FP7-ICT Collaborative Project PHEME (No. 611233)

Computing Veracity Across Media, Languages, and Social Networks

D2.3 Spatio-Temporal Algorithms

Leon Derczynski, University of Sheffield
Kalina Bontcheva, University of Sheffield

Abstract.
FP7-ICT Collaborative Project ICT-2013-611233 PHEME
Deliverable D2.3 (WP2)

The deliverable describes the results of Task 2.4 in WP2 on extracting spatio-temporal entities from Twitter streams. Following the description of work, spatial and temporal entities are extracted from streams of social media text across multiple languages (English, German and Bulgarian). In addition, cross-lingual resources are provided for future development. We provide methods for extracting spatial information both from within messages (i.e. location mentions) and also at message level (i.e. geolocation). The research from Task 2.4 enables spatio-temporal constraint of claims, which is critical to reliable fact-checking in PHEME.

Keyword list: Twitter, event extraction, temporal annotation, spatial annotation, annotation projection, unsupervised feature extraction, geolocation

Project	PHEME No. 611233
Delivery Date	13 July 2015
Contractual Date	30 June 2015
Nature	Prototype/Report
Reviewed By	Arkaitz Zubiaga, University of Warwick; Maria Liakata, University of Warwick
Web links	http://www.pheme.eu
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

The research in this deliverable is focused on identifying and extracting spatial and temporal entity expressions from social media text. This has not been attempted before, with the majority of temporal information extraction and spatial role labelling tasks being over newswire. As demonstrated in our earlier research, reported in D2.2., analysing social media automatically is a particularly challenging task, due to the noisy, irregular, high volume, and strongly contextual nature of the genre.

This deliverable addresses several research questions. Firstly, we examine how the social media text type impacts spatio-temporal discourse annotation. Secondly, we address methods for the effective semantic annotation of spatial and temporal entities in social media. Thirdly, the challenge of processing new under-resourced languages is addressed through projection.

The novel contributions presented here are: *(i)* minimal spatio-temporal entity annotation guidelines for social media; *(ii)* a Twitter dataset annotated for spatio-temporal entities; *(iii)* automated methods for cross-genre spatio-temporal entity extraction; *(iv)* a method for grounding of locations in documents to linked data through Nomenclature of Territorial Units for Statistics (NUTS) subdivisions; and, finally, *(v)* a method for cross-lingual spatio-temporal entity projection, to enable low-overhead adaptation of our methods to less resourced languages.

The results of this work comprise:

- A tagger for spatio-temporal entities in social media text;
- A method for grounding locations in NUTS regions;
- Corpora annotated with spatial and temporal entities, over social media text in three languages (English, German, Bulgarian);
- Guideline refinements for ISO semantic annotation when applied to microblog social media text.

Contents

Chapter 1

Introduction

Recently people have started using social media not only to keep in touch with family and friends, but also increasingly as a news source. However, knowledge gathered from online sources and social media comes with a major caveat – it cannot always be trusted. Rumours, in particular, tend to spread rapidly through social networks, especially in circumstances where their veracity is hard to establish. For instance, during an earthquake in Chile rumours spread through Twitter that a volcano has become active and there was a tsunami warning in Valparaiso (Marcelo et al., 2010). Researchers have found that people read untrusted sources for various reasons, the main ones being their interestingness, entertainment value, a friend's online recommendation, or a search engine result (Ennals et al., 2010).

Each rumour centres around events, actors and context. These are all vitally important to the definition of the rumour. Context can be thought of the place and time in which a rumour takes place; e.g. for the rumours surrounding Vladimir Putin's disappearance, there was a strict temporal context – March 2015 – and a soft spatial context: Russia. Therefore, spatio-temporal information provides critical information to understanding the rumour.

In addition, social media craves grounding context. Messages are short, and authors tend to assume that they will be read soon by someone in their social network (whether defined by explicit friendship / follower relations, or through homophily). These factors create implicit context, meaning that messages can be short and vague, in the knowledge that the reader will likely have enough common context to understand the message at the time that they read it. This leads to incomplete information in the message. As a result, actors involved in rumours are often referred to in vague terms, and need to be grounded before they can be automatically processed.

This deliverable describes methods for performing spatio-temporal information extraction and grounding in social media text. To account for the gamut of social media text types, we begin applying newswire extraction techniques over Twitter text, which as a platform is both a tough NLP genre Derczynski et al. (2013a) and also the model organism of social media Tufekci (2014). We include tweets as a counter to newswire because they are much tougher to process than other sources, and rich in colloquialisms. Forums and blog posts are in nature easier as they are closer to newswire (Baldwin et al., 2013), and so readily accessible by tools that excel at both news and tweets. Specifically, we examine the use of spatio-temporal language in four PHEME rumours. In particular, we want to extract spatial and temporal entity mentions from user-generated content; we want to spatially ground tweets in linked data, for processing by other components in the project; and we want to develop resources for spatial and temporal information extraction in the project languages outside English – Bulgarian and German.

Spatial and temporal grounding are key for the following research in visualising spatial and temporal properties of rumours (D5.3), in integrating spatio-temporal knowledge within the PHEME framework (WP7), and in presenting longitudinal models of events and users.

To capitalise on prior research and the resources already gathered in PHEME, we take unconventional approaches to spatio-temporal information extraction and projection. Instead of the traditional supervised model of building a dataset, annotating it, extracting features and learning a classifier, we use the Twitter-specific tools from D2.2 to process our large corpora and then proceed with unsupervised feature extraction over this data. To manage the costly transition in linguistic processing from newswire to tweets, having already developed tools, we use both these for English but also blend news with non-news data during our unsupervised feature extraction (with an advanced version of Brown clustering). This results in knowledge that crosses both text types and is effective at many entity recognition tasks in general, and thus can be readily applied outside of PHEME. In addition, the unsupervised feature extraction can now also be applied to new languages given the tools developed in D2.2, easing the otherwise large challenges of crossing not only text type from news to user-generated context but also the crossing from English to new languages, German and Bulgarian.

Our approach results in automatic tools for spatio-temporal information extraction over tweets and resources in new languages.

We detail our spatio-temporal entity extraction in Chapter 2. The performance is acceptable, being a bit below state-of-the-art newswire performance (e.g. to 69% F1 for event recognition in tweets with our system, from 72% F1 for events in newswire as the best score in the multiple-participant TempEval-3 shared task (UzZaman et al., 2013)). A drop in performance is expected given the usual step down in moving from newswire to Twitter processing. In fact, as state-of-the-art tools like the Stanford Named Entity Recognition kit drop from 89% to 41% F1 on tweets (Derczynski et al., 2013a) – a 48% drop in absolute performance – we consider our results in this domain strong, and exceptionally efficient, as well as easy to expand and transfer to new domains and languages.

Chapter 3 details our work in linking spatial entities found in tweets to linked data. We detail the integration of a state-of-the-art entity linking system with geographic data sources, leading to three sources of location grounding. This includes an analysis of the different levels of granularity available for grounding social media posts. In general, we found that more specific items became harder and harder to pinpoint, but could be done so with greater accuracy. Given that notable rumours are more likely to be spatially contextualised at the level of city or above, the work is suitable for PHEME.

In Chapter 4, we detail our method for projecting annotations from existing gold-standard resources to new languages. Specifically, we map the ACE SpatialML annotations and TimeBank 1.2 from English into German and Bulgarian. These resources have been developed from mature and now ISO-grade annotation schemata; in addition, they have been around for a while and in the case of TimeBank received and integrated multiple iterations of considerable constructive community feedback, e.g. (Boguraev and Ando, 2006; Boguraev et al., 2007). This led to very solid resources. The annotation projection technique is selected based on the property that both annotation standards aim for minimal-length annotations of the expressed concepts. Finally, we use a crowdsourcing step to screen out bad projected annotations.

1.1 Relevance to PHEME

The PHEME project aims to detect and study the emergence and propagation of rumours in social media, which manifest as dubious or false claims. In order to determine the context and precise meaning of a claim, we *must* know its spatio-temporal context.

