DELIVERABLE SUBMISSION SHEET

To: Susan Fraser *(Project Officer)*

EUROPEAN COMMISSION
Directorate-General Information Society and Media
EUFO 1165A
L-2920 Luxembourg

From:

Project acronym: PHEME Project number: 611233

Project manager: Kalina Bontcheva

Project coordinator The University of Sheffield (USFD)

The following deliverable:

Deliverable title: Algorithms for Detecting Disputed Information: Final Version

Deliverable number: D4.2.2

Deliverable date: 31 March 2016

Partners responsible: Universitaet de Saarlands (USAAR)

Status: ☒ Public ☐ Restricted ☐ Confidential

is now complete. ☒ It is available for your inspection.
☒ Relevant descriptive documents are attached.

The deliverable is:

☐ a document
☐ a Website (URL: .....)
☐ software (.....)
☐ an event
☒ other (Prototype)

Sent to Project Officer:	Sent to functional mail box:	On date:
Susan.Fraser@ec.europa.eu	*CNECT-ICT*-611233*@ec.europa.eu*	31 May 2016

Algorithms for Detecting Disputed Information: Final Version

Piroska Lendvai (USAAR), Isabelle Augenstein (USFD), Dominic Rout (USFD), Kalina Bontcheva (USFD), Thierry Declerck (USAAR)

Abstract.
FP7-ICT Collaborative Project ICT-2013-611233 PHEME
Deliverable D4.2.2 (WP4)

This deliverable reports on building and evaluating resources for contradiction detection in the context of information verification in user-generated content for the PHEME project. The resources include special-purpose Recognizing Textual Entailment (RTE) datasets and machine learning-based systems trained on these datasets. The systems are evaluated in terms of baseline strategies, utilizing existing RTE collections and systems, and compared with scores reported in the literature. The best scores on classifying contradictory text pairs achieve comparable F-scores across the datasets and the systems that we have investigated. We conclude that our results on contradiction detection in social media data are state-of-the-art.
Keyword list: textual entailment, social media, verification

Project	PHEME No. 611233
Delivery Date	May 31, 2016
Contractual Date	March 31, 2016. Extension granted to May 31, 2016
Nature	Prototype
Reviewed By	Laura Tolosi-Halacheva (Ontotext)
Web links	http://www.pheme.eu/wp-content/uploads/2016/04/pheme_rte_datasets_2016.zip
Dissemination	PU

PHEME **Consortium**

This document is part of the PHEME research project (No. 611233), partially funded by the FP7-ICT Programme.

University of Sheffield
Department of Computer Science
Regent Court, 211 Portobello St.
Sheffield S1 4DP
UK
Contact person: Kalina Bontcheva
E-mail: K.Bontcheva@dcs.shef.ac.uk

MODUL University Vienna GMBH
Am Kahlenberg 1
1190 Wien
Austria
Contact person: Arno Scharl
E-mail: scharl@modul.ac.at

ATOS Spain SA
Calle de Albarracin 25
28037 Madrid
Spain
Contact person: Tomás Pariente Lobo
E-mail: tomas.parientelobo@atos.net

iHub Ltd.
NGONG, Road Bishop Magua Building
4th floor
00200 Nairobi
Kenya
Contact person: Rob Baker
E-mail: robbaker@ushahidi.com

The University of Warwick
Kirby Corner Road
University House
CV4 8UW Coventry
United Kingdom
Contact person: Rob Procter
E-mail: Rob.Procter@warwick.ac.uk

Universitaet des Saarlandes
Campus
D-66041 Saarbrücken
Germany
Contact person: Thierry Declerck
E-mail: declerck@dfki.de

Ontotext AD
Polygraphia Office Center fl.4,
47A Tsarigradsko Shosse,
Sofia 1504, Bulgaria
Contact person: Georgi Georgiev
E-mail: georgiev@ontotext.com

King's College London
Strand
WC2R 2LS London
United Kingdom
Contact person: Robert Stewart
E-mail: robert.stewart@kcl.ac.uk

SwissInfo.ch
Giacomettistrasse 3
3000 Bern
Switzerland
Contact person: Peter Schibli
E-mail: Peter.Schibli@swissinfo.ch

Executive Summary

In this deliverable, we report on building and evaluating resources for contradiction detection in the context of information verification in user-generated content for the PHEME project. The resources include special-purpose Recognizing Textual Entailment (RTE) datasets and algorithms trained on these datasets. The newly built datasets are available at the Pheme page[1].

In Chapter 1, we describe the RTE framework and its relation to contradiction detection. An overview is provided about existing data collections and tools for contradiction detection. We subsequently report on extensive experiments on contradiction detection: in Chapter 2 we describe the transformation of two project-internal corpora of annotated microblog texts into 3-way RTE collections, in which tweet pairs are assigned a label that characterizes their relationship in terms of three possible categories: Entailing, Contradicting, and Unknown. In Chapter 3, the collections are put to use to retrain two systems that are based on machine learning algorithms: a maximum entropy-based system from an open source RTE platform and a logistic-regression-based system that utilizes word embeddings. Both systems are evaluated in terms of baseline strategies, utilizing existing RTE collections and systems, and compared with scores reported in the literature.

We find that collecting and utilizing two different types on contradiction data originating from the same social media platform helped gain new insights into the nature of contradictions, and supported the evaluation of different contradiction types appearing in social media, and that it was crucial for classification robustness to identify, build and evaluate both types of data.

The best scores on classifying contradictory text pairs achieve comparable F-scores across the datasets and the systems that we have investigated. We conclude that our results on contradiction detection in social media data are state-of-the-art.

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

Contents

Chapter 1

Contextualizing contradiction detection

In this deliverable, we report on building and evaluating resources for contradiction detection 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. The resources include special-purpose Recognizing Textual Entailment (RTE) datasets and algorithms trained on these datasets. The newly built data sets are available at the Pheme page[1]

Regarding contradiction as a building block of rumourousness, the presence of contradictory claims in social media posts can be indicative of misinformation, disinformation, controversy or speculation, which are important elements of factuality assessment and related veracity checking procedures. Our aim is to produce resources that allow for identifying contradictory claims about newly emerging events reported on social media platforms.

1.1 Recognizing Textual Entailment (RTE)

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. An entailment relation holds between two text snippets if the claim present in snippet A is present in snippet B. The contradiction relation applies when the claim in A and the claim in B cannot be simultaneously true. Entailment recognition approaches are useful for application domains such as information extraction, question answering or summarization, for which evidence from multiple sentences needs to be combined.

