Lukasik et al., 2015) or did not separate training claims from testing claims (Qazvinian et al., 2011).

We develop a stance classification approach based on multiclass logistic regression, using features extracted from the article headline and the claim, achieving an accuracy of 73% on our test data-set, also demonstrating that features relying on syntax, word alignment and paraphrasing contribute to the performance. Since the task bears similarities with textual entailment, we compare it against the Excitement Open Platform (Magnini et al., 2014) which achieved a substantially lower accuracy of 47%.

2 The Emergent data

The claims in Emergent are collected by journalists from a variety of sources such as rumour sites, e.g. snopes.com, and Twitter accounts such as @Hoaxalizer. Their subjects include topics such as world and national U.S. news and technology stories. Once a claim is identified, the journalist searches for articles that mention the claim and decides on the *stance* of each such article:

- *for*: The article states that the claim is true, without any kind of hedging.
- *against*: The article states that the claim is false, without any kind of hedging.
- *observing*: The claim is reported in the article, but without assessment of its veracity.

The journalist also summarises the article into a headline. In parallel to the article-level stance detection, a claim-level veracity judgement is reached as more articles associated with the claim are examined. The veracity of each claim is initially *unverified*, later becoming either *true* or *false* when the journalist decides that adequate evidence from the associated articles has been compiled. Finally, the source and the number of times each associated article is shared are recorded. An example of a claim verified on Emergent appears in Figure 1.

There are a number of tasks for which the Emergent data can be useful for development and evaluation. The article-level stance labels can be used to develop a stance detection system between the claim and an associated article. The claim-level veracity labels would be straightforward to use for fact-checking. Finally, the article headlines can be used for focused summarization.

Claim: Robert Plant ripped up an $800 million contract offer to reunite Led Zeppelin

Source: mirror.co.uk (shares: 39,140)
Headline: Led Zeppelin's Robert Plant turns down £500MILLION to reform supergroup
Stance: *for*

Source: usnews.com (shares: 850)
Headline: No, Robert Plant Didn't Rip Up an $800 Million Contract
Stance: *against*

Source: forbes.com (shares: 3,360)
Headline: Robert Plant Reportedly Tears Up $800 Million Led Zeppelin Reunion Contract
Stance: *observing*

Veracity: *False*

Figure 1: Example verification taken from `http://www.emergent.info/led-zeppelin-contract`. The full text of the articles is omitted for brevity.

In this paper we focus on stance detection of an article with respect to the claim using the headline provided by the journalist. For this purpose we obtained a database dump from the developers of Emergent and extracted all claims and associated article headlines. We made no attempt to exclude a claim or article based on grammatical errors or complex syntactic structure. Our final dataset contains 300 claims, and 2,595 associated article headlines, with an average ratio of 8.65 (7.31) articles per claim; the minimum number of articles per claim is 1 and the maximum number is 50. The class distribution of article stances is 47.7% *for*, 15.2% *against* and 37.1% *observing*. This dataset was split into training and test set parts, containing 2,071 and 524 instances respectively, ensuring that each claim appeared in only one of the parts. Both the database dump and the extracted claim-article headline dataset are available from `https://github.com/willferreira/mscproject`.

3 Stance Classification

We treat stance classification as a 3-way classification task using a logistic regression classifier with

L_1 regularization (Pedregosa et al., 2011)[3] and we explore two types of features: those extracted solely from the article headline and those extracted by combining the headline and the claim. The former are aimed at capturing the cases in which the stance of headline can be determined without consulting the claim, which is often the case with *observing* cases, as they often use hedging. The latter are aimed at determining the entailment relation between them. All feature engineering was conducted using 10-fold cross-validation on the training data. Our implementation is available from `https://github.com/willferreira/mscproject`.

Headline features The features extracted from the headline are the commonly used bag of words representation (**BoW**) and whether it ends in a question mark (**Q**). In addition, we added two features representing the minimum distance from the root of the sentence to common refuting (e. g. deny) and hedging/reporting (e.g. claim, presumably) words (**RootDist**). As an example of the **RootDist** feature, consider the dependency parse in Figure 2. The minimum number of edges from the root to a hedging/refuting word ("not" in the example) is three. The dependency parses were obtained using Stanford CoreNLP (Manning et al., 2014) and the word lists were compiled using online resources.