1.1.1 Relevance to project objectives

Visualising rumours over space and time is a key goal of PHEME, and so we require information that provides that spatial and temporal information. In addition, grounding rumours in linked data is critical to formally reasoning about them and fact-checking, and so this connection of events and places to the semantic web is vital. This is also important for the use cases, as it enable users to inspect, analyse, and refine information needs around emerging phemes.

1.1.2 Relation to forthcoming and prior research in Pheme

This is the penultimate deliverable in WP2, Ontologies, Multilinguality, and Spatio-Temporal Grounding. It builds on the linguistic pre-processing and multilinguality tools in D2.2, making use of these to develop resources useful for general linguistic processing as well as specifically for spatio-temporal and for Twitter processing. For example, the tokenisers developed for English, German and Bulgarian are crucial to effective Brown cluster generation; without them, we would not have come across the novel, cross-language, unsupervised approach for building spatio-temporal entity extraction tools presented here. This is likely to be of great utility to many other researchers.

The spatio-temporal information extracted here enriches the social science research detailed in T2.1. We can now add explicit spatial and temporal context to the observations made, leading to a rich and empirical analysis of rumour sources and diffusers in D2.4. Additionally, as we have chosen Twitter, the hardest social media venue to automatically process, for our development, our tools should be comfortably robust enough to handle less challenging sources such as web forums, e.g. Patients Like Me from the biomedical use-case.

If we have NLP performance problems here, we have in T2.4 ended up developing algorithms that not only satisfy the description of work but also are flexible in text type through their unsupervised feature extraction. The upshot of this is that re-adapting to another specific domain requires only text data, and no manual labelling. As we already have a large body of text data for each venue and even each forum from T7.2 and T8.2, we are now in a perfect position to adapt to these texts trivially if needed, using the tools in D2.2 and the techniques presented here in Chapter 2.

1.1.3 Relation to other work packages

The techniques, software and resources presented in this deliverable use the pre-processing developed in WP2. They can be readily applied in the use-cases, WP7 and WP8. In addition, they help contextualise the event clusters generated in WP3, and can be readily linked with the PHEME veracity framework (WP6). Finally, the PHEME visual dashboard is informed by the spatio-temporal information used to enrich tweets in this work package and shared via integration, aiding presentation of spatial and longitudinal information (WP5).

Chapter 2

Spatio-Temporal Entity Extraction

The goal of this work is to identify and extract spatial and temporal entity expressions from social media text. This has not been attempted before, with the majority of temporal information extraction and spatial role labelling tasks being over newswire. However, PHEME focuses on social media and other less-curated forms of online language, where rumours are present – in contrast to the structured newswire articles previous technology has worked on. We are by now aware of the general challenges present in social media text: increased lexical diversity; orthographic deviation (both intentional and by mistake); terseness; unstable capitalisation; a lack of annotated datasets; and a lack of context.

The text type is intrinsically important to address. While we understand newswire to a certain degree, it is highly constrained, having biases not only in terms of its canonicality, but also stemming from explicit rules (e.g. editorial guidelines) and implict sources (socio-economic bias) (Eisenstein, 2013). Further, text being generated on social media is of huge volume, because it is essentially a (biased) sampling of all human discourse. The insights and analyses possible over such a rich and large resource are only just starting to be realised (Derczynski et al., 2013b). As a result, just as space and time are critical parts of context in natural language, so is it essential to understand the expressions times and places in social media text.

Both spatial and temporal semantic annotation can generally be divided into two parts: annotation of entities, and annotation of the relations that obtain between them. These parts are each deeply challenging. As the documents in the most plentiful (and one of the toughest) sources of social media – Twitter – are short, relations are likely to be inter-document and require entity coreference to address, which adds an intermediate stage to the process. In this section, we describe approaches to the fundamental first part: entity annotation.

The research questions addressed are:

- How does the social media text type impact spatio-temporal discourse annotation;
- How can one effectively semantically annotate spatial and temporal entities in social

media text?

The novel contributions presented here are: minimal spatio-temporal entity annotation guidelines for social media; a Twitter dataset annotated for spatio-temporal entities; and automated methods for cross-genre spatio-temporal entity extraction. This leads to delivery of the following artifacts:

- A tagger for spatio-temporal entities in social media text;
- Corpora annotated with spatial and temporal entities, over social media text;
- Guideline refinements for ISO semantic annotation when applied to tweets.

2.1 Annotation

The first goal is to concretely define what will be extracted. It is prudent to be careful in this regard and take the target text type into account, instead of picking up an entire existing annotation standard and applying it in a new challenge. The standards we start from are ISO-TimeML (Pustejovsky et al., 2010) and ISO-Space (Pustejovsky et al., 2011), well recognised community standards for temporal and spatial annotation respectively. These provide full-featured schemas for both entities and relations, though as mentioned, our focus is on entity annotation only. The parts that we choose for entity annotation are described below.

There is a lack of certainty about what a "named entity" is. Indeed, it is generally application dependent; cf. Kripke (Kripke, 1972). This provides strong support for having adapted our own guidelines to spatio-temporal annotation over the social media text type.

Following recent research on customising annotation (Schneider, 2015), we reduce the many diverse types of entity supported by these ISO standards down to the set that is both applicable to social media text and also fits within the scope of our task. Annotating the full standards would be very painful and superfluous to needs; however, taking a strict subset of the standards is much cheaper than rebuilding a spatio-temporal annotation and encoding standard, especially given the large volume of guidelines, edge case guidance and other supplementary material that has accumulated for ISO-TimeML and ISO-SpatialML. This means that, for example, we will ignore all temporal relation information and data about spatial paths and other relations. While generally critical to the understanding of text, these relations are not immediately necessary to the spatial grounding of concepts and entities that is required in PHEME.

In addition, we do not attempt to address the annotation of signals in this work, as they are primarily intended as intermediaries for relation annotation. However, we do exploit their nature of being frequently bisemous with spatial and temporal senses in Section 2.2.2.

This leaves temporal and spatial entity annotations only. The specific choices made are discussed in the two subsections below.

As a general point, we approach this semantic annotation over social media text just as we have done previous in linguistic annotation work. We do not use a customised annotation process, or expect lower quality of annotations due to the inherently informal nature of social media text. It is our prior experience that annotators can engage equally well with newswire and social media text; this is supported by recent pilot studies (Plank et al., 2015).

2.1.1 Temporal entities

ISO-TimeML describes two entity types: temporal expressions (*timexes*) and events. Briefly, temporal expressions are explicit mentions of periods "*for three days*" or times "*next April*", and a sophisticated scheme for representing them is provided; events are single words describing events or eventualities, of many kinds. Following the standard, we include TIMEX3-style timexes, as they are terser than the TIMEX2 variety, which capably generates single annotations longer than a whole tweet. Event annotation is more nuanced. Events can be of many types.

For temporal expressions, we annotated as per the standard. For events, we removed some kinds of event that can be confusing to annotators (in general) or are more useful for relation annotation than the entity annotation goal. This means that the following classes of event are included (descriptions from the TimeML annotation guidelines):

- **Reporting** – Reporting events describe the action of a person or an organization declaring something, narrating an event, informing about an event, etc.
- **Perception** – This class includes events involving the physical perception of another event. Such events are typically expressed by verbs.
- **Aspectual** – A grammatical device of aspectual predication, which focuses on different facets of event history.
- **I_Action** – An Intensional Action. An I ACTION introduces an event argument (which must be in the text explicitly) describing an action or situation from which we can infer something given its relation with the I ACTION.
- **Occurrence** – This class includes all the many other kinds of events describing something that happens or occurs in the world.

And these are excluded:

- **State** – States describe circumstances in which something obtains or holds true, with specific caveats; a broad notion that goes beyond just eventualities and is notoriously hard for annotators to learn, as well as often difficult to ground in the manner required by PHEME (which generally concerns events which are spatially groundable or bounded).
- **I_State** – Similar to I_Actions, this class includes states that refer to alternative or possible worlds.

In the case of both event and temporal expression annotations, we did not annotate types or other values, sticking just to entity boundary recognition. Once identified, the annotation of entity attributes is well-understood, with a large supply of tools performing just this task (e.g. for timexes, (Llorens et al., 2012; Bethard, 2013)).