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

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

A typology of contradictions, and a complex system for detecting contradictions in the RTE sense are set out in De Marneffe et al. (2008) who state that applications that benefit from contradiction detection are those that seek to highlight discrepancies or incompatibility in descriptions of the same event (De Marneffe et al., 2008). The PHEME Journalism usecase fits exactly this scenario, whereby it aims to support the daily work of journalists via alerting for controversial or unreliable information appearing on social media platforms. Our approach is thus to perform contradiction detection within the RTE setup, applying it to social media texts.

A bottleneck for the contradiction detection task is obtaining training data. From the literature one can observe that the creation of natural language data annotated for inference phenomena is a nontrivial and largely manual procedure, yielding expensive resources that are nonetheless not straightforward to port to new text genres and application domains. Existing initiatives have often created RTE data via syntactic and lexical transformations with predictable effects, asking annotators to (re)write sentences taken from gold standards[3] 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). We could not identify research datasets available for RTE on microblogs data in general, and for contradiction detection in such data in particular.

This deliverable reports on extensive experiments on contradiction detection: we describe the transformation of two project-internal corpora of annotated microblog texts into 3-way RTE collections, in which tweet pairs are assigned a label that characterizes their relationship in terms of three possible categories: Entailing, Contradicting, and Unknown. The collections are put to use to retrain two instances of machine learning algorithms: the maximum entropy algorithm from the Excitement Open Platform (EOP, **?**)[4], as well as a logistic-regression-based system that utilizes word embeddings. Both the maxent-based system and the word embeddings-based system is evaluated in terms of baseline strategies, utilizing existing RTE collections and systems, and compared with scores reported in the literature.

1.2 Contradictions in microblog data

In the PHEME project it is so far data from Twitter that is being analyzed and utilized. We observe two different contexts on Twitter, in which contradictions may emerge: threaded discussions and independently posted tweets. In threaded discussions the unit of RTE focus is a source tweet – reply tweet pair, where the reply tweet is a direct response given to the source tweet, and the response typically expresses immediate denial, rejection, debunking and related phenomena as a reaction to the information (claim, statement) present in the source tweet. In independently posted tweets, contradictions would typically be

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

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

emerging in a larger discourse setting, i.e. across time, across documents or threads, or across sources. In such source tweet – source tweet pairs, two individual positive claims are present, typically without explicit, classical rejection markers or contradiction-related cues in terms of modality and speculation phenomena.

We hypothesize that the textual data, i.e. the tweet pairs related to these contexts are differently shaped both in terms of general language phenomena, and in terms of contradiction-related phenomena. Manual analysis pointed us to distributional differences too, i.e. that in naturally occurring data explicit contradictions tend to have a lower frequency than implicit ones. Explicit contradictions are exemplified in Figure 1.1, implicit contradictions in Figure 1.2. Furthermore, we observe from the literature that there is less evidence for contradictions in the data relative to the other two RTE relation types in a corpus. The ratio of contradictions in existing RTE sets is about 10-15%. We therefore aimed at collecting and utilizing contradiction data form PHEME from both scenarios, and aim to boost the amount of contradiction pairs.

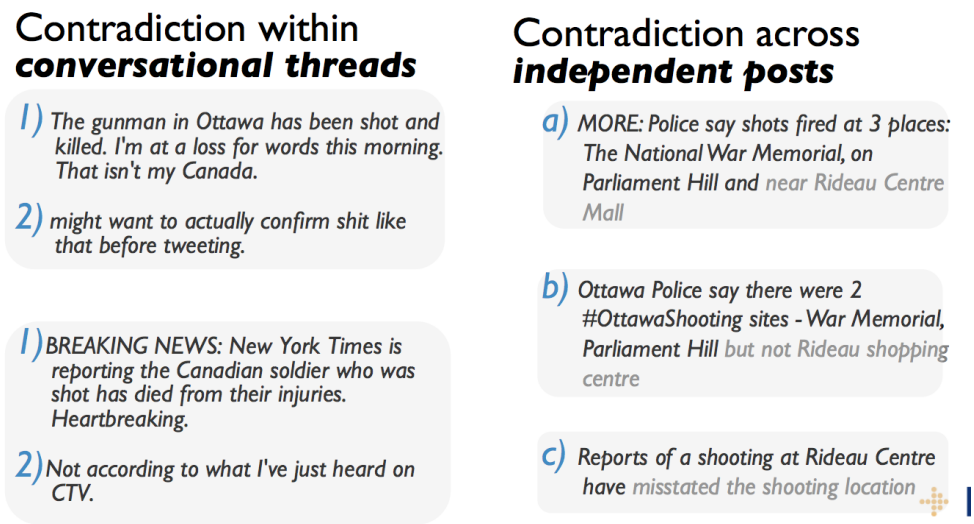

Figure 1.1: Explicit contradictions in threaded discussions and in independent posts in the PHEME Twitter collection.

1.3 Previous work and benchmarks

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; for example, ”A man inspects the uniform of a figure in some East Asian country.” vs ”The man is

Figure 1.2: Implicit contradictions in threaded discussions and in independent posts in the PHEME Twitter collection.

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').

The RTE-3 dataset is the first resource that labels paired text snippets in terms of 3-way RTE judgements, and is comprised of general newswire texts rather than non-social media data, and had been previously annotated for contradiction by De Marneffe et al. (2008) at Stanford. This data is taken from the corpora provided for the first three PASCAL Recognising Textual Entailment (RTE) challenges, which focusses on web-based news (**???**). This data will be referred to as the Stanford contradictions corpus.

RTE and its resources 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).

We are aware of two resources of stance detection data that could be mapped to the RTE task: the Emergent dataset of news headlines[5], and the SemEval 2016 Twitter Stance Detection corpus[6], which is a resource containing social media posts.

In the contradiction literature, the detailed definition of what counts as a contradiction is often project- and task-dependent, which may lead to marked differences in the semantics of existing datasets that contain contradiction(-related) instances. Such differences influence performance scores in automated contradiction detection, and make it difficult

[5]http://eprints.whiterose.ac.uk/97416/
[6]http://alt.qcri.org/semeval2016/task6/

to directly perform cross-project evaluation.

In Figure 1.3 we provide a simple approach to comparing the textual similarity between the RTE-3 development dataset's RTE classes and the Emergent project's dataset's RTE classes. Textual similarity is assessed in terms of the longest overlapping unnormalized token sequence (LCS) ratio[7].

The boxplots corresponding to development/training corpora of the two projects (RTE-3 and Emergent) show that the distribution of LCS values that characterize similarity in terms of lexical overlap are roughly in line across the stance vs contradiction task. Although the differences between each pairs of the three classes (ENT, CON, UNK) are statistically significant, there is considerable token overlap across the classes, which signals that separating the three classes based on shared tokens and alignment phenomena – in line with classical RTE methods – is a difficult task.

