EC Project 610829

A Decarbonisation Platform for Citizen Empowerment and Translating Collective Awareness into Behavioural Change

D2.2.2: Text Analytics Tools for Environmental Information Extraction v.2

23 May 2016

Version: 1.0

Version history

Version	Date	Author	Comments
0.1		Diana Maynard	Initial version.
0.2		Diana Maynard	Comments from Miriam added
0.3		Diana Maynard	Addressed reviewer comments

Peer reviewed by: Miriam Fernandez, OU

Dissemination Level: PU – Public

This document is part of the DecarboNet research project, which receives funding from the European Union's 7th Framework Programme for research, technology development and demonstration (Grant Agreement No 610829; ICT-2013.5.5 CAPS Collective Awareness Platforms for Sustainability and Social Innovation).

Executive Summary

This deliverable provides a report to accompany the tools for environmental information extraction delivered. The tools are enabled as web services which perform entity disambiguation, recognition of environmental terms in English and German. Open source versions of the tools are also available.

The report explains how to use the web services, describes the applications and the underlying natural language processing tools used, and details some experiments carried out to evaluate the performance of these tools. The evaluation datasets have been made available publicly. It also explains how these tools have been used in case studies in WP4 for analysis of the Earth Hour campaigns and COP21, to help understand user engagement with these campaigns. Finally, it outlines some remaining improvements and issues to be investigated during the remainder of the project and beyond.

1. Introduction 4

2. ClimaTerm: a web service for term recognition 4

2.1. Improvements to the English term extraction 5
2.1.1. Terms from DBpedia 5
2.1.2. Reclassification of the terms 5
2.1.3. Generating term variants on the fly 6
2.1.4. Using word embeddings to enrich the lexicons 7
2.1.5. Evaluation 9
2.2. Extracting German terms 10
2.2.1. Expanding the German lexicon 11
2.2.2. Evaluation 13
2.3. Software availability 14
3. Extraction of linguistic events 15

3.1. Types of Linguistic Event 15
3.1.1. Questions 16
3.1.2. Directives 17
3.1.3. Conditionals 18
3.1.4. Use of first/second person 19
3.2. Automatic recognition of linguistic modalities 19
3.3. Summary 21
4. Actor recognition in tweets 21

4.1. Related Work 22
4.2. The @Mention Dataset 22
4.3. Methods 24
4.3.1. Baselines 24
4.3.2. The Random Forest Mention Classifier 26
4.3.3. Features 26
4.3.4. Tweet-based Features 28
4.4. Results 28
Table 15: Results on the development dataset 29
4.5. Summary 31
***5.* Recognyze: a service for named entity recognition** 32

5.1. Software availability 32
5.2. Evaluation of Recognyze 33
5.3. **Ongoing Work** 36
5.4. Summary 37
6. **Conclusions and further work** 37

B. List of Tables 39

C. List of Abbreviations 40

D. References 41

1. Introduction

This deliverable describes the second version of the text analytics tools for extracting various kinds of information related to the environment and climate change. There are four main strands of work: improvements to and extension of the term recognition tools (ClimaTerm) described in D2.2.1, including a German version; improvements to and evaluation of the entity recognition tools (Recognyze) described in D2.2.1; new tools for the recognition of actors in tweets; and new tools for the recognition of linguistic events in tweets. The term recognition tools have been evaluated and the datasets made available publicly.

The work described here is closely related to work in WP4 where the tools have been used directly in the case studies about the Earth Hour and COP21 events, in two ways: (1) to provide linguistic criteria for the automatic identification of different stages of behaviour towards climate change based on the users' social media contributions; and (2) in experiments to automatically categorise users into behavioural stages using the analysis produced by the tools. The work is also closely related to WP1, which aims to better understand the nature of behavioural change and to align the theoretical with the practical. WP1 uses the insights from the case studies described in WP4 and the analytics tools described in this report to provide handles to gauge the levels of awareness, engagement, and willingness to change and influence people's behaviour with respect to climate change. A paper has been published resulting from this collaboration [Fernandez 2016].