2.1.2 Spatial entities

ISO-Space describes two kinds of entity: location and spatial_entity. The boundary between these two is not precisely defined, but in general, locations tend to be the union of geo-political entities and of geographic features that have a non-human origin. Conversely, the spatial_entity has the broader definition of “anything participating in a spatial relation”. Under this, something like “My spoon is in the bowl” would have two spatial_entity annotations: *spoon* and *bowl*. The general assumption here is that the location is of coarser spatial granularity than spatial_entity. While something like this is useful in e.g. recipe tasks (Regneri et al., 2010; Kusmierczyk et al., 2015), it does not have a place in our scenario.

Gaizauskas et al. discuss the boundaries between location and spatial_entity (Gaizauskas et al., 2012), finding it to be somewhat subjective and not consistent either at different spatial scales or over time. For example, geo-political entities (GPEs) like islands are traditionally locations, whereas restaurants are deemed somewhat more transient and therefore a spatial_entity; however, there are volcanic islands whose locations shift and even that appear and disappear in timespans shorter than some restaurants’ existence. In this case, we should be careful to pick something that fits the social media task well, especially considering that we will not annotate spatial relations at all. Therefore, we opt for a constrained interpretation of spatial_entity.

In this instance, we include the traditional notion of location, and the ACE concept of GPE (ACE, 2004). In addition, we included the ACE/Freebase concept of “facility” (Bollacker et al., 2008), which is often used in social media entity recognition exercises (Ritter et al., 2011; Baldwin et al., 2015). We exclude some parts of the spatial_entity definition, giving objects a rough minimum size of about $10m^2$. This is to avoid including items not useful for contextualising rumours, e.g. the bowl and spoon from the example above. Finally, ISO-Space includes an event entity type; also it is implemented differently to that in ISO-TimeML, we override this and use just the ISO-TimeML event definition (refined as above).

2.2 Approach

The datasets, features representations and classifiers used are detailed in this section. Our general approach is supervised, using in-type and out-of-type training data, represented with unsupervised features and features specific to the problem. We cast the problem as sequence labelling, where one attempts to predict labels one after each other in a defined sequence. In this case, the sequence is seen as the words (tokens) in each sentence, and the labels show whether the token is an entity or not. This approach recognises the structure inherent in using sentence-level units and in the order of words within each sentence. Labels are in two groups, learned by two different classifiers. For the temporal entity extraction task, the labels are [TIMEX, EVENT, O]; for the spatial entity extraction task, the labels are [LOCATION, SPATIAL_ENTITY, O]. In both cases, O corresponds to "outside", meaning that a token is outside of any entity, i.e., a non-entity word.

2.2.1 Datasets

We use four distinct datasets in this exercise. Firstly, the data which is manually annotated is drawn from the rumours gathered in deliverable D8.2, as detailed in Section 2.2.6. Secondly, there are two pre-annotated datasets used to support this: the W-NUT/Ritter NE annotations (Ritter et al., 2011; Baldwin et al., 2015) for spatial, and the TempEval-2 data for temporal (Verhagen et al., 2010). Finally, we perform unsupervised feature extraction through Brown clustering using a sample of tweets from Twitters 10% feed, which is a fair sample (Kergl et al., 2014), drawn between 2009 and 2015 to induce resilience against entity drift (Masud et al., 2011; Fromreide et al., 2014) – the garden hose archive (GHA).

2.2.2 Spatio-temporal bisemy

Certain words that have a spatial or temporal sense are well-known to have a spatio-temporal bisemy: that is, they occur in either a spatial or temporal sense. This has been useful in prior temporal annotation work, where function tags specifying spatial function were strong negative indicators of a temporal sense. However, there is insufficient context in tweets to attempt a classical word sense disambiguation approach, and scant manually-annotated data for this purpose. To this end, we attempt to identify spatial and temporal signal words and flag them in tweets, using these flags as features.

Specifically, we add a feature for each signal word encountered, describing whether it is suspected to occur in a spatial or temporal sense. This is based on existing lists of temporal signals extracted from prior work (Derczynski and Gaizauskas, 2011) or the

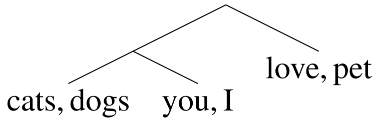


Figure 2.1: A binary, hierarchical clustering of semantically similar entries. Each leaf corresponds to a cluster of words (i.e., a "class").

ground truth; the sense distribution of each word in newswire; and the immediate context of the word. It is represented as a feature only present if a word has a temporal signal sense, weighted to the prior probability of an instance of that word actually having that sense, from the TB-sig TimeBank corpus variant. For example, 69.4% of occurrences of the word "until" had a temporal sense in this corpus, so when this word is found, a feature *tempsignal=0.694* is added. Conversely, while the word "into" sometimes has a temporal sense, this does not make for the majority of its occurrences; and accordingly, the low-weighted feature *tempsignal=0.048* is added to its representation.

2.2.3 Cross-genre transfer

Semi-supervised approaches have worked well for Twitter annotation in the past (Ritter et al., 2011; Johannsen et al., 2014). This can be attributed to the relative lack of annotated ground truth data for this text type, coupled with the very large amount of text available. These semi-supervised approaches have seen success by inducing distributional representations over text of the target type, from for example Brown clustering or Collobert & Weston embeddings (Turian et al., 2010).

We already have newswire corpora with spatio-temporal entity annotations. We expect that distributional representations induced over one text type will be less effective when applied to another, as observed already by (Maas et al., 2011). Therefore, we hypothesise that inducing such representations over a mixed-genre dataset will afford better cross-genre resilience.

Accordingly, we introduce Brown clusters in a variety of configurations. Brown clustering builds a hierarchical binary tree, with leaves clustering distributionally similar words, for a preset number of leaves m (Brown et al., 1992). This generates features unsupervised, by extracting paths to leaves containing a given word. As we rely heavily on this technique, a short introduction to it follows.

Brown clustering uses distributional information to group similar words. Unsupervised, it induces a hierarchical clustering over words to form a binary tree (e.g., Figure 2.1). Brown clustering uses mutual information to determine distributional similarity, placing similar words in the same cluster and similar clusters nearby in the tree.

In practice, Brown clustering takes an input corpus and number of classes m, and uses mutual information to assign each term to one of the m classes. Ideally, each class contains highly semantically-related words, by virtue of words being distributed according to their meaning (Wittgenstein, 1953). Each class is a leaf on an unbalanced binary tree. The path from the root to each leaf can be described as a bit string, where the i'th bit is 0 iff the path branches left at depth i (e.g., *you,I* is on the path `01` in Figure 2.1). Brown clustering posits that leaves with longer common path prefixes are more semantically related. For example, here, the *cats,dog* and *you,I* classes are more similar than either is to the *love,pet* class.

The fact that Brown clustering as implemented only relies upon bigram information means that the only linguistic tool required to use it is a tokeniser. As we have developed tokenisers for social media text in *all* of the project languages as part of D2.2, we can apply this tool directly without further manual intervention, using the data gathered with tools such as those from D6.1.1.

We build models based on 6000 classes, as opposed to the classical 800-1000 used in NLP. This is designed to permit better capture of the lexical variety expected in social media text. Following the intuition of the Stanford NER tool in ignoring PoS labels as the provided features are a superset of those given to its PoS tagger, we also decided against including separate PoS-tag features here, knowing that we already have a lot of data, including Brown clusters, which are known to support good discriminative PoS taggers themselves (Blunsom and Cohn, 2011).

Our approach is to use Brown clustering based on the RCV1 corpus (Rose et al., 2002) mixed with English-language tweets. We take 64M tokens of the *cleaned* RCV1 corpus (Liang, 2005), and mix this with 64M tokens of English tweets from an archive of the Twitter "garden hose", a 10% feed of all tweets, restricted to just English-language tweets using langid.py (Lui and Baldwin, 2012). We chose this as langid.py was found to be the highest-performing tool in an empirical comparison of entity recognition in tweets (Derczynski et al., 2015c).