Claim-headline features While the article headline often provides adequate features to classify its stance, we also need to take into account its entailment relation with the claim. Therefore, based on the work by Rus and Lintean (2012) we compute an alignment using the Paraphrase Database (PPDB) (Pavlick et al., 2015) and the Kuhn-Munkres algorithm (Kuhn, 1955; Munkres, 1957) as follows. For each word pairing between the claim and the headline an edge is created and assigned a score by the following scheme:

- if the stems of the words are identical, assign *maxScore*
- else, if the words are paraphrases according to PPDB, assign their maximum paraphrase score
- else, assign *minScore*

[3]Specifically, we used the sklearn `LogisticRegression` classifier with the default parameters, and L_1 penalization.

maxScore and *minScore* were set to +10 and -10 respectively. Running the Kuhn-Munkres algorithm on this graph finds the maximum scoring 1-to-1 word alignment and the score of this alignment, normalized by the length of the claim or headline, whichever is the shorter. An additional feature is extracted to indicate if in an aligned pair of words, one of them — either in the claim or or the article headline — is negated according to the parser. Furthermore, we extracted the subject-verb-object (SVO) triples from the claim and the article headline (typically one in each) and matched them as follows. For each component of the triples we extracted from PPDB the following labels: *equivalence*, *forwardEntailment*, *backwardEntailment*, *independence* or *noRelation*. Thus the matching of an SVO triple in the claim to one in the headline is represented by a concatenation of three labels, each corresponding to the relation between the subjects, the verbs and the objects (**SVO**). Finally, we computed the cosine similarity between the vector representations for the claim and the headline (**word2vec**). The representations were calculated by multiplying the word2vec vectors (Mikolov et al., 2013a) for each word, which we found to perform better than addition. We utilised pre-trained vectors trained on part of the Google News dataset, comprising 300-dimensional vectors for 3 million words and phrases (Mikolov et al., 2013b).

4 Results

Since none of the stance labels dominates the label distribution, we evaluate the performance primarily using accuracy, also reporting per-class Precision and Recall. A majority baseline would achieve 47%, but would always predict *for*. For a better baseline we used the lexical overlap between the claim and the article headline, which we defined as the percentage of the ratio of the number of lemmas in common between them to the number of lemmas in their union. Using the training data we calculated the average overlap for each stance and found that *for* instances exhibit higher overlap, followed by *observing* and then by *against*. Following this, we defined two overlap thresholds, *minFor* and *maxAgainst*. If the overlap of a claim-headline pair is higher than *minFor* it is labeled *for*, if lower than *maxAgainst* it

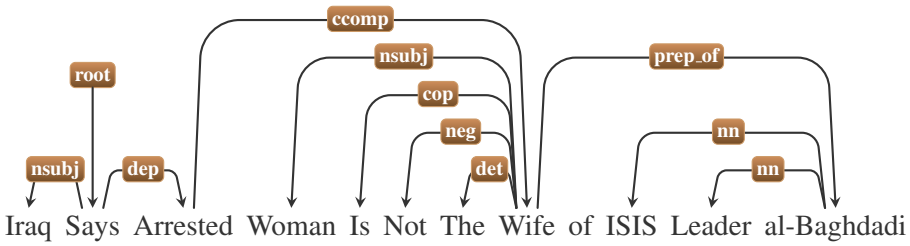

Figure 2: Dependency structure for sentence containing a refuting word.

method	acc.	*for*	*against*	*observing*
overlap	32%	50%/42%	18%/52%	32%/9%
EOP	47%	52%/77%	100%/1%	34%/29%
classifier	73%	71%/89%	82%/70%	74%/54%

Table 1: Test set accuracy and per stance precision and recall.

is labeled *against*, otherwise *observing*.