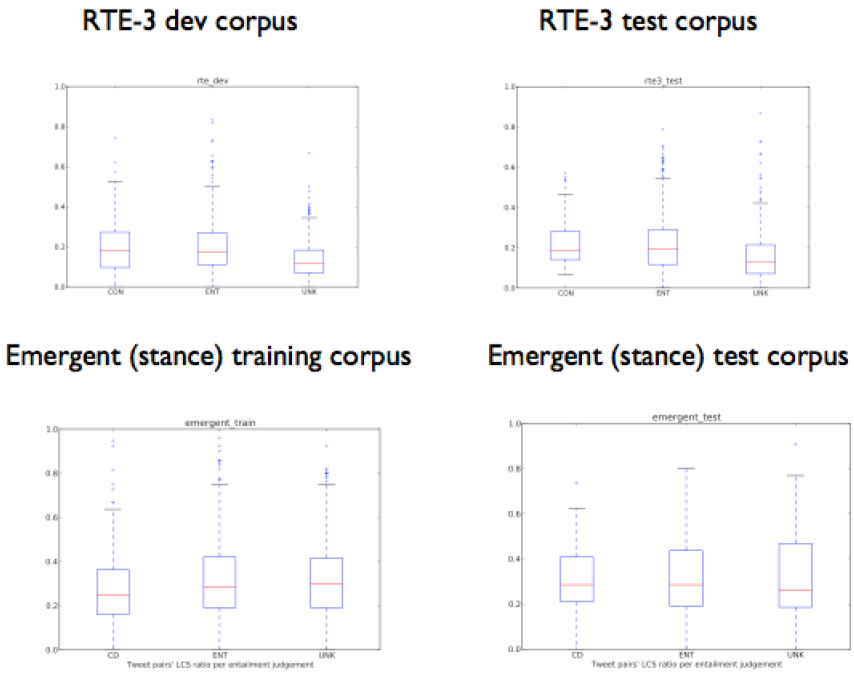

Figure 1.3: Textual similarity in terms of the longest overlapping token sequence (LCS) ratio in four external datasets: RTE-3 development and test corpus, Emergent project training and test corpus.

As regards to tools for contradiction detection, existing approaches utilized statistical models, i.e. supervised machine learning.

The EOP platform integrates several entailment decision algorithms, that emerged in the past decade. Out of these, the Maximum Entropy-based model (Wang and Neumann,

[7]https://en.wikibooks.org/wiki/Algorithm_Implementation/Strings/Longest_common_subsequence

2007) is available for 3-way RTE classification, which is an improvement over the classical binary Entailment vs Nonentailment RTE scenario. This model implements state-of-the-art linguistic preprocessing augmented with lexical resources (WordNet, VerbOcean), and uses the output of part-of-speech and dependency parsing in its structure-oriented, overlap-based approach for classification.

A standalone contradiction detection system was implemented by De Marneffe et al. (2008). This system uses a number of complex rule-based features which are combined using hand-set weights to carry out contradiction detection. Due to the complexity of the features and rules in this system, it would be extremely difficult to adapt it successfully to the domain of tweets.

1.4 Relevance to PHEME

1.4.1 Relevance to project objectives

We report on 3-way-judgement RTE datasets that are used in the development of statistical approaches for end tasks drawing on semantic inference across microblog texts. The RTE text pairs are built from naturally occurring social media data by a method that is portable across languages and domains, but 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.

Our current efforts include further development of the reported approach and the curation of project-internal data in other languages, in order to release[8] several monolingual RTE benchmark resources. The training of standalone classifiers is additionally ongoing within the project.

1.4.2 Relation to other work packages

The released Pheme RTE dataset is created using manual annotations assigned within WP8 by the Swissinfo partner. We compare the utility of this dataset with a potentially useful second resource that is generated from threaded discussions, utilizing manual annotations assigned by the University of Warwick partner in WP2, more specifically reported

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

in T2.1 corresponding to D2.4 ”Qualitative Social Science Analysis of Rumour across Media and Languages” and the final Twitter annotated corpus that was released[9].

In the context of WP4 ”Detecting Rumours and Veracity”, the Ontotext partner is planning to use contradiction detection output for rumour detection and fact-checking, utilizing inference. The PHEME RTE dataset reported as released in this deliverable can already be imported in Ontotext’s GraphDB, while the output of the EOP classifier may need to be tailored to the PHEME pipeline.

[9] https://figshare.com/articles/PHEME_rumour_scheme_dataset_journalism_use_case/2068650

Chapter 2

Development of social media RTE collections

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 first PHEME RTE dataset is centered on contradictory claims present in the independent posts data, and is extended to the other two classes to a limited extent. The resulting independent posts dataset is balanced across the three classes. The dataset was made publicly available as reported in Lendvai et al. (2016). We compare this dataset to a resource that we generated from threaded discussions, drawing on project-internal annotations.

The raw corpus in both cases was collected from the Twitter social media platform[1]. It consists of a large number of tweets that report on several world news events, out of which we picked four crisis events: the Charlie Hebdo shooting[2] (*chebdo*), the Ottawa shooting[3] (*ottawa*), the Sydney Siege[4] (*ssiege*), and the Germanwings crash[5] (*gwings*). Tweets were collected by filtering on event-related keywords and hashtags in the Twitter Streaming API.

2.1 Contradiction within independent posts

In the corpus, each tweet was manually annotated as relating to one specific rumourous claim – a plausible but at a certain point in time officially unconfirmed statement, lexical-

[1] twitter.com

[2] https://en.wikipedia.org/wiki/Charlie_Hebdo_shooting

[3] https://en.wikipedia.org/wiki/2014_shootings_at_Parliament_Hill,_Ottawa

[4] https://en.wikipedia.org/wiki/2014_Sydney_hostage_crisis

[5] https://en.wikipedia.org/wiki/Germanwings_Flight_9525

ized 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.

2.1.1 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[6] based approach implemented by the community[7] that estimates the probability of a given text to be written in a number of languages and selects the highest scoring language.

2.1.2 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 hundred contradictory tweet pairs may be removed for each event by this step. Significantly, in this step every contradiction instance that involves explicit contradiction in terms of negation cues is removed from the data. This implies that only subtler and implicit contradiction are contained in the resulting PHEME RTE dataset, increasing the specificity of this resource.

2.1.3 Creating the Contradiction relation

For each of the four crisis events, we manually paired labeled claims ('Stories') that could be regarded as contradictory. In the datasets, 10-26% of all identified claims were judged contradictory: 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,

[6] http://www.nltk.org/book/ch02.html

[7] http://blog.alejandronolla.com/2013/05/15/detecting-text-language-with-python-and-nltk/

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. Since the goal was to detect that the targeted information in two text snippets is contradictory, even though the two sentences can be true simultaneously, we aimed to supply linguistic evidence for analyzing contradictory texts, and not to represent how real-life events unfold during a crisis. The below pairs exemplify contradictory claims.