2.2.4 Window features

In addition to the clusterings used above, we use n-gram window features, word shape window features, and part-of-speech tag window features. Although dependencies would link sentences together better, n-gram window features have given state-of-the-art performance in temporal boundary detection before (Llorens et al., 2011), and so we stick with the simpler representation rather than potentially introduce noise through tweet dependencies. The window used is [-2,2] with unigram and bigram features. This is implemented through an extension of the CRFsuite API, using the same hardness as USFD developed for their W-NUT entry (Derczynski et al., 2015a).

Task	Precision	Recall	F1
using blended data			
Event recognition	68.55	69.29	68.92
Timex recognition	59.57	52.83	56.00
Location recognition	81.25	64.36	71.82
Spatial entity recognition	48.15	18.06	26.26
using only rumour data			
Location recognition	67.86	42.22	52.05
Spatial entity recognition	28.57	5.88	9.76

Table 2.1: Spatio-temporal entity recognition in tweets

2.2.5 Classifier and features

We attempt to reach high performance by modelling the language used in newswire and tweets concurrently. To achieve this, we use unsupervised feature extraction through Brown clusters, blending both newswire and Twitter input. Additionally, we increase the number of input clusters in order to achieve better cluster resolution, as we are using features that extract paths and not just the eventual clusters (Derczynski et al., 2015b).

We train our model using CRFsuite's linear chain implementation, using both L-BFGS updates (standard) and also passive-aggressive updates to compensate for Twitter noise, which have been shown to work well with single entity type annotation in tweets (Derczynski and Bontcheva, 2014).

2.2.6 Training data

Annotations are made over 400 tweets taken from four rumour instances studied across PHEMEand delivered in D8.2. These tweets are also annotated for their role in rumour spread elsewhere in WP2, and for named entities by SWI. The rumours from which English tweets were drawn are: the Charlie Hebdo shootings; Putin's 2015 disappearance; the Germanwings CFIT event; and the shootings in Ferguson. These are annotated for ISO-TimeML event and temporal expression, and ISO-Space location and spatial_entity (with the above constraints from Section 2.1 applied). The resulting dataset contained 605 events, 122 timexes, 139 spatial_entities and 223 locations.

These were combined with existing gold-standard datasets. For the temporal entity annotation, the TempEval-2 data (Verhagen et al., 2010) was mixed in. For the spatial entity annotation, the W-NUT data (Baldwin et al., 2015) was added, mapping facility to spatial_entity and geo-loc to location.

The data was split 80% / 20% training/evaluation.

2.3 Evaluation

Results for spatial and temporal entity annotation are given in Table 2.2.6. This is the first attempt at spatio-temporal entity annotation in tweets, and so there are no universal baselines to compare to. However, in newswire, the closest equivalent evaluation for temporal annotation is TempEval-3 (UzZaman et al., 2013). In this instance, the best F1 for temporal expression extraction was 90.32, drawn between NavyTime and SUTime. For events, the best F1 was 81.05, from the ATT system. Bear in mind that these figures are over newswire, which is both constrained in variety of expression (Eisenstein, 2013), and also has a much larger preceding body of research and resources. In terms of location and spatial entity recognition, the comparable system is that from W-NUT; here, the best system (Ousia) reached F1 of 34.48 for spatial_entity equivalents and 66.42 for locations. Indeed, we found the spatial_entity extraction task very difficult as well; it may need better definition. Of greater interest is the high performance of location extraction, which due to its larger granularity is likely to have a significantly greater impact when it comes to contextualising and grounding events.

The extra spatial data used to bolster the rumours was drawn from a dataset where only named entities are annotated. It is still possible for non-named entities to be useful spatial_entities; for example, in "The road to Misrata", the road is an important entity. Thus, there is a decent chance of false negative spatial_entities in the W-NUT data used to supplement training data for the spatial recognition task. However, not using the extra data leaves us with a very small dataset, which is hard for a classifier to generalise over. We evaluated the impact of this extra named data by training a model on spatial information in just rumour tweets, holding out the Putin rumours dataset as test data. This model performed worse than the model using named entity data only, though location recognition was still somewhat reasonable.

Finally, we note that in almost every case, recall was lower than precision – an interesting constant in Twitter entity extraction (Derczynski et al., 2013a) that is also apparent in this new task.

Our system is the first attempt at ISO-compliant spatio-temporal entity grounding over tweets, and performs reasonably well – while other systems have performed better, these comparisons are only an approximation to how they might to do on this semantic annotation task, and included for the sake of completeness.

2.4 Conclusion

We have presented novel and reasoned approaches to spatio-temporal semantic annotation in tweets. These include representations for crossing the text type boundary between social media and newswire, and developing word representations that span this significant gap in text processing. As well as being shared within PHEME, our tools and datasets are made available as standalone, open-source utilities and datasets, furthering research in this challenging and important area.

The resulting data can be found at: http://gate.ac.uk/data/st-socmed-tool.html and also from the http://pheme.eu website.

Chapter 3

Spatial Grounding

This chapter discusses the grounding of locations in documents to linked data through Nomenclature of Territorial Units for Statistics (NUTS) subdivisions (c.f. ISO 3166 / Eurostat). This is useful because it provides an explicit, unambiguous location for a document, accessible in universal unified linked data formats. Such grounding is not intrinsic in typical linked data grounding, to e.g. DBpedia or WikiData.

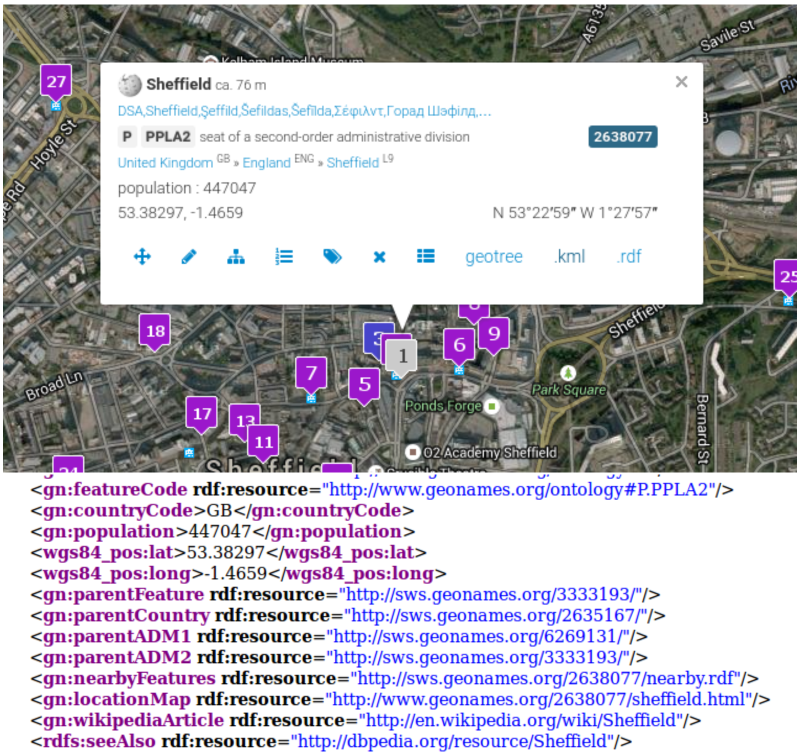
Being able to spatially ground documents – and the rumours that they may contain – is critical to building a machine-readable representation of the rumour and its immediate context. Without it, locations and therefore the context of the rumour become ambiguous, and it is then difficult to fact check claims. For example, if we cannot tell which Paris a given tweet refers to in its claim "Paris taxes set to double", then we do not know if this claim refers to Paris, Texas or Paris, France. This can lead to an inability to interpret claims in rumours, due to incorrect contextualisation.

The ability to ground things accurately in space is also important. Failing to do this can lead to rumours being mis-identified, and incorrect clarifications presented. Continuing the example, if taxes are set to double in Paris Texas but the lcoation is mis-identified and a refutation issued citing tax plans for Paris France, the authority of the fact-checking is diminished, and this always takes a long time to recover. Unfortunately, while hard in news documents, disambiguation of spatial context in tweets and other social media streams is even harder (Derczynski et al., 2013b).