The comparison between the baseline and the L_1-regularized logistic regression classifier with the features described in the previous section appears in Table 1. As it can be observed, the proposed classifier performs much better in accuracy with substantial gains in all stances. Both approaches are mostly challenged by instances of the *observing* class, since the article headlines with that stance are quite similar to the claim, which is also the case for the more populous *for* class. We also compare our classifier against the Excitement Open Platform (EOP) textual entailment classifier (Magnini et al., 2014). In particular, we used the MaxEntClassificationEDA classifier with the RTE-3 pre-trained model which we found to be the best performing one among those available achieving 33% accuracy. Finally, we trained the same classifier on the Emergent training data achieving 47%, which is 26% lower than the proposed method.

In order to assess the contribution of the features developed we conducted an ablation analysis and the results appear in Table 2. The L_1 regularization used enforces sparsity which helps highlight the features relevant for each stance. The **RootDist** feature has a substantial contribution as it helps distinguish the *observing* from the *for* class. We also evaluated a model using only **BoW**, **Q** and **word2vec** features and the performance was 3% lower than using the complete feature set, thus highlighting the contribution of the features relying on alignment, syntax and

Feature	**10-fold cv**	**test**	chosen for stance
-BoW	1.66%	5.15%	ALL
-Q	1.85%	0.19%	*observing*
-RootDist	2.02%	2.48%	*for*, *observing*
-PPDB	0.47%	0.76%	*for*
-Neg	0.29%	0%	*for*, *against*
-SVO	0.20%	0.19%	*for*, *observing*
-word2vec	0.049%	-0.19%	*against*

Table 2: Ablation results: each row represents the drop in accuracy caused by removing the corresponding feature(s). The last column shows for which stance label(s) the feature(s) had non-zero weight(s).

the PPDB. Finally, the fact that **-word2vec** did not help, especially when compared to **PPDB**, can be partly attributed to the inability of methods relying solely on contexts to learn antonymy.

5 Related work

The task defined by the Emergent dataset differs from recent work in stance classification (Qazvinian et al., 2011; Lukasik et al., 2015; Zhao et al., 2015) not only in the number of claims from which the article headlines are derived, but also in that correct prediction requires considering entailment relation between the claim and the headline. It also differs from work on target-specific stance prediction in debates (Walker et al., 2012; Hasan and Ng, 2013), since the targets considered there are topic labels such as abortion, instead of event claims as in this work.

Emergent, being derived from the workflow of journalists is more realistic than data-sets designed for textual entailment such as FraCas (Cooper et al., 1996) and SICK (Marelli et al., 2014) that are constructed artificially. Compared to the crowdsourced dataset of Bowman et al. (2015), it is smaller but of a different nature, since the former assumes that

all sentences are visual representations, while news tend be more varied.

Stance detection in the context of Emergent is one component in the process of fact-checking claims appearing in the news which are usually more complex than the entity-relation-entity or entity-property-number triples considered in previous work (Nakashole and Mitchell, 2014; Vlachos and Riedel, 2015). The choice of claims to fact-check is a task in its own right, as shown by Hassan et al. (2015). Finally, the only other use of data from the Emergent project is by Liu et al. (2015); however their focus was not on the NLP aspects of the task but on using Twitter data to assess the veracity of the claim, ignoring the articles and their stances curated by the journalists.

6 Conclusions - Future work

In this paper we proposed Emergent, a new real-world dataset derived from the digital journalism project Emergent which can be used for a variety of NLP tasks in the context of fact-checking. We focus on stance detection, for which the large number of claims in the dataset compared to previous work allows for more reliable assessment of the generalization capabilities of the methods evaluated. We proceed to develop a model for stance classification using multiclass logistic regression and show how features beyond the typically used bag of words can be beneficial, achieving accuracy 26% better than an RTE system trained on the same data. We make both the datasets and our code available.