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 tweet 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 did not hold for our 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>

2.1.4 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>

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

The pairs labeled as unknown 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>

The characteristics of the PHEME RTE independent posts dataset are shown in Table 2.1. From about 500 English tweets related to 70 unique claims we compiled 5.4k RTE pairs. The proportion of contradiction instances in the dataset is 25%. The approach yielded only a handful of RTE pairs for our second targeted language, German, as there is a disproportionally small amount of German tweets in the claim-annotated data; these belong to the few contradictory claim pairs identified for the *gwings* event.

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 2.1: The PHEME RTE independent posts 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*).

2.2 Contradiction within conversational threads

The RTE data we generated from Twitter conversational threads uses English data published in the final Twitter annotated corpus (D2.4: "Qualitative Analysis of Rumours, Sources, and Diffusers across Media and Languages")[8]. The conversational threads RTE dataset was created based on the manually assigned Response Type labels that characterize Source Tweet – Reply Tweet relations in terms of four categories. We mapped these four categories onto three RTE labels: *Agreed* was mapped to *Entailment*, *Disagreed* was mapped to *Contradiction*, *AppealforMoreInfo* and *Comment* were mapped to *Unknown*. Only direct replies to source tweets relating to the same four events as in the independent posts RTE dataset were kept. The characteristics of the obtained dataset are described in Table 2.2. The proportion of contradiction instances in this dataset amounts to 7% only.

The annotations of the conversational threads dataset are closer to the stance detection task than to the RTE task, meaning that the obtained resource contains more noise than the independent posts RTE dataset. In the original annotations, Response Type was used to designate the support of response tweets towards a source tweet that introduces a rumourous story, where the value of *Agreed* applies when the author of the response supports the statement they are responding to, *Disagreed*, when the author of the response disagrees with the statement they are responding to, etc. (cf. pp. 36 of deliverable D2.4).

The amount of tweet pairs is three times larger in the independent posts dataset (5,419 pairs) than in the threads dataset (1,850 pairs), but the textual data are much more varied in the threads dataset. Figure 2.1 shows the distribution of the three classes in both datasets.

In Figure 2.2 we provide a comparison of textual similarity in terms of the longest overlapping (unnormalized) token sequence (LCS) ratio between the threads dataset and the independent posts dataset. We observe differences per entailment judgement between

[8] https://figshare.com/articles/PHEME_rumour_scheme_dataset_journalism_use_case/2068650

event	ENT	CD	UNK	#uniq clms	#uniq tws
chebdo	143	34	486	36	736
gwings	39	6	107	13	176
ottawa	79	37	292	28	465
ssiege	112	59	456	37	697
total	373	136	1341	114	2074

Table 2.2: Pheme threads RTE 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*).

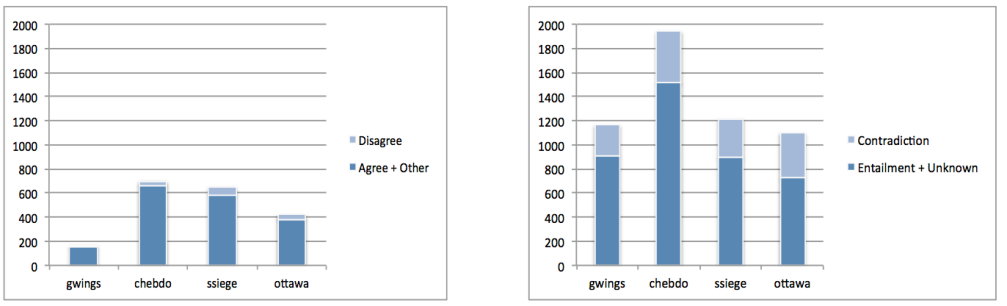

Figure 2.1: Stacked counts of contradiction pairs vs entailment and unknown pairs in the threads (left) vs independent posts (right) datasets.

the two sets, as well as collection-internally, that may point at the inherent differences between the collection origin scenario as well as the annotation's end goal (stance detection data vs RTE data).

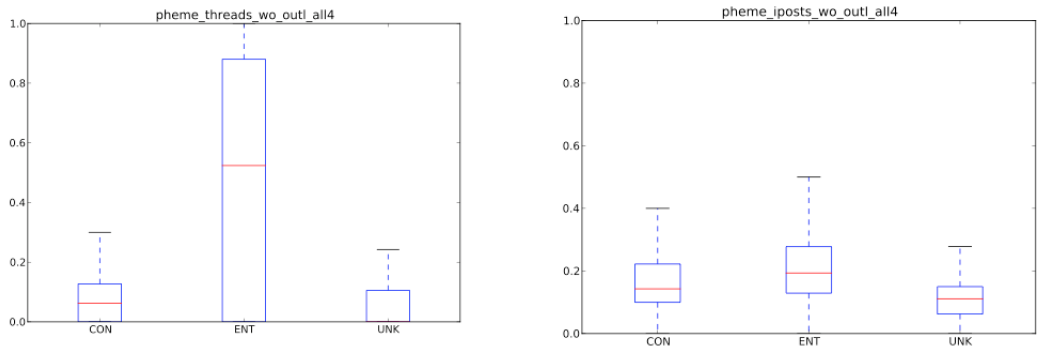

Figure 2.2: Textual similarity between the threads dataset and the independent posts dataset in terms of LCS ratio.

The workflow of the data creation approach is semi-automatic. The currently implemented method requires PHEME metadata (claim annotations, response annotations, contradictory claim identification) that so far were manually assigned. The automatization of the metadata creation is ongoing in the project, and is going to be extended to data in German.

The PHEME RTE dataset can be downloaded from `http://www.pheme.eu/wp-content/uploads/2016/04/pheme_rte_datasets_2016.zip` under a CC-BY license, while Twitter retains the ownership and rights of the content of the tweets.

Chapter 3

Development of contradiction detection algorithms

3.1 Experiments with EOP

One of our setups for contradiction detection in the PHEME data is pairwise classification of tweets: tweet pairs labelled with one of the three entailment judgements are passed to the EOP platform, where the text snippets in the pair are analysed on various linguistic levels (e.g. token, lemma, part-of-speech, syntactic chunks and structure, named entities, lexical semantic information) and scored in terms of how well these linguistic phenomena can be aligned in the two snippets.

The maximum entropy classifier (*maxent*) in EOP is based on a prototype system called TIE (Textual Inference Engine) developed in the Language Technology lab of DFKI GmbH[1]. maxent uses linguistic alignment scores in the training phase as features to learn a classification model. In the test phase, the same features are extracted from unseen pairs in the test data, and are used to guess the relation label of unseen test pairs. The algorithm works in a language-independent way.