We attempt to address this, connecting location mentions in social media text to NUTS subdivisions, thus spatially grounding them.

3.1 Resources

The resources we use are mostly open source, and consist of linked data repositories and systems for grounding entities. In addition, we use the spatio-temporal entity recognition systems developed in Chapter 2.



Sheffield ca. 76 m

DSA,Sheffield,Şeffild,Šefildas,Šefīlda,Σέφιλντ,Горад Шэфілд,...

P PPLA2 seat of a second-order administrative division 2638077

United Kingdom ^{GB} » England ^{ENG} » Sheffield ^{L9}

population : 447047

53.38297, -1.4659 N 53°22′59″ W 1°27′57″

```
<gn:featureCode rdf:resource="http://www.geonames.org/ontology#P.PPLA2"/>
<gn:countryCode>GB</gn:countryCode>
<gn:population>447047</gn:population>
<wgs84_pos:lat>53.38297</wgs84_pos:lat>
<wgs84_pos:long>-1.4659</wgs84_pos:long>
<gn:parentFeature rdf:resource="http://sws.geonames.org/3333193/"/>
<gn:parentCountry rdf:resource="http://sws.geonames.org/2635167/"/>
<gn:parentADM1 rdf:resource="http://sws.geonames.org/6269131/"/>
<gn:parentADM2 rdf:resource="http://sws.geonames.org/3333193/"/>
<gn:nearbyFeatures rdf:resource="http://sws.geonames.org/2638077/nearby.rdf"/>
<gn:locationMap rdf:resource="http://www.geonames.org/2638077/sheffield.html"/>
<gn:wikipediaArticle rdf:resource="http://en.wikipedia.org/wiki/Sheffield"/>
<rdfs:seeAlso rdf:resource="http://dbpedia.org/resource/Sheffield"/>
```

Figure 3.1: Geonames entry and XML for Sheffield. Note the DBpedia link.

Spatial grounding is achieved through three non-DBpedia mechanisms. Firstly, we ground in terms of latitude and longitude if these are available in the social media post's metadata (which accounts for a growing proportion of work; (Sadilek et al., 2012)). Secondly, we use the Geonames linked data resource to connect location mentions to this extensive formal set of conurbations, political regions and geographical features (Figure 3.1). Finally, we map locations into the EU NUTS location space.

3.2 Method

The goal is to connect mentions of spatial objects to canonical, unambiguous identifiers. This aids computational processing and reasoning about these mentions.

We make use of the the YODIE state-of-the-art LOD-based entity grounding tools, developed at USFD in the TrendMiner project,[1] and published recently (Gorrell et al., 2015). These dereference entity mentions in tweets to linked data entries in DBpedia, where available. In order to link the automatically-detected locations from the CRF-

[1] http://www.trendminer-project.eu/

based classifier in Section 2.3, we include any candidates that YODIE generates which match the token spans corresponding to entities marked as spatial locations. This enables the recognition of entities that might otherwise have been discounted had YODIE been run on its own, and is intended to overcome the recall problems dominant in social media entity extraction (Derczynski et al., 2015c).

Having found a DBpedia reference for locations detected in a document, this can then be linked to other resources. GeoNames includes explicit references to DBpedia entries, and so can be mapped. NUTS can also be mapped, though is slightly more difficult as fine-grained entries do not have exact references. However, YODIE generates enough information to select NUTS regions for spatial entities grounded in the EU (i.e. the NUTS-covered region), enabling linking.

To do the linking, we avoid conventional triple stores like OWLIM, in favour of custom direct lookups. Many commercial and enterprise-grade triple stores are designed for a single large initial load of updates and then for smaller subsequent changes, with fast reads being a performance priority. To achieve this priority, they compute all inferences between triples entered, thus pre-caching this information, in order to speed up reads and queries. However, as we are doing direct lookups and do not require any transitive reasoning between the linked data resources which we are using, this is a wasteful technique; it takes weeks to run on datasets at our scale, consuming many terabytes of storage. Instead, we perform direct lookups once, building a pair-store mapping DBpedia to GeoNames and NUTS.

3.3 Conclusion

The outcome is an enrichment of entities in tweets, grounded spatially in different manners. This comprises our spatial contextualisation of the content. We found that the (less-specific) spatial_entity type (and there also the facility NE type) were grounded less regularly. These tend to describe either vague regions, or small and precise regions like bars or car repair garages; there is not enough coverage for this kind of location.

3.3.1 Analysis

This finding, that locations were grounded more regularly, led to two observations. The approach is appropriate at the level of granularity used for grounding and disambiguating rumours – usually city-level or sub-city-level, but not house-level. There is also a manifestation of the classic tradeoff between precision and recall: the broader, less-precise locations are covered well, but the very-precise positional information is harder to link in a semantic way, usually just available as GPS co-ordinates. Currently, in our datasets, while 55% of tweets included some kind of co-ordinates, only 15% of tweets came from a user who had enabled geolocation, and just 2% had point-level accuracy from the source device – suggesting that the majority of geographic information embedded in Twitter metadata is of questionable accuracy, relying on things like user profile and coarse IP geolocation. This drives demand for indirect ways of spatially grounding content.

3.3.2 Future work

Linking a more fine-grained resource, like Foursquare, to the spatial entities discussed in tweets, is possible using GPS co-ordinates when available. This can lead to the creation of new entries in geographical resources. In the interim, some placeholder technique for handling these "unlisted" resources such that they can be use as disambiguation/grounding resources while tracking a rumour would also be useful – after all, once we have a few location names and concurrent GPS co-ordinates, a location can confidently be described. However, the city-level grounding and entity name capture from the tokens used to represent a location in social media posts is ample for the purposes of PHEME.

Task T6.2 will contain an extended evaluation of the spatio-temporal grounding deliverable, including spatial grounding.

Chapter 4

Cross-lingual Spatio-Temporal Entity Projection

Identifying mentions of places and times is a difficult and expensive annotation task. It has mostly only been thoroughly explored for English, and for newswire. This has lead to a dearth of resources in two of the project's three languages.

To eliminate annotation cost and effort, annotations in one language can be projected into another. As we already have reference annotations in English (Chapter 2), it should be possible to map these onto new languages. The ISO-TimeML framework intended to be language-independent, though largest body of work using it comprised of applications to English.

This chapter details the projection of spatial and temporal annotations from English to German and Bulgarian text. The end result is linguistic resources: models and techniques for achieving the projections, and the final annotated documents themselves.

4.1 Corpus projection

Many modern approaches to semantic annotation use supervised machine learning, to build quality NLP tools. Annotating spatial and temporal entity mentions is just such a semantic annotation problem, and indeed many approaches to the task have been based on supervised machine learning. However, training data annotated manually by experts is difficult to come by and expensive to create. In an attempt to overcome this shortage, we use *projection* to map annotations over a corpus in one language to a version of that same corpus in another language. This increases the amount of annotated information available, for use in e.g. supervised machine learning approaches, while also removing the conceptual monolingual barrier.

While some systems do exist for automatic temporal or spatial annotation of non-

English text, these are often rule-based and the number of languages supported is low. In particular, there are no resources for Buglarian, and only temporal expression annotation is supported in German, as far as we are aware.

We present an approach to projecting spatial and temporal entity annotations from English to the two other project languages, German and Bulgarian. The entity types that we project are those defined in Chapter 2: events, timexes, locations, and spatial entities. These are taken from the relevant standards, ISO-TimeML (Pustejovsky et al., 2010) and ISO-Space (Pustejovsky et al., 2011). The two semantic annotation standards include sophisticated means of linking these spatio-temporal entities using extra-document annotation (for example, descriptions of the time order of events, or how entities spatially relate to each other, and already have a tracking in the kind of general entity linking work (Burman et al., 2011) used elsewhere in PHEME. These links can be directly applied to project annotations and do not require additional linguistic processing.

