Monolingual Social Media Datasets for Detecting Contradiction and Entailment

Piroska Lendvai*, Isabelle Augenstein, Kalina Bontcheva**, Thierry Declerck***
*Dept. of Computational Linguistics, Saarland University, Germany
**Dept. of Computer Science, University of Sheffield, GB
piroska.r@gmail.com, {i.augenstein,k.bontcheva}@sheffield.ac.uk, declerck@dfki.de

Abstract

Entailment recognition approaches are useful for application domains such as information extraction, question answering or summarisation, for which evidence from multiple sentences needs to be combined. We report on a new 3-way judgement Recognizing Textual Entailment (RTE) resource that originates in the Social Media domain, and explain our semi-automatic creation method for the special purpose of information verification, which draws on manually established rumourous claims reported during crisis events. From about 500 English tweets related to 70 unique claims we compile and evaluate 5.4k RTE pairs, while continue automatizing the workflow to generate similar-sized datasets in other languages.

Keywords: textual entailment, social media, verification

1. Introduction

In this paper, we report on building a special-purpose Recognizing Textual Entailment (RTE) dataset in the context of information verification in user-generated content (Mendoza et al., 2010; Qazvinian et al., 2011; Procter et al., 2013) for the PHEME project[1]. The dataset is compiled based on naturally occurring contradiction in manually labeled claims in crisis events discussed on Twitter, and to our knowledge is the first resource for RTE in the social media and verification domain.

The detection of semantic inference phenomena between natural language text snippets, such as contradiction, entailment, and stance, is targeted by a number of research communities. Its most focused interest group formalizes inference tasks in the generic framework of RTE[2]. RTE is applied to benefit several Natural Language Processing (NLP) tasks, such as information retrieval or text summarization. The task of RTE is to recognise the relationship between sentence pairs, specifically if they entail or contradict each other, or neither of those.

A bottleneck for this task is obtaining training data. The creation of natural language data annotated for inference phenomena is so far a nontrivial and largely manual procedure, yielding expensive resources that are nonetheless problematically portable to new text genres and application domains. Existing initiatives have often created RTE data by syntactic and lexical transformations with predictable effects asking annotators to (re)write sentences taken from gold standards for other tasks such as question answering (Bar-Haim et al., 2006) and image and video description (Bowman et al., 2015; Marelli et al., 2014).

RTE tasks may involve 2-way or 3-way inference judgements. In case of a 2-way judgement, the class to guess is either *Entailment* or *Nonentailment*. On the 3-way judgement scheme the *Nonentailment* class is further differentiated into *Contradiction* and *Unknown*. The presence of contradictory statements in social media can be indicative for mis-/disinformation, controversy or speculation, which are important triggers in veracity checking procedures. Our present contribution therefore addresses the transformation of a project-internal corpus of annotated microblog texts into a 3-way RTE dataset.

Contradiction is the RTE relation when two texts falsify one another's truth value; its typology and an RTE assessment method are set out in (De Marneffe et al., 2008). To represent the phenomenon of contradiction, existing dataset often incorporate artificially created text pairs, generated e.g. by adding negative markers to existing benchmark texts[3]. Recently, the RTE task received attention through a large annotated corpus (Bowman et al., 2015), providing the basis for research on deep models for understanding entailment without the need for manual feature engineering (Wang and Jiang, 2015; Rocktäschel et al., 2016). *Contradiction* pairs in this corpus tend to be rather generic however; for example, "A man inspects the uniform of a figure in some East Asian country." vs "The man is sleeping.", which features a rather broad contrast: 'observing' and 'sleeping' are indeed not plausible to simultaneously take place, so the judgement is justified – but outside of the image captioning task it would not be straightforward to characterize a situation in which this contradiction would naturally emerge (as opposed to the more intuitive pair 'awake' vs 'sleeping').

Next to describing our workflow and the properties of the PHEME RTE pilot dataset we built, we present its first evaluation results by retraining an off-the-shelf RTE classifier on the resource and comparing it with the same classifier trained on RTE benchmark data. We conclude the paper by observing encouraging results for English, and the need to obtain more data in other languages for the method to yield similar-sized output.