We have used the EOP platform's version 1.2.3 for our experiments, in which linguistic preprocessing is carried out by the DKPro tool[2] within the UIMA framework[3]. We used maxent with the following linguistic preprocessing settings: bag-of-words scoring, bag-of-lemmas scoring, bag-of-lexical relations scoring, bag-of-dependencies (with and without part-of-speech tags) scoring. The resources that we utilized for linguistic preprocessing are the following: TreeTagger[4] is used for part-of-speech tagging, MaltParser[5]

[1] cf. https://github.com/hltfbk/EOP-1.2.3/wiki/MaxEntClassificationEDA

[2] https://dkpro.github.io/dkpro-core/

[3] https://uima.apache.org

[4] http://www.cis.uni-muenchen.de/∼schmid/tools/TreeTagger/

[5] https://dkpro.github.io/dkpro-core/releases/1.7.0/apidocs/index.html?de/tudarmstadt/ukp/dkpro/core/maltparser/MaltParser.html

for dependency parsing, for lexical semantic relations VerbOcean[6] (strongerThan, CanResultIn, Similar relations among verbs) and WordNet[7] (synonym, hypernym, holonym relations among nouns, verbs, adverbs and adjectives). Maximum entropy classification was parametrized in terms of maximum iteration number (default = 10000) and the cutoff threshold (default = 1).

We evaluated the algorithm's performance in terms of precision, recall, and F-score values for each of the three classes (whereby we report it only for the contradiction class), and the mean of each such value is weighted by support on each class, which we report as overall scores.[8]. In line with De Marneffe et al. (2008), we do not report the accuracy, because neither of the evaluated datasets is balanced across classes.

3.1.1 Baseline experiment

```
<pair id="41" entailment="UNKNOWN" task="IE"
    length="short" >
    <t>He said that "there is evidence that
        Cristiani was involved in the murder of the
        six Jesuit priests" which occurred on 16
        November in San Salvador.</t>
    <h>Cristiani killed six Jesuits.</h>
</pair>
<pair id="42" entailment="NO" task="IE" length=
    "short" >
    <t>"There were 20 survivors, the majority of
        them are in critical condition" said
        Cyprian Gatete, Rwanda's assistant
        commissioner of police, during a telephone
        conversation with Reuters.</t>
    <h>Cyprian Gatete is an employee of Reuters.</
        h>
</pair>
<pair id="43" entailment="YES" task="IE" length=
    "short" >
    <t>Parents also had to contribute "much more
        fully", while "coasting" schools would be
        tackled, Ms Kelly told MPs. But shadow
        education secretary David Cameron, who
        backed some ideas, said other parts were a
        "complete muddle".</t>
    <h>David Cameron works as the shadow education
        secretary.</h>
```

Figure 3.1: Annotated examples from the RTE-3 development dataset. Contradiction judgements are encoded by the *entailment="NO"* property.

To generate baseline scores, we have trained the maxent model on the RTE-3 development dataset[9] and tested it separately on the independent posts and the threads data. In this experiment we assess how the EOP classifier's performance generalizes over events,

[6] http://demo.patrickpantel.com/demos/verbocean/

[7] https://wordnet.princeton.edu

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

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

as well as over the domains of newswire data vs social media data, because the RTE-3 data on which we train maxent contains classical newswire RTE pairs.

To ease comparison, these results are going to be reported in each evaluation table in Section 3.1.2 in the *rte3* rows. We are going to see that the performance values in general show a large difference to the advantage of PHEME-internal training and testing (cf. Section 3.1.2), proving the utility of the two newly built PHEME RTE corpora for the PHEME project.

3.1.2 Cross-event validation

To assess the performance generalization over events for each of the PHEME datasets from the two data collection scenarios (independent posts and conversational threads), the labeled instances relating to our four distinct world events (chebdo, gwings, ottawa, ssiege) are going to be used in supervised learning in a leave-one-event-out setup. This is going to be a collection4-fold domain-internal cross-validation (CV) evaluation, where we always train on three of the event datasets and test on the held-out event dataset. The mean values of the 4-fold cross-validation are given in the bottom section of each EOP evaluation table.

We run these 4-fold CV experiments separately on the independent posts data and on the conversational threads data, which allows us to compare performances on PHEME collections originating from the same domain, i.e. same genre (social media texts) and even the same web platform, but different collecting scenarios (independent posts vs conversational threads).

4-fold CV on independent posts

As expected, we observe in the cross-event validation runs that the scores vary depending on which event is used for testing. The overall performance scores for the independent posts dataset are given in Table 3.1. The mean F score (.51%) is a large improvement over the baseline where training data from a different domain was used, proving the utility of the newly built PHEME RTE data.

The performance scores on the contradiction class for the independent posts dataset are given in Table 3.2. The 4-fold cross-validation average when using PHEME data shows a 4-points F-score improvement over the baseline, and yields a .25% F.

4-fold CV on conversational threads

Cross-event validation scores for the threads collection is given in Table 3.3. Again, the benefits of having this collection can be observed when comparing the values to the

test	train	P	R	F
chebdo	rte3	.6063	.3201	.2034
	iposts	.5734	.5907	.5639
gwings	rte3	.5972	.3742	.2731
	iposts	.6328	.6120	.6207
ottawa	rte3	.4557	.4718	.3908
	iposts	.4766	.3291	.3260
ssiege	rte3	.3294	.2792	.1780
	iposts	.5715	.5717	.5388
mean	rte3	.4971	.3613	.2613
	iposts	.5635	.5258	.5123

Table 3.1: Training data generalizability in terms of events represented in **independent posts**. Evaluation of the maxent model (EOP platform) on unseen events in the iPosts dataset. maxent was trained on general RTE data vs event-based held-out iPosts data (4-fold CV). Precision, recall and F measurements for each fold are **averaged over the 3 RTE labels** (Entailment, Contradiction, Unknown), weighted by support (the number of true instances for each label). The bottom lines express the mean values of the 4-fold CV.

test	train	P	R	F
chebdo	rte3	.2222	.1358	.1686
	iposts	.4750	1780	.2589
gwings	rte3	.2436	.2218	.2322
	iposts	.3375	.4202	.3743
ottawa	rte3	.3277	.1538	.2094
	iposts	.3170	.0689	.1132
ssiege	rte3	.3919	.1830	.2495
	iposts	.5758	.1798	.2740
mean	rte3	.2963	.1736	.2149
	iposts	.4263	.2117	.2551

Table 3.2: Independent posts evaluation, scoring the **contradiction class**.

test	train	P	R	F
chebdo	rte3	.7172	.3074	.3072
	threads	.6770	.5956	.6234
gwings	rte3	.6027	.2876	.2759
	threads	.6036	.5229	.5459
ottawa	rte3	.6652	.3000	.3137
	threads	.6051	.5829	.5872
ssiege	rte3	.5898	.2638	.2447
	threads	.6423	.6066	.6210
mean	rte3	.6437	.2897	.2853
	threads	.632	.577	.5943

Table 3.3: Threads collection evaluation – overall scores.

test	train	P	R	F
chebdo	rte3	.0301	.1429	.0498
	threads	.0303	.1143	.0479
gwings	rte3	0	0	0
	threads	0	0	0
ottawa	rte3	.0682	.1622	.0960
	threads	.0435	.0811	.0566
ssiege	rte3	.0488	.1000	.0656
	threads	.0481	.0833	.0610
mean	rte3	.0367	.1012	.0528
	threads	.0304	.0696	.0413

Table 3.4: Threads collection evaluation, scoring the **contradiction class**.

baseline classifier: the mean F score of the 4-fold cross-validation improves from .29% to .59%collection.