Despite its advantages, the dataset collected is rather small to learn all the nuances of the task. Thus in future work we will explore ways of incorporating large amounts of raw text in training stance classification models, possibly using a neural network architecture. Finally, stance detection is one of the tasks in the fact-checking process of Emergent. In future work we will develop methods for the other tasks involved, such as classifying the stance of a whole article towards a claim and truth assessment.

Acknowledgments

Many thanks to Craig Silverman and the Tow Center for Digital Journalism at Columbia University for allowing us to use the data from the Emergent Project. The research reported in this paper was conducted while the first author was at University College London.

References

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 (EMNLP)*. Association for Computational Linguistics.

Sarah Cohen, Chengkai Li, Jun Yang, and Cong Yu. 2011. Computational Journalism: a call to arms to database researchers. *Proceedings of the 5th Biennial Conference on Innovative Data Systems Research (CIDR 2011) Asilomar, California, USA.*, pages 148–151.

Robin Cooper, Dick Crouch, Jan Van Eijck, Chris Fox, Josef Van Genabith, Jan Jaspars, Hans Kamp, David Milward, Manfred Pinkal, Massimo Poesio, Steve Pulman, Ted Briscoe, Holger Maier, and Karsten Konrad. 1996. Using the framework. Technical report, The FraCas Consortium.

Kazi Saidul Hasan and Vincent Ng. 2013. Stance Classification of Ideological Debates: Data, Models, Features, and Constraints. In *IJCNLP*, pages 1348–1356. Asian Federation of Natural Language Processing / ACL.

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

Harold W. Kuhn. 1955. The Hungarian Method for the assignment problem. *Naval Research Logistics Quarterly*, 2:83–97.

Xiaomo Liu, Armineh Nourbakhsh, Quanzhi Li, Rui Fang, and Sameena Shah. 2015. Real-time rumor debunking on twitter. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, pages 1867–1870.

Michael Lukasik, Trevor Cohn, and Kalina Bontcheva. 2015. Classifying tweet level judgements of rumours in social media. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2590–2595, Lisbon, Portugal, September. Association for Computational Linguistics.

Bernardo Magnini, Roberto Zanoli, Ido Dagan, Kathrin Eichler, Günter Neumann, Tae-Gil Noh, Sebastian Pado, Asher Stern, and Omer Levy. 2014. The excitement open platform for textual inferences. In *Proceedings of the ACL 2014 System Demonstrations*. ACL, 6.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60.

Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. 2014. A sick cure for the evaluation of compositional distributional semantic models. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may. European Language Resources Association (ELRA).

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc.

James Munkres. 1957. Algorithms for the assignment and transportation problems. *Journal of the Society for Industrial and Applied Mathematics*, 5(1):32–38.

Ndapandula Nakashole and Tom M Mitchell. 2014. Language-Aware Truth Assessment of Fact Candidates. *Acl*, pages 1009–1019.

Ellie Pavlick, Pushpendre Rastogi, Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2015. Ppdb 2.0: Better paraphrase ranking, fine-grained entailment relations, word embeddings, and style classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 425–430, Beijing, China, July. Association for Computational Linguistics.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Vahed Qazvinian, Emily Rosengren, Dragomir R. Radev, and Qiaozhu Mei. 2011. Rumor has it: Identifying misinformation in microblogs. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1589–1599.

Vasile Rus and Mihai Lintean. 2012. A comparison of greedy and optimal assessment of natural language student input using word-to-word similarity metrics. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, pages 157–162, Stroudsburg, PA, USA. Association for Computational Linguistics.

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

Andreas Vlachos and Sebastian Riedel. 2014. Fact Checking: Task definition and dataset construction. *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pages 18–22.

Andreas Vlachos and Sebastian Riedel. 2015. Identification and verification of simple claims about statistical properties. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*.

Marilyn Walker, Pranav Anand, Rob Abbott, and Ricky Grant. 2012. Stance Classification using Dialogic Properties of Persuasion. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 592–596.

Zhe Zhao, Paul Resnick, and Qiaozhu Mei. 2015. Enquiring minds: Early detection of rumors in social media from enquiry posts. In *Proceedings of the 24th International Conference on World Wide Web*, pages 1395–1405.