2. ClimaTerm: a web service for term recognition

As described in D2.2.1, this web service aims to annotate documents with terms related to climate change. In the first version, we investigated various relevant ontologies available as Linked Open Data and chose the two which appeared to be the most relevant: GEMET and REEGLE. The web service takes as input a document or set of documents, and outputs those documents as XML files annotated with term and URI information. The underlying application is developed in GATE[1] and contains the following processing stages:

- linguistic pre-processing: tokenisation, sentence splitting, part-of-speech tagging, morphological analysis
- term extraction: matching against known terms, plus some recognition of morphological and synonym variants
- export as XML (inline annotation)

In the second version, we have improved the results of the term extraction for English, as described in Section 2.1, and added a tool for German term extraction, as described in Section 2.2. The web service for German is identical in design to that of the English one.

[1] http://gate.ac.uk

2.1. Improvements to the English term extraction

Our initial application, described in the previous deliverable D2.2.1, achieved excellent precision but only moderate recall when compared with the gold standard set. In the second version of the tool, we have improved the term recognition in several ways:

- added selected terms from DBpedia to supplement the existing terms (Section 2.1.1);
- removed many spurious terms that were deemed too general (Section 2.1.2);
- added some extra terms semi-automatically (Section 2.1.2);
- incorporated some optionality into the tool to enable switching on or off of different term sets (Section 2.1.2);
- added a component for generation of term variants on the fly (Section 2.1.3);
- investigated ways of extending the lexicons, such as using word embeddings (Section 2.1.4).

Finally, we have also re-evaluated the term generation, and show an improvement on previous evaluations (Section 2.1.5), though work is still ongoing to make further improvements.

2.1.1. Terms from DBpedia

The set of existing terms was expanded by searching for relevant terms from DBpedia. Terms related to the concepts "environment" and "environmental issues" were collected automatically from DBpedia (all subclasses and instances of these concepts in the ontology), manually verified, and added to the gazetteers. The list corresponding to the "environment" concept contains 65 new entries, including terms such as: "anthropocene", "Earth Hour", "Environmental Performance Index", and "eco-industrial development". The list corresponding to the "environmental issues" concept contains 66 new entries, including terms such as: "soil contamination", "water scarcity", "hot stain"[2] and "land degradation". As with the other ontologies (GEMET and REEGLE), when relevant DBpedia terms are matched, they are annotated also with the URI.

2.1.2. Reclassification of the terms

The set of existing terms was also improved by manual addition of some missing terms from our training data (described in D2.2.1), which where possible were linked to an existing URI, using the "prefLabel" feature to indicate that the existing term in the ontology is the preferred term and that this is an alternative variant (in the same way that existing term variants in the various ontologies are handled). For example, the term "fossil fuel" is linked to the Reegle term "fossil energy".

[2] A "hot stain" is a region where safe drinking water has been depleted.

Some clearly spurious terms were manually deleted. Finally, a manual analysis of terms from the two ontologies, Reegle and GEMET, revealed that most of the terms in GEMET were too general to be included, and were only relevant if occurring in a known climate change context. This means that while they are useful in the analysis of tweets about e.g. Earth Hour, they are not relevant to use as indicators that a tweet is about climate change.

This distinction is important because in WP4, the results of the term extraction are used precisely for this case. In our study of Earth Hour and COP21, we used the Twitter IDs of the participants of these events to generate a second collection of data, and to gather historical tweets from their timelines, providing information for up to several years for some users. Naturally, these users post about environmental issues, but they also post about their jobs, hobbies, personal experiences, and so on. To identify which of the tweets produced by the users relates to their environmental behaviour, we used the ClimaTerm tool. For Earth Hour 2015, out of over 56 million tweets, 750,538 were identified as being climate-related according to the tool. These filtered tweets were then used to automatically categorise users into different behavioural stages over time.