The performance scores on the contradiction class for the threads posts dataset are given in Table 3.4. The 4-fold cross-validation average scores extremely low on this dataset, and there is no improvement over the baseline when using PHEME data.

As a general trend, we observe that maxent retrained on PHEME data yields a large performance improvement over the baseline score. We get comparable performances on the independent posts and the threads data, both in terms of overall performance and of the contradiction class scores, except for the extremely low scores on the contradiction class in the threads data.

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 news vs social media genrecollection, as well as the generic RTE vs special-purpose RTE (i.e., PHEME) scenario.

A known issue with respect to the currently evaluated EOP maxent configuration is that the sophisticated linguistic analysis based on which maxent generates features typically fails to be extracted from the social media data encoded in the PHEME datasets, since tweets are notoriously hard to parse for grammatical properties (e.g. part-of-speech and syntax information).

Another caveat of using EOP maxent is that the snippets in a pair in RTE-3 feature markedly different lengths, whereby the *text* is longer than the *hypothesis* (see Figure 3.1). Such directionality is anticipated by the EOP maxent classifier, while it is insignificant and thus cannot be put to use in PHEME data, where typically both snippets are of comparable length due to the 160 character length superimposed by the Twitter platform.

The retained maxent algorithm scores best on getting all three relations right in threaded conversations (.59 F), whereas learning the contradiction class is performed better on independent posts (.26 F).

We argue that collecting and utilizing two different types on contradiction data originating from the same social media platform helped gain new insights into the nature of contradictions, and supported the evaluation of the two different contradiction types in the new PHEME collections, and that it was crucial to identify and generate both datasets to increase classification robustness.

3.1.3 Cross-collection validation

Another experiment we conducted aimed to assess the impact of the differences between the classification models learned on the threads vs the independent posts collections. We trained maxent on the entire independent posts collection and tested on the entire threads collection and vice versa. We observe that cross-collection validation has a negative outcome, i.e. maxent performs significantly worse when cross-collection-trained than when collection-internally trained (see: 4-fold CV scores in Section 3.1.2).

On independent posts, the F score is so badly affected that it is falls under the baseline. This means that the RTE examples in the threads collection are unable to contribute a good RTE model for entailment and contradiction occurring in independent social media posts. For the threads collection however, the overall cross-training score is still a large improvement over the baseline, and performance on the contradiction class is able to improve with respect to both the baseline and the collection-internal classification (i.e., event-based 4-fold CV), nonetheless still staying poor. We conclude that cross-scenario validation indicates that the entailment and contradiction model learned by maxent from the independent posts collection is able to contribute to classifying entailment and contradiction occurring in threads, but not vice versa. The relatively small size of the threads collection may bias the scores, which needs to be explored via scaling up of the experi-

test	**train**	scoring	**P**	**R**	**F**
threads	iposts	overall	.6266	.4246	.4845
		CON	.0674	.2727	.1081
iposts	threads	overall	.5173	.3786	.2140
		CON	.2105	.0029	.0058

Table 3.5: **Cross-collection validation** scores by maxent (ME). Weighted means for all 3 classes, and scoring the contradiction class.

ments.

3.1.4 Cross-domain validation

De Marneffe et al. (2008) report to achieve .2295 precision and .1944 recall on the contradiction class in the RTE-3 test dataset, when training their complex contradiction detection system on an aggregate of 5 RTE datasets. One of those is the RTE-3 development set that we used for training the baseline maxent model. This baseline model achieves .0714 precision and .1111 recall values on the same RTE-3 test set that De Marneffe et al. (2008) report. When we train on the newly created PHEME independent posts collection that has proved its generalizability in Section 3.1.3, we obtain on this test set .1022 precision and .1944 recall, where both are an improvement over the baseline, whereas precision is half of what De Marneffe et al. (2008) achieve, and recall is the exact same.

Our best performance on the contradiction class using maxent was obtained on one of the newly created PHEME social media collections (the independent posts collection) in event-based 4-fold CV: .2963 precision and .1736 recall. EOP maxent utilizes roughly the same level of linguistic complexity as De Marneffe et al. (2008). maxent by itself encodes less world knowledge (e.g. in terms of event coreference information) than the system of De Marneffe et al. (2008), but it might encode such knowledge via the way the event-specific collections were created. We deem our results encouraging and conjecture that our performances are at a comparable level. In follow-up experiments we are planning to study non-default configurations for maxent in terms of linguistic preprocessing and parameter settings.

3.2 Experiments with word embeddings

Experiments documented in the earlier sections of this chapter are based on lexicalised models implemented as part of the EOP framework. As discussed in Section 1.3, recent research in the area of recognising textual entailment has focussed on models which, unlike the EOP models available as part of the framework, do not rely on hand-crafted

external resources such as WordNet (**?**) which are then used for feature extraction. Instead, they are based on learning embeddings which are used as input for classifiers. We therefore report additional experiments based on models which utilise word embeddings.

Our models are an extension to the "sum of word embeddings" word embedding model reported as a baseline in Bowman et al. (2015), where we modify the approach used for feature generation. The final system in this work did not greatly outperform the baseline.

In our system, we use word2vec models (**?**), which are trained in an unsupervised way on a large collection of text. We discuss the training of these models in detail in Sections 3.2.2 and 3.2.2.

3.2.1 Baseline system

For comparison within this part of the work we reimplemented most of the De Marneffe contradiction detection system, as described in De Marneffe et al. (2008). This system is designed for operation on newswire, and consists of a number of complex features for classification. Features are combined using manually set weights. Significantly, the system uses typed dependency parsing for many of its features. Although there are dependency parsers for tweets, to our knowledge, no pre-existing dependency parsers produce dependency type labels identical to those produced by Stanford's typed dependency parser. As such, the De Marneffe system would require considerable adaptation to operate correctly on tweets, as in the PHEME RTE data.