2. Building the PHEME RTE dataset

The raw corpus was collected from the Twitter social media platform[4]. It consists of a large number of tweets that report on several world news events, out of which we picked

[1] http://www.pheme.eu/

[2] http://www.aclweb.org/aclwiki/index.php?title=Recognizing_Textual_Entailment

[3] http://www-nlp.stanford.edu/projects/contradiction/

[4] twitter.com

four crisis events: the Charlie Hebdo shooting[5], the Ottawa shooting[6], the Sydney Siege[7], and the Germanwings crash[8]. Tweets were collected by filtering on event-related keywords and hashtags in the Twitter Streaming API.
Each tweet was manually annotated as relating to one specific rumourous claim – a plausible but at a certain point in time officially unconfirmed statement, lexicalized by a concise proposition, e.g. '12 people died in connection with the Charlie Hebdo attack', 'NORAD on high-alert posture', 'The Sydney Opera House has been evacuated', 'There are no survivors in Germanwings crash'. The rumour annotation procedure was performed by journalists as described in Zubiaga et al. (2015). The manually assigned rumourous claim labels were used to create the PHEME RTE data by the following pipeline.

2.1. Language identification

The raw data includes a handful of European languages, out of which we kept only English and German tweets. We adopted a simple NLTK-stopwords[9] based approach implemented by the community[10] that estimates the probability of a given text to be written in a number of languages and selects the highest scoring language.

2.2. Normalization

Data preprocessing involved screen name and hashtag sign removal, URL masking, and selected punctuation removal. Since the manual annotations have been applied irrespective of a tweet supporting or denying a claim, we removed tweets containing lexical items that, when present in a tweet, would reverse the RTE relation between tweet and claim. E.g. the tweet "DEVELOPING: MPs tweeting that gunman has been shot dead. CBC has not confirmed this. Condition of soldier also unknown." contains uncertainty which makes the assumed *Entailment* relationship with its labeled claim Suspected shooter has been killed/is dead invalid. Such tweets were filtered based on a cue list of about twenty items that we obtained from the literature and by observing the data (e.g. 'false', 'wrong', 'not', 'unclear', 'cannot', 'didn't', 'contrar', 'oppose', 'incorrect', 'retract', '?', etc.). A few dozen tweets are removed for each event by this step.

2.3. Creating the Contradiction relation

For each of the four crisis events, we manually identified labeled claims that could be regarded as contradictory, this left us with 1 contradictory claim pair for the *gwings* data, 6 for *ottawa*, 7 for *chebdo* and 8 for *ssiege*. The notion of contradiction was employed in the semantic contrast sense, i..e., the claims regarded as contradictory for the PHEME special-purpose RTE task could have taken place simultaneously in real life, as the rumour pair in the first example that features world-knowledge-level named entity mismatch, or were not produced by tweeters truly simultaneously, as the rumour pair in the second example, featuring lexical-level semantic opposition. It was not our goal to represent how real-life events unfold during a crisis, but to supply linguistic evidence for analyzing contradictory texts.

1. 'Parliament Hill is on lockdown' – 'The University of Ottawa is on lockdown'

2. 'Shooter is still on the loose' – 'Suspected shooter has been killed/is dead'

Contradictory snippet pairs for RTE – termed the *text* and the *hypothesis* – were generated by pairing each of the tweets annotated with a certain claim with each of the tweets annotated by its manually identified counterpart claim. Directionality does not hold for our current project purpose; to conform to the RTE format, the longer tweet was chosen to be the *text* (*t*), the shorter tweet was designated to be the *hypothesis* (*h*). The procedure resulted in *Contradiction* pairs such as

- <t>12 people now known to have died after gunmen stormed the Paris HQ of magazine CharlieHebdo URL URL</t> <h>Awful. 11 shot dead in an assault on a Paris magazine. URL CharlieHebdo URL</h>