In order to preserve the distinction between terms from different ontologies and therefore different levels of usefulness, we add a feature to our annotations noting which ontology the term is related to, and we also have a (manually operated) switch in our application that enables only the terms we are confident about to be annotated in cases where the context is not specifically climate-related.

2.1.3. Generating term variants on the fly

We also developed a module in GATE to extend the set of known terms, by matching variants found in the text which are related to a member of this set. This uses a set of JAPE pattern-matching rules to match not only terms present in the lexicons, but also other terms which are Noun Phrases (as identified by our POS tagger and Noun Phrase recogniser) and of which a part of the Noun Phrase matches a head or modifier word in a list. If we find the head of a known term from our lexicon (e.g. “flooding”), we check for additional modifiers in the text, e.g. “snowmelt”, to give the term “snowmelt flooding”. If we find a known modifier from our lexicon (e.g. “environmental”), we check the text for possible heads that it could modify (e.g. “environmental damage”). If either of these matches occur, we create a match on the whole term (head + modifier). This means that we do not have to pre-specify every possible term in the text in advance, as this matching can be done on the fly.

Each term matched gets allocated the same URI as the related term in the lexicon; this latter is also assigned as the value of “prefLabel” as with the other term variants. This enables the possibility of later grouping together variations of terms, which might be useful for some specific analysis (e.g. tweets about a particular aspect of climate change, or for helping to assess the engagement level of the user (see WP4). This grouping was used precisely in the case study about Earth Hour, where we collected tweets during the campaign and then provided analysis about how people were tweeting (e.g. more frequent terms used in the tweets). This analysis will be described in a forthcoming deliverable in WP4. Figure 1 shows an example of a matched term variant, where the term “smart meter” has been matched in the text and annotated with

the prefLabel (preferred Reegle term) “smart metering devices”, along with the URI of the English term in Reegle.

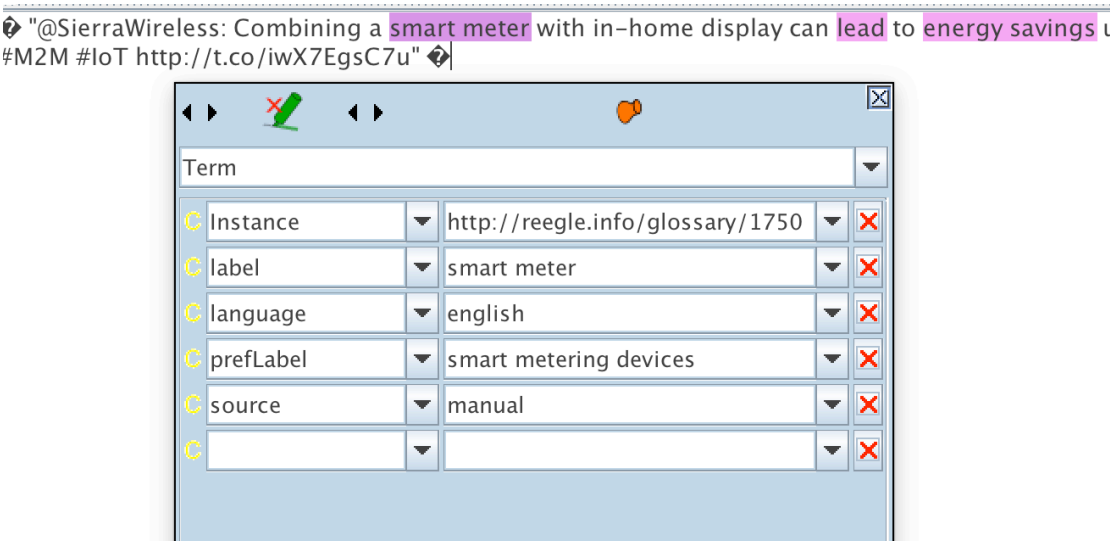

Figure 1: Annotated English term variant

One caveat about this module is that while it improves the recall considerably, it does have a slight tendency