3.2.2 Experimental setups

Two experimental setups were utilised in this work. Firstly, we carried out experiments in entailment detection using the general RTE data set described in De Marneffe et al. (2008). This data allowed us to evaluate both a number of systems and determine which to evaluate in our second experimental setup, using the PHEME test datasets.

Non-social-media setup

Our initial experiments were carried out on the Stanford contradictions corpus. The purpose of these experiments was to compare the performance of a contradiction detection method based on machine learning and word embeddings with that of our existing reimplementation of the De Marneffe et al. (2008) contradiction detection system. This comparison allows us to decide whether to undertake the extensive task of adapting the De Marneffe et al. (2008) approach to work on the tweets of our Twitter RTE data. If our learned model, which is far simpler to train or retrain, is shown to perform favourably

compared to the De Marneffe et al. (2008) approach on the Stanford contradictions corpus, we will adapt that to tweets instead.

Since the domain of the Stanford contradictions corpus is web-based news, we were able to use the pre-trained word2vec model trained on the Google News corpus[10]. Tokenisation is carried out using the Stanford tokeniser from CoreNLP (**?**). Our own contradiction classification model is trained on the development portion of the Stanford contradictions corpus, using all three years' worth of RTE data.

Social media setup

Tweets are preprocessed as follows. Twitter-based tokenisation is performed with twokenize[11]. Afterwards, tokens are normalised to lower case and stopwords are filtered, using the nltk[12] English stopword list, punctuation characters, plus Twitter-specific stopwords. The latter is manually created and consists of: "rt", "thats", "im", "'s", "...", "via", "http". The first six have to be an exact token match, the last one has to match the beginning of a token. Finally, phrases are detected, using an unsupervised method that creates 2-grams of commonly occurring multi-word expressions (**?**)[13]. At application time, if two subsequent tokens are identified as a phrase, those tokens are merged to one token.

For the social media setup, we train our own word2vec model based on unlabelled tweets from the Ferguson riots. We also experiment with the same word2vec model based on Google news as used in the first experimental setup. Our own word embeddings are trained as a skip-gram word2vec model (dimensionality 300, 5 min words, context window of 5). The contradiction classification model for this part of the work is trained separately for each event using the labelled data for all the other events (event-based 4-fold cross-validation).

3.2.3 Classification model

An overview of the model architecture can be found in Figure 3.2. Classification is carried out over features based on vectors which are derived automatically using the unsupervised word2vec method.

Prior to carrying out this process, we must obtain a suitable word2vec model. These models map from terms onto vectors which represent their meaning semantically using word embeddings. The vectors are intended to be compared mathematically - for example words that have vectors that are close to one another are thought to be strongly related. Word2vec models can be trained automatically using large amounts of unlabelled data,

[10] https://drive.google.com/file/d/0B7XkCwpI5KDYNlNUTTlSS21pQmM
[11] https://github.com/leondz/twokenize
[12] http://www.nltk.org/
[13] https://radimrehurek.com/gensim/models/phrases.html

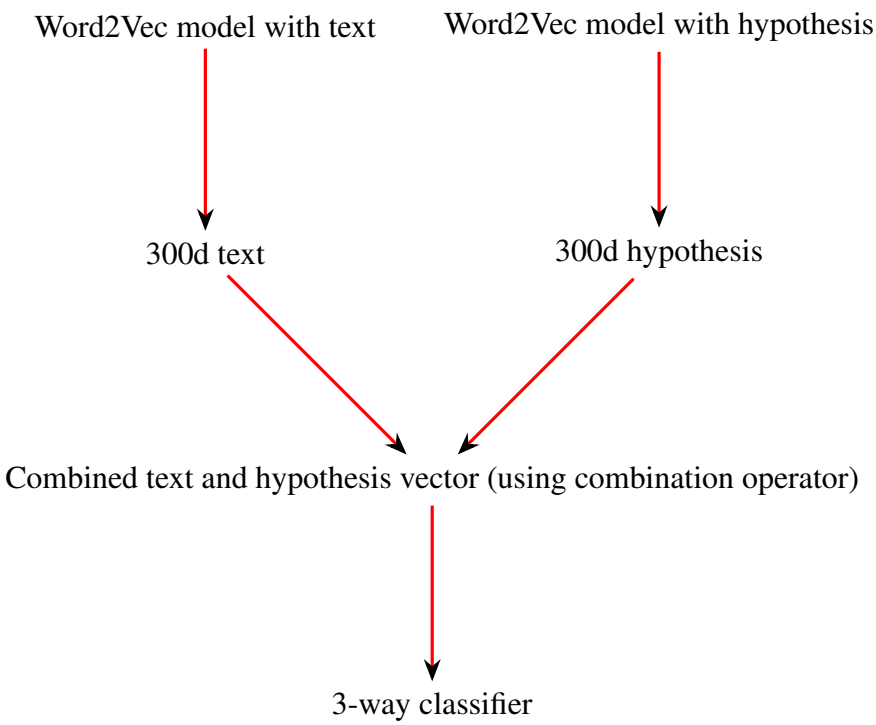

Figure 3.2: Model for contradiction classification using word2vec embeddings

though pre-trained models are already distributed online. In this work we use the pre-trained Google News word2vec models[14] in our work, along with a model trained on tweets from the Ferguson riots.

The vectors produced by word2vec are fixed length (300 in our experiments), with one vector per word. Starting with a hypothesis or target sentence, we convert its words into a series of vectors. Since we are using a simple classifier rather than a sequence model, we combine the vectors for each word by summing them, giving one vector of size 300 to represent the entire sentence.

Since the RTE pairs consist of a target and hypothesis (as shown in Figure 3.1), after transforming them into vectors, we have to combine the two representations before feeding them into a classifier. We experiment with two different strategies for that: concatenating the text and hypothesis representations or taking the outer product of the text and hypothesis vectors. The latter gives a much larger feature space than the former (90,000 vs 600), which is a disadvantage for learning on small data sets, but allows more complex relationships between hypothesis and text to be learned. We do not evaluate the outer product features with the Stanford contradictions corpus as there is not enough training data to learn from so many features.

After generating the features for the contradiction pair, we train a Bag-of-Word-

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

Test	System	P	R	F1
RTE-1	De Marneffe	.3111	.0940	0.1443
	Our System	.2450	.2483	0.2467
RTE-2	De Marneffe	.1818	.0577	.0876
	Our System	.2132	.2788	.2417
RTE-3	De Marneffe	.4062	.1806	.2500
	Our System	.1545	.2361	.1868

Table 3.6: Quality of bag-of-word-embeddings system in comparison to De Marneffe et al. (2008) contradiction detection. Results are for the contradiction class only.

Vectors system (BOWV), in which the representations are fed into a 3-way logistic regression classifier with L2 regularisation (**?**). Class weights are determined based on their prior distributions in the training set. The decision of this classifier is the final result for our system.