- <t>Several MPs tweeting that lone gunman shot dead in Centre Block. All MPs reportedly safe. cdnpoli ottawa</t> <h>More shots being fired near parliament in Ottawa, suspect still at large: TV</h>

2.4. Creating the Entailment relation

We assumed that tweets annotated with one and the same claim would be entailing each other's content. Positive entailment judgement cases were created by pairing tweets belonging to those claims based on which the *Contradiction* set was made. This restriction is assumed to keep the final dataset balanced across the three entailment judgment instances, and to enable systematic feature assessment in classification experiments. Examples of the resulting *Entailment* pairs are

- <t>Germanwings Airbus A320 en route from Barcleona to Dusseldorf crashes in southern French Alps - 148 people on board URL</t> <h>Received news that a Germanwings Airbus A320 plane crashed in southern France, carrying 142 passengers + 6 crew onboard.</h>

- <t>SYDNEY ATTACK - Hostages at Sydney cafe - Up to 20 hostages - Up to 2 gunmen - Hostages seen holding ISIS flag DEVELOPING..</t> <h>Up to 20 held hostage in Sydney Lindt Cafe siege URL URL</h>

2.5. Creating the Unknown relation

The third class in the data carries the neutral judgement label called *Unknown*, because the tweets in such a pair are neither entailing nor contradicting each other. The two text snippets might be topically related (as in the PHEME

[5]https://en.wikipedia.org/wiki/Charlie_Hebdo_shooting
[6]https://en.wikipedia.org/wiki/2014_shootings_at_Parliament_Hill,_Ottawa
[7]https://en.wikipedia.org/wiki/2014_Sydney_hostage_crisis
[8]https://en.wikipedia.org/wiki/Germanwings_Flight_9525
[9]http://www.nltk.org/book/ch02.html
[10]http://blog.alejandronolla.com/2013/05/15/detecting-text-language-with-python-and-nltk/

event	ENT	CD	UNK	#uniq clms	#uniq tws
chebdo	647	427	866	27	199
gwings	461	257	447	4	29
ottawa	555	377	168	18	125
ssiege	332	317	565	21	143
total	1995	1378	2046	70	496

Table 1: The PHEME RTE English dataset compiled from 4 crisis events: amount of pairs per entailment type (*ENT, CD, UNK*), amount of unique rumourous claims (*#uniq clms*) used for creating the pairs, amount of unique tweets corresponding to claims (*#uniq tws*).

dataset), or they might be unrelated, as in classical RTE data.

The pairs labeled as *Unknown* in PHEME RTE data were built by taking all claims that received the *Contradiction* label, pairing each of them with a randomly chosen claim in the raw dataset that was not part of the contradictory claim set. For example, the below claim pairs are regarded to express the neutral relation.

- 'The Sydney airspace has been closed' – 'A police officer has a gunshot wound to the head/is injured'
- 'At least two dead in hostage-taking in Porte de Vincennes' – 'Kosher restaurants /Jewish shops (and schools, synagogues, etc.) are closing in Paris in wake of Porte de Vincennes hostage-taking'.

The resulting *Unknown* pairs are e.g.

- <t>BREAKING: NSW police have confirmed the siege in Sydney's CBD is now over, a police officer is reportedly among the several injured.</t> <h>Update: Airspace over Sydney has been shut down. Live coverage: URL sydneysiege</h>
- <t>Update - AFP reports at least two people killed after shooting at kosher grocery in eastern Paris in which at least five were taken hostage</t> <h>BREAKING: Police order all shops closed in famed Jewish neighborhood in central Paris far from attacks.</h>

3. PHEME RTE pilot dataset

The characteristics of the PHEME RTE dataset are shown in Table 1. From about 500 English tweets related to 70 unique claims we compiled 5.4k RTE pairs. The approach yields only a handful of RTE pairs for our second targeted language, German, as we have a disproportionally small amount of German tweets in the raw data so far; these belong to the few contradictory claim pairs identified for the *gwings* event.