Projected annotations often contain a degree of noise, due to both gold standard errors and errors during word alignment between languages. Additionally, the possibility of achieving coherent projections requires that the lexicalisation of the concept being projected follows the direct correspondence assumption, or DCA (Hwa et al., 2002), which stipulates a homomorphism between source and the literal translation in the target text. Fortunately, the majority of entities being mapped are often short, and some by definition are only one token (e.g. ISO-TimeML event mentions). However, deviation from the DCA cannot be ruled out, and occurs regularly. To deal with this inevitable noise, we adopt two techniques. Firstly, we use data engineering techniques first proposed by (Spreyer and Frank, 2008) in order to limit splitting of tokens and contexts. Secondly, we filter entity mentions through a crowdsourced “sanity check”, asking groups of crowd workers to rule out suspicious annotations.

4.2 Related Work

Projection of linguistic structure has recently been addressed in treebanks (Tiedemann et al., 2014). This sophisticated and powerful technique has informed our approach. However, as we do not rely on capturing syntactic relations, but rather on mapping surface lexicalisations from one language to the another and then re-attaching semantic information, the linguistic level at which this work operated is denser than required, and we can afford to use a shallower system without losing anything.

For instance, in Example 4.2, the word ordering is different in each language. However, item 1 happens temporally before item 2 in each instance, and this transcends the choice of tongue used to express the idea. Therefore, the temporal ordering information (which is semantic) doesn’t depend on the surface forms. The only thing that is important for projection is to get the items 1 and 2 mapped to the correct words.

(4.1) The newspaper was printed **today**$_1$ and will be burned **tomorrow**$_2$

(4.2) Die Zeitung wurde **heute**$_1$ gedruckt und **morgens**$_2$ gebrannt

Similar techniques have been used for mapping between two reference translations and applying an automatic tool then converting its results to the target language (Spreyer and Frank, 2008). This approach is accurate enough to enable learning of a labeller from the projections; however, we find the idea of taking automatically created semantic annotations and then projecting them for use as training data to be too lossy. We can afford to start from gold-standard newswire corpora and only lose information during projection, rather than initial annotation as well. As automatic semantic annotation of spatio-temporal information is still difficult (Bethard et al., 2015; Pustejovsky et al., 2015), there is still an inherently large risk in using automatically-generated source annotations. However, we do adopt some of the logic-based approaches in this work for handling conflicting and split annotations.

4.3 Methods and data

Using the adapted annotation criteria described in Section 2.1, we isolate a sub-set of ISO-TimeML and ISO-Space entity annotations over the input corpora. These are the spatio-temporal entity lexicalisations that will be projected into the target languages.

The source texts are then translated automatically into the destination language, using the SDL language cloud [1]. This gives a parallel corpus. Word alignments are then extracted using an unsupervised tool, cdec (Dyer et al., 2010). We choose this tool because it is unsupervised and therefore readily adaptable to any target language. These alignments create token-to-token mappings between the corpora.

The spatio-temporal annotations and mappings can then be used to build new annotations on the translated documents. These projections are filtered in two passes.

The corpora and entities projected were as follows:

- TimeBank 1.2: 61k tokens, 7900 events, 1400 temporal expressions;
- ACE 2005 SpatialML annotations: 290k tokens, 6900 places as follows: 15 celestial, 616 civil, 86 continent, 3557 country, 538 facility, 6 mountain, 6 mountainous, 3 postal code, 677 populated, 244 populated A, 689 populated B, 288 region, 64 road, 2 vehicle, 68 water.

We divide the SpatialML types, from an earlier iteration of the ISO-Space standard, into spatial entities and locations as follows:

- *spatial_entity*: facility, postal code, road, vehicle

[1] https://languagecloud.sdl.com/

- *location*: celestial, civil, continent, country, mountain, mountainous, populated, populated A, populated B, region, water.

4.3.1 Projection filtering 1: Data engineering

The token mappings are filtered according to some basic constraints regarding contiguous annotations, as specified in (Spreyer and Frank, 2008). In this case, mappings from single words to other single words are fine, as are mappings from contiguous blocks of words to single words, or mappings from single words to contiguous blocks of words. Problems arise when there are non-contiguous aligned spans, or when multiple tags map to the same token.

In the case of the first problem, going from a single word to a broken sequence of words, we ignore the break if it is just one or two tokens long, assuming that the target language needs some extra words to express the same concept. If the required break is longer, the longest contiguous block is chosen, or the first block in case of a tie. For the second problem, multiple annotations colliding on one destination annotation, the choice is made arbitrarily, and given priority. As locations and temporal expressions are (respectively) generally less ambiguous than spatial entities and events, we prioritise these two types.

4.3.2 Projection filtering 2: Crowdsourcing

For the second quality assurance pass, we implement a semi-manual "sanity check". This uses crowd sourcing, posing a task through the GATE crowd sourcing plugin (Bontcheva et al., 2014) where workers are given a rough idea of an entity and presented with a projected annotation. They are asked if the highlighed, projected entity matches the description (e.g. "Could the highlighted phrase describe a time or date?"). We are content to include uncertain annotations as positives, as the task is known to be difficult for experts, and crowd recall is difficult to achieve (Trushkowsky et al., 2013). The task is presented in the target language, and geographically restricted to the target language's region; i.e., only workers in Bulgaria were asked to judge Bulgarian annotations.

4.4 Results and analysis

4.4.1 Translation and alignment

The documents are translated using the SDL API, in chunks of 4000 characters, to remain within API limits. Here's an example input and output:

- *English from gold standard:* If Israel is asked to uh stretch itself on matters that are vital to its security concerns, then we must see an equal effort on the other side.
- *German output:* Wenn Israel aufgefordert wird, uh selbst dehnen auf Fragen, die so wichtig sind für ihre Sicherheit sorgen, dann müssen wir eine gleiche Anstrengung auf der anderen Seite.
- Bulgarian output: Ако Израел е помолен да uh разтягане себе си по въпроси, които са от жизненоважно значение за своята загриженост по отношение на сигурността, тогава ние трябва да видите равна усилия на другата страна.

After this, cdec is used to generate token-level alignments between the source and target translations. It takes a set of pairs of sentences, and outputs a sequence of token-to-token offset relations from the source to target text. So, for the German and Bulgarian examples in the sentence above:

- EN-DE: 0-0 1-1 2-2 3-3 4-4 5-5 6-6 8-7 9-8 13-9 10-10 11-11 13-12 13-13 14-14 16-15 17-16 18-17 19-18 20-19 21-20 22-21 23-22 24-23 25-24 26-25 27-26
- EN-BG: 0-0 1-1 2-2 3-3 1-4 4-5 5-6 6-7 7-8 8-9 9-10 9-11 12-14 13-15 14-17 15-18 13-19 17-20 18-22 19-23 20-24 21-25 24-26 23-27 24-28 25-29 25-30 26-31 27-32

By this point, we have automatic translations and also token-to-token mappings from the original text to the target. Some noise is guaranteed – these systems are automatic – and so there are later filterings to remove this. The remaining step is the projection.

4.4.2 Event and timex projection

Using alignments from cdec and token indices from CAVaT (Derczynski and Gaizauskas, 2012), we created a CoNLL-style output of events and timexes from English TimeBank to Bulgarian and German. The loaded data looks like this (for events, in this case):

```
162|e1|OCCURRENCE|snapping|snap|246|1|39
162|e2|OCCURRENCE|loss|loss|303|1|47
162|e4|REPORTING|said|say|326|1|53
162|e5|OCCURRENCE|earned|earn|334|1|55
162|e6|REPORTING|reported|report|447|2|8
162|e240|STATE|$538.5|$538.5|465|2|11
162|e241|STATE|$388.5|$388.5|524|2|25
162|e8|OCCURRENCE|earned|earn|587|3|7
162|e10|REPORTING|said|say|716|5|6
162|e11|I_STATE|expected|expect|724|5|8
```

Where the last two values describe the document sentence and token ID respectively. Using the token ID and the cdec alignments, we project the corresponding annotation (and its attributes) into the target language.