3.2.4 Results

We report the same metrics as for the EOP results, giving only the scores for the contradiction class. The results for our non-Twitter experimental setup, comparing our system with the reimplemented De Marneffe et al. (2008) approach, are shown in Table 3.6. Our system outperforms De Marneffe on F-score for all but the third RTE dataset, on which it is beaten on precision.

Given that our system is shown to be comparable to De Marneffe in the worst case and stronger than it in the best case, we further develop this word-embeddings based system, retraining and evaluating it on the Pheme contradictions data.

The results for our second experimental setup, based on the Pheme corpus of contradictions in tweets, are shown in Table 3.7. We evaluated the system with existing word-embeddings provided by Google and trained on a corpus of news, as well as our own word embeddings trained on unlabelled data from the Ferguson riots. In addition, we also evaluated for each dataset the performance when taking the outer product of all features for both the target tweet and the hypothesis tweet (giving a rich but very large feature space) and when simply adding the two average word embeddings for the tweets.

Unfortunately, no variant of the system consistently outperformed the others. Although the system with in-domain word embeddings and added features gave the strongest performance for the *ssiege* and *ottawa* data, it was beaten when using the Google News embeddings on both the *chebdo* and *gwings* data.

Test	Embeddings	Features	P	R	F1
chebdo	Google News	Outer	0.281976744	**0.227166276**	**0.251621271**
	Google News	Add	**0.282828283**	0.131147541	0.1792
	Ferguson	Outer	0.208860759	0.077283372	0.112820513
	Ferguson	Add	0.221621622	0.096018735	0.133986928
gwings	Google News	Outer	**0.222222222**	0.015564202	0.029090909
	Google News	Add	0.2000	0.023346304	**0.041811847**
	Ferguson	Outer	0.090909091	0.003891051	0.007462687
	Ferguson	Add	0.171428571	**0.023346304**	0.04109589
ottawa	Google News	Outer	0.32	0.021220159	0.039800995
	Google News	Add	0.25	0.01061008	0.020356234
	Ferguson	Outer	0.384615385	0.013262599	0.025641026
	Ferguson	Add	**0.4**	**0.021220159**	**0.040302267**
ssiege	Google News	Outer	0.258169935	0.221288515	0.238310709
	Google News	Add	0.257627119	0.212885154	0.233128834
	Ferguson	Outer	0.25	0.210084034	0.228310502
	Ferguson	Add	**0.286729858**	**0.338935574**	**0.310654685**
mean	Google News	Outer	**0.270592225**	**0.121309788**	**0.139705971**
	Google News	Add	0.247613851	0.09449727	0.118624229
	Ferguson	Outer	0.233596309	0.076130264	0.093558682
	Ferguson	Add	0.269945013	0.119880193	0.131509943

Table 3.7: Quality of bag-of-word-embeddings system evaluated on Pheme data. Results are for the contradiction class only. Models shown are trained using Google news word embeddings, and word embeddings trained on unannotated tweets related to the Ferguson riots

Chapter 4

Discussion and Conclusion

The best scores on classifying contradictory text pairs achieve comparable F-scores across the datasets and the systems that we have investigated:

- our reimplementation of de Marneffe's system on RTE-3: .2500 F
- our bag-of-word-embeddings system on RTE-1: .2467 F
- our bag-of-word-embeddings system on RTE-2: .2417 F
- our retraining of maxent in EOP on the newly built PHEME RTE independent posts collection: .2551 F.

We conclude that our results on contradiction detection in social media data using the maximum entropy algorithm in EOP are state of the art, and the retrained maxent model can be suggested to be integrated in the PHEME pipeline. Since EOP is designed to be a general RTE component compatible with NLP pipeline, for software documentation we refer to the Excitement Open Platform developer repository[1]. For obtaining the newly built PHEME RTE independent posts collection, we refer to the PHEME project's software download site[2].

Given the low scores we observed in cross-domain experiments, we argue that the new, semi-automatically created PHEME RTE dataset is targeting the contradiction task better and possibly yielded more varied lexicalisations of contradictory text pairs occurring in real-life social media data than hand-seeded Hearst-like patterns that we foresaw in the project's DoW. The employment of use case specific controversies as seeds is going to be explored in follow-up experiments nonetheless, in which cross-media features (e.g. links to other sources, trustworthiness and authority) are to be investigated as well.

[1] https://github.com/hltfbk/EOP-1.2.3/wiki/MaxEntClassificationEDA
[2] http://www.pheme.eu/software-downloads/

Bibliography

Augenstein, I., Vlachos, A., and Bontcheva, K. (2016). USFD: Any-Target Stance Detection on Twitter with Autoencoders. In Mohammad, S. M., Kiritchenko, S., Sobhani, P., Zhu, X., and Cherry, C., 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. (2006a). The second pascal recognising textual entailment challenge. In *Proceedings of the second PASCAL challenges workshop on recognising textual entailment*, volume 6, pages 6–4.

Bar-Haim, R., Dagan, I., Dolan, B., Ferro, L., Giampiccolo, D., Magnini, B., and Szpektor, I. (2006b). 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 Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Rosner, M., and Tapias, D., 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. Association for Computational Linguistics.

Dagan, I., Glickman, O., and Magnini, B. (2006). *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment: First PASCAL Machine Learning Challenges Workshop, MLCW 2005, Southampton, UK, April 11-13, 2005, Revised Selected Papers*, chapter The PASCAL Recognising Textual Entailment Challenge, pages 177–190. Springer Berlin Heidelberg, Berlin, Heidelberg.

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

Fellbaum, C. (1998). *WordNet – An Electronic Lexical Database*. MIT Press.

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

Giampiccolo, D., Magnini, B., Dagan, I., and Dolan, B. (2007). The third pascal recognizing textual entailment challenge. In *In Proceedings of the ACLPASCAL Workshop on Textual Entailment and*.

Lendvai, P., Augenstein, I., Bontcheva, K., and Declerck, T. (2016). Monolingual social media datasets for detecting contradiction and entailment. In *LREC*.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., and McClosky, D. (2014). The Stanford CoreNLP Natural Language Processing Toolkit. In Bontcheva, K. and Jingbo, Z., editors, *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, Baltimore, Maryland. Association for Computational Linguistics.

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.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013a). Distributed Representations of Words and Phrases and their Compositionality. In Burges, C., Bottou, L., Welling, M., Ghahramani, Z., and Weinberger, K., editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013b). Distributed Representations of Words and Phrases and their Compositionality. In Burges, C., Bottou, L., Welling, M., Ghahramani, Z., and Weinberger, K., editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc.

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.

Padó, S., Noh, T.-G., Stern, A., Wang, R., and Zanoli, R. (2015). Design and Realization of a Modular Architecture for Textual Entailment. 21(02):167–200.

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

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

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

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