3.1. Pilot assessment

To assess the PHEME RTE pilot dataset, we use the MaxEnt classifier-based model (Wang and Neumann, 2007) distributed with the Excitement Open Platform (EOP, Pado et al. 2014)[11]. The benefit of the entailment platform is the use of external resources for training, which can lead to a relatively high performance even with a small training data size. The MaxEnt model is augmented with lexical resources shipped with EOP (WordNet and VerbOcean), and uses the output of part-of-speech and dependency parsing in its structure-oriented approach for classification.

test	train	P	R	F1
chebdo	rte3	.3329	.8609	**.4802**
	pheme	.5734	.5907	.5639
gwings	rte3	.5972	.3742	.2731
	pheme	.6328	.6120	**.6207**
ottawa	rte3	.4557	.4718	.3908
	pheme	.4766	.3291	.3260
ssiege	rte3	.3294	.2792	.1780
	pheme	.5715	.5717	.5388

Table 2: Evaluation of the English PHEME RTE dataset with the Excitement MaxEnt model, retrained on the respective training corpora. Precision, recall and F1 measurements are averaged over the 3 RTE labels (Entailment, Contradiction, Unknown), weighted by support (the number of true instances for each label).

Data from the four distinct events in the PHEME RTE dataset can be conveniently used in different training/test scenarios; for our pilot evaluation we choose a leave-one-event-out setup, always training on three datasets and testing on the held-out dataset. The results are reported in Table 2[12]. To compare the scores obtained from training and testing on the PHEME data, we have retrained the MaxEnt model on the RTE-3 development dataset[13] and tested it against the data in the four events.

We observe for both cross-validation runs that the scores vary depending on which event is used for testing. Training and testing on PHEME data performs between 33 (*ottawa*) - 62 (*gwings*) F score points, training on RTE-3 and testing on PHEME yields between 18 (*ssiege*) - 48 (*chebdo*) F score points. We hypothesize that the poorer performance when training on RTE-3 is to be explained by portability issues between the RTE-3 data and the PHEME data properties in terms of the newswire vs social media genres, as well as the generic RTE vs special-purpose RTE scenario. We are going to report on specific error analysis and performance comparison in follow-up publications.

4. Discussion and Future Work

We reported on a new 3-way-judgement RTE resource for the Natural Language Processing, Social Media, and Recognizing Textual Entailment communities that enables the development of statistical approaches for end tasks drawing on semantic inference across microblog texts. The RTE pairs are built from naturally occurring data by a method that is portable across languages and domains, but

[11] http://hltfbk.github.io/Excitement-Open-Platform/

[12] generated by the sklearn metric classification_report, see http://scikit-learn.org

[13] http://nlp.stanford.edu/RTE3-pilot/RTE3_dev_3class.xml

requires event and claim annotations. The manual effort spent to create such annotations is feasible to replace by automatic means which are currently being implemented in the project.

Bentivogli et al. (2010) stress the importance of creating specialized data sets for RTE, in order to facilitate more targeted assessment and decomposition of the RTE task's complexity. In our resource, the text snippets that form a RTE pair deliberately keep reoccurring across all three judgement labels in systematically varied pairings, allowing to investigate, model and evaluate linguistic and extra-linguistic phenomena that underly semantic inference in the misinformation detection scenario.

Previous RTE research has mainly focused on achieving good performance on the *Entailment* relation, whereas our method is motivated by the need for a resource that facilitates the development of statistical processing approaches specifically targeting the *Contradiction* relation. Therefore, the procedure we used for building the PHEME RTE dataset is centered on contradictory claims present in the data, and is extended to the other two classes to a limited extent. The resulting pilot dataset is balanced across the three classes. In future work we will investigate and evaluate the relevance of our data and compilation approach with respect to the RTE-5 Entailment Search pilot task[14] and the RTE-6 Entailment Summarisation task[15], in which RTE systems are required to find all sentences in a document or a set that entail a given Hypothesis.