Note that the TimeML makeinstance type is discarded. This type is used to label lexicalisations of events and eventualities, and there are some edge cases where more than one temporal event is evoked by (or represented in) the same word. However, this multiple instantiation is so rare that makeinstance was de facto dropped after ISO-TimeML, and its attributes overloaded onto the event type. The event instance ID attributed is retained, however, and multiple instantiations are now evoked by means of a standoff event that uses the same event instance ID (eiid).

For timexes, we try to map the entire phrase. Timexes can (and often are) more than one token long. When the phrase is split, we include up to two interlocuting words (as long as the timex is three or longer words), or take the longest or earliest group, as above.

4.4.3 Location and spatial entity projection

Projection of the spatial phrases – locations and spatial entities – worked just as with the events and timexes. These phrases were often longer, so were more prone to being projected broken; however, in the target languages, the phrases were actually split or rearranged less often than temporal expressions. So, while at greater risk of being hard to project, the actual proportion of contiguous projections was higher, yielding higher-confidence results.

Note that we also transduce the corpus from SpatialML to ISO-Space through the splitting of the "place" type to location and spatial_entity.

4.4.4 Crowdsourced filtering

As a final step, we used the crowd to filter out clearly wrong results. For each entity, we created a CrowdFlower job (Biewald, 2012) using the GATE Crowdsourcing plugin (Bontcheva et al., 2014). As crowd recall is problematic (Trushkowsky et al., 2013), and affirmative and negative statements have different response rates (Wu and Marian, 2014), we phrased the question so that workers would only highlight items that they were sure were wrong; if there's uncertainty, the item stays in. This is to compensate for the complexity in training annotators for spatio-temporal marking, and to avoid throwing out crowd-debatable items that have come from a correct gold-standard corpus.

4.5 Conclusion

This chapter detailed the projection of semantic annotations from English into the project's two other languages. ISO-standard annotated resources were created, annotated for both spatial and temporal entities, in Bulgarian and English. The resulting data can be found at: http://gate.ac.uk/data/st-proj-corpora.html and also from the http://pheme.eu website.

4.5.1 Future work

The resulting corpora comprise a substantial amount of high-quality data, semantically annotation for space and time using a dominant standard. The German data is the first TimeML gold-standard based dataset; previous work used automatically-created annotations from TARSQI, an automatic temporal annotation system (Verhagen et al., 2005). The spatial annotations in both languages, and the Bulgarian temporal annotations, are all the first of their kind.

The resulting projected corpora can be used as training data for machine-learning systems for spatio-temporal entity extraction. These would be the first tools of their kind for these languages. Additionally, with the adaptations and techniques described in Chapter 2, these can be adapted for social media usage.

In PHEME, the next and final step with these resources is to evaluate their performance in unseen, real-world data, as part of task T6.4, Accuracy and Scalability Evaluation.

Chapter 5

Conclusion and Future Work

Spatio-temporal grounding allows us to contextualise and therefore correctly recognise the content of claims on the web. Extraction this kind of information helps us determine the meaning of claims made on the web, by mining the semantics behind mentions made by people. We introduced a text-type-insensitive technique for unsupervised entity recognition, and used this to merge two drastically different text types – newswire and tweets. In turn, this allowed us to achieve good performance in entity extraction with a relatively modest amount of manual annotation, by leveraging prior gold-standard annotations and also a large volume of plain text that was gathered previously in PHEME and pre-processed with the tools in D2.2.

We use state-of-the-art techniques (hyperparameter-tuned Brown clustering and passive-aggressive CRF) for spatio-temporal entity extraction. The entities found are then grounded through intersection with a new, high-performance information extraction framework, YODIE. The resulting documents are to be integrated with the Apache Kafka framework. PHEME uses Apache Kafka (Kafka, 2011) as the integration framework. Kafka provides a high-throughput, low-latency, distributed platform for handling real-time data streams. It follows a producer-consumer model where the producers send messages (topics) over the network to the Kafka cluster which in turn serves them to the consumers. This allows easy and desynchronised annotation and sharing of data and documents between all partners, so that spatio-temporal annotation can be decoupled from a pipeline.

Future work will investigate the effectiveness of the new feature representations presented and apply them to other tasks in these languages, to determine the likelihood of rapidly developing high-performance tools for social media processing. In addition, we can use the tools presented here to connect with event clusters (WP6), providing extra spatial and longitudinal information for their visualisation (WP5) as well as grounding them in linked data.

Finally, we will apply these tools to the use cases in WP7 and WP8, adding spatial and temporal context to the claims and rumours found within these diverse and impactful scenarios – community medical information and formal journalistic reporting.

Bibliography

ACE (2004). *Annotation Guidelines for Entity Detection and Tracking (EDT)*. Available at http://www.ldc.upenn.edu/Projects/ACE/.

Baldwin, T., Cook, P., Lui, M., MacKinlay, A., and Wang, L. (2013). How noisy social media text, how diffrnt social media sources. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 356–364.

Baldwin, T., Han, B., de Marneffe, M. M. C., Kim, Y.-B., Ritter, A., and Xu, W. (2015). Findings of the 2015 Workshop on Noisy User-generated Text. In *Proceedings of the Workshop on Noisy User-generated Text (WNUT 2015)*. Association for Computational Linguistics.

Bethard, S. (2013). A synchronous context free grammar for time normalization. In *EMNLP*, pages 821–826.

Bethard, S., Derczynski, L., Pustejovsky, J., and Verhagen, M. (2015). SemEval-2015 Task 6: Clinical TempEval. In *Proceedings of the 9th International Workshop on Semantic Evaluation*. Association for Computational Linguistics.

Biewald, L. (2012). Massive multiplayer human computation for fun, money, and survival. In *Current trends in web engineering*, pages 171–176. Springer.

Blunsom, P. and Cohn, T. (2011). A hierarchical Pitman-Yor process HMM for unsupervised part of speech induction. In *Proc. ACL*, pages 865–874.

Boguraev, B. and Ando, R. K. (2006). Analysis of TimeBank as a resource for TimeML parsing. In *Proceedings of LREC*, pages 71–76.

Boguraev, B., Pustejovsky, J., Ando, R., and Verhagen, M. (2007). TimeBank evolution as a community resource for TimeML parsing. *Language Resources and Evaluation*, 41(1):91–115.

Bollacker, K., Evans, C., Paritosh, P., Sturge, T., and Taylor, J. (2008). Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 1247–1250. ACM.

Bontcheva, K., Roberts, I., Derczynski, L., and Rout, D. (2014). The GATE Crowdsourcing Plugin: Crowdsourcing Annotated Corpora Made Easy. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*. Association for Computational Linguistics.

Brown, P., Della Pietra, V., de Souza, P., Lai, J., and Mercer, R. (1992). Class-based n-gram models of natural language. *Computational Linguistics*, 18:467–479.

Burman, A., Jayapal, A., Kannan, S., Kavilikatta, M., Alhelbawy, A., Derczynski, L., and Gaizauskas, R. (2011). USFD at KBP 2011: Entity linking, slot filling and temporal bounding. In *Proc. Text Analysis Conference*.

Derczynski, L., Augenstein, I., and Bontcheva, K. (2015a). USFD: Twitter NER with Drift Compensation and Linked Data. In *Proceedings of the Workshop on Noisy User-genereated Text (W-NUT)*. ACL.

Derczynski, L. and Bontcheva, K. (2014). Passive-aggressive sequence labeling with discriminative post-editing for recognising person entities in tweets. *EACL 2014*, page 69.

Derczynski, L., Chester, S., and Bøgh, K. S. (2015b). Tune your Brown Clustering, Please. In *To appear*. Association for Computational Linguistics.

Derczynski, L. and Gaizauskas, R. (2011). A Corpus-based Study of Temporal Signals. In *Proceedings of the 6th Corpus Linguistics Conference*.

Derczynski, L. and Gaizauskas, R. (2012). Analysing temporally annotated corpora with CAVaT. In *Proc. LREC*, pages 398–404.

Derczynski, L., Maynard, D., Aswani, N., and Bontcheva, K. (2013a). Microblog-Genre Noise and Impact on Semantic Annotation Accuracy. In *Proceedings of the 24th ACM Conference on Hypertext and Social Media*. ACM.

Derczynski, L., Maynard, D., Rizzo, G., van Erp, M., Gorrell, G., Troncy, R., Petrak, J., and Bontcheva, K. (2015c). Analysis of named entity recognition and linking for tweets. *Information Processing & Management*, 51(2):32–49.

Derczynski, L., Yang, B., and Jensen, C. (2013b). Towards Context-Aware Search and Analysis on Social Media Data. In *Proceedings of the 16th Conference on Extending Database Technology*. ACM.

Dyer, C., Weese, J., Setiawan, H., Lopez, A., Ture, F., Eidelman, V., Ganitkevitch, J., Blunsom, P., and Resnik, P. (2010). cdec: A decoder, alignment, and learning framework for finite-state and context-free translation models. In *Proceedings of the ACL 2010 System Demonstrations*, pages 7–12. Association for Computational Linguistics.

Eisenstein, J. (2013). What to do about bad language on the internet. In *Proceedings of NAACL-HLT*, pages 359–369.

Ennals, R., Trushkovsky, B., and Agosta, J. (2010). Highlighting disputed claims on the web. In *WWW'10*.

Fromreide, H., Hovy, D., and Søgaard, A. (2014). Crowdsourcing and annotating NER for Twitter #drift. *European language resources distribution agency*.

Gaizauskas, R., Barker, E., Chang, C., Derczynski, L., Phiri, M., and Peng, C. (2012). Applying ISO-Space to Healthcare Facility Design Evaluation Reports. In *Proceedings of the Joint ISA-7, SRSL-3 and I2MRT Workshop on Semantic Annotation and the Integration and Interoperability of Multimodal Resources and Tools*, pages 31–38.

Gorrell, G., Petrak, J., and Bontcheva, K. (2015). Using @Twitter conventions to improve #lod-based named entity disambiguation. In *The Semantic Web. Latest Advances and New Domains*, pages 171–186. Springer.

Hwa, R., Resnik, P., Weinberg, A., and Kolak, O. (2002). Evaluating translational correspondence using annotation projection. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 392–399. Association for Computational Linguistics.

Johannsen, A., Hovy, D., Alonso, H. M., Plank, B., and Søgaard, A. (2014). More or less supervised supersense tagging of Twitter. *Lexical and Computational Semantics (* SEM 2014)*, page 1.

Kafka, A. (2011). http://kafka.apache.org/.

Kergl, D., Roedler, R., and Seeber, S. (2014). On the endogenesis of Twitter's Spritzer and Gardenhose sample streams. In *Advances in Social Networks Analysis and Mining (ASONAM), 2014 IEEE/ACM International Conference on*, pages 357–364. IEEE.

Kripke, S. A. (1972). *Naming and necessity*. Springer.

Kusmierczyk, T., Trattner, C., and Nørvåg, K. (2015). Temporality in online food recipe consumption and production. In *Proceedings of the 24th International Conference on World Wide Web Companion*, pages 55–56. International World Wide Web Conferences Steering Committee.

Liang, P. (2005). Semi-supervised learning for natural language. Master's thesis, Massachusetts Institute of Technology.

Llorens, H., Derczynski, L., Gaizauskas, R. J., and Saquete, E. (2012). TIMEN: An open temporal expression normalisation resource. In *LREC*, pages 3044–3051.

Llorens, H., Saquete, E., and Navarro, B. (2011). Syntax-motivated context windows of morpho-lexical features for recognizing time and event expressions in natural language. In *Natural Language Processing and Information Systems*, pages 295–299. Springer.

Lui, M. and Baldwin, T. (2012). langid.py: An off-the-shelf language identification tool. In *Proceedings of the ACL 2012 system demonstrations*, pages 25–30. Association for Computational Linguistics.

Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., and Potts, C. (2011). Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 142–150. Association for Computational Linguistics.

Marcelo, M., Barbara, P., and Carlos, C. (2010). Twitter under crisis: Can we trust what we RT? In *1st Workshop on Social Media Analytics (SOMA)*.

Masud, M. M., Gao, J., Khan, L., Han, J., and Thuraisingham, B. (2011). Classification and novel class detection in concept-drifting data streams under time constraints. *Knowledge and Data Engineering, IEEE Transactions on*, 23(6):859–874.

Plank, B., Alonso, H. M., and Søgaard, A. (2015). Non-canonical language is not harder to annotate than canonical language. In *The 9th Linguistic Annotation Workshop held in conjuncion with NAACL 2015*, page 148.

Pustejovsky, J., Kordjamshidi, P., Moens, M.-F., Levine, A., Dworman, S., and Yocum, Z. (2015). SemEval-2015 Task 8: SpaceEval. In *Proceedings of the 9th International Workshop on Semantic Evaluation*. Association for Computational Linguistics.

Pustejovsky, J., Lee, K., Bunt, H., and Romary, L. (2010). ISO-TimeML: An international standard for semantic annotation. In *LREC*.

Pustejovsky, J., Moszkowicz, J. L., and Verhagen, M. (2011). ISO-Space: The annotation of spatial information in language. In *Proceedings of the Sixth Joint ISO-ACL SIGSEM Workshop on Interoperable Semantic Annotation*, pages 1–9.

Regneri, M., Koller, A., and Pinkal, M. (2010). Learning script knowledge with web experiments. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 979–988. Association for Computational Linguistics.

Ritter, A., Clark, S., Mausam, and Etzioni, O. (2011). Named entity recognition in tweets: An experimental study. In *Proc. of Empirical Methods for Natural Language Processing (EMNLP)*.

Rose, T., Stevenson, M., and Whitehead, M. (2002). The Reuters Corpus Volume 1 - from Yesterday's News to Tomorrow's Language Resources. In *LREC*, volume 2, pages 827–832.

Sadilek, A., Kautz, H., and Silenzio, V. (2012). Modeling spread of disease from social interactions. In *International AAAI Conference on Weblogs and Social Media (ICWSM)*, pages 322–329. AAAI.

Schneider, N. (2015). What i've learned about annotating informal text (and why you shouldn't take my word for it). In *Proceedings of The 9th Linguistic Annotation Workshop*, pages 152–157. ACL.

Spreyer, K. and Frank, A. (2008). Projection-based acquisition of a temporal labeller. In *IJCNLP*, pages 489–496. ACL.

Tiedemann, J., Agic, Z., and Nivre, J. (2014). Treebank translation for cross-lingual parser induction. In *CoNLL*, pages 130–140.

Trushkowsky, B., Kraska, T., Franklin, M. J., and Sarkar, P. (2013). Crowdsourced enumeration queries. In *Data Engineering (ICDE), 2013 IEEE 29th International Conference on*, pages 673–684. IEEE.

Tufekci, Z. (2014). Big questions for social media big data: Representativeness, validity and other methodological pitfalls. *arXiv preprint arXiv:1403.7400*.

Turian, J., Ratinov, L., and Bengio, Y. (2010). Word representations: a simple and general method for semi-supervised learning. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 384–394. Association for Computational Linguistics.

UzZaman, N., Llorens, H., Derczynski, L., Verhagen, M., Allen, J. F., and Pustejovsky, J. (2013). SemEval-2013 Task 1: TempEval-3: Evaluating Time Expressions, Events, and Temporal Relations. In *Proceedings of the 7th International Workshop on Semantic Evaluations*.

Verhagen, M., Mani, I., Sauri, R., Knippen, R., Jang, S. B., Littman, J., Rumshisky, A., Phillips, J., and Pustejovsky, J. (2005). Automating temporal annotation with TARSQI. In *Proceedings of the ACL 2005 on Interactive poster and demonstration sessions*, pages 81–84. Association for Computational Linguistics.

Verhagen, M., Sauri, R., Caselli, T., and Pustejovsky, J. (2010). Semeval-2010 task 13: Tempeval-2. In *Proceedings of the 5th international workshop on semantic evaluation*, pages 57–62. Association for Computational Linguistics.

Wittgenstein, L. (1953). *Philosophical Investigations*. Basic Blackwell, London.

Wu, M. and Marian, A. (2014). Corroborating facts from affirmative statements. In *EDBT*, pages 157–168.