RTE and its resources also tend to be utilized in the recently emerging task of stance detection (Mohammad et al., 2016), i.e. classification of the standpoint of an expression such as "Climate change is a real concern" towards a piece of (social media) text as either supportive, denying, or neutral (Augenstein et al., 2016; Ferreira and Vlachos, 2016). It remains to be evaluated if the approaches built for stance detection are reusable or need specific adaptation to our goal of RTE in social media verification.

Our current efforts include further development of the reported approach and the curation of project-internal data in other languages, in order to release[16] several monolingual RTE benchmark resources.

5. Acknowledgements

Work presented in this paper has been supported by the PHEME FP7 project (grant No. 611233).

6. Bibliographical References

Augenstein, I., Vlachos, A., and Bontcheva, K. (2016). USFD: Any-Target Stance Detection on Twitter with Autoencoders. In Saif M. Mohammad, et al., editors, *Proceedings of the International Workshop on Semantic Evaluation*, SemEval '16, San Diego, California.

Bar-Haim, R., Dagan, I., Dolan, B., Ferro, L., Giampiccolo, D., Magnini, B., and Szpektor, I. (2006). The Second PASCAL Recognising Textual Entailment Challenge. In *Proceedings of the Second PASCAL Challenges Workshop on Recognising Textual Entailment*.

Bentivogli, L., Cabrio, E., Dagan, I., Giampiccolo, D., Leggio, M. L., and Magnini, B. (2010). Building Textual Entailment Specialized Data Sets: a Methodology for Isolating Linguistic Phenomena Relevant to Inference. In Nicoletta Calzolari, et al., editors, *LREC*. European Language Resources Association.

Bowman, S. R., Angeli, G., Potts, C., and Manning, C. D. (2015). A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal, September. Association for Computational Linguistics.

De Marneffe, M.-C., Rafferty, A. N., and Manning, C. D. (2008). Finding contradictions in text. In *ACL*, volume 8, pages 1039–1047.

Ferreira, W. and Vlachos, A. (2016). Emergent: a novel data-set for stance classification. In *Proceedings of NAACL*.

Marelli, M., Menini, S., Baroni, M., Bentivogli, L., Bernardi, R., and Zamparelli, R. (2014). A sick cure for the evaluation of compositional distributional semantic models. In *LREC*, pages 216–223.

Mendoza, M., Poblete, B., and Castillo, C. (2010). Twitter Under Crisis: Can We Trust What We RT? In *Proceedings of the First Workshop on Social Media Analytics (SOMA'2010)*, pages 71–79, New York, NY, USA. ACM.

Mohammad, S. M., Kiritchenko, S., Sobhani, P., Zhu, X., and Cherry, C. (2016). SemEval-2016 Task 6: Detecting stance in tweets. In *Proceedings of the International Workshop on Semantic Evaluation*, SemEval '16, San Diego, California.

Procter, R., Vis, F., and Voss, A. (2013). Reading the riots on Twitter: methodological innovation for the analysis of big data. *International Journal of Social Research Methodology*, 16(3):197–214.

Qazvinian, V., Rosengren, E., Radev, D. R., and Mei, Q. (2011). Rumor has it: Identifying misinformation in microblogs. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 1589–1599.

Rocktäschel, T., Grefenstette, E., Hermann, K. M., Kočiskỳ, T., and Blunsom, P. (2016). Reasoning about Entailment with Neural Attention. In *International Conference on Learning Representations (ICLR)*.

Wang, S. and Jiang, J. (2015). Learning Natural Language Inference with LSTM. *CoRR*, abs/1512.08849.

Wang, R. and Neumann, G. (2007). Recognizing textual entailment using a subsequence kernel method. In *AAAI*, volume 7, pages 937–945.

Zubiaga, A., Liakata, M., Procter, R., Bontcheva, K., and Tolmie, P. (2015). Towards Detecting Rumours in Social Media. *CoRR*, abs/1504.04712.

[14] http://www.nist.gov/tac/2009/RTE/RTE5_Pilot_Guidelines.pdf

[15] http://www.nist.gov/tac/2010/RTE/

[16] http://www.pheme.eu/software-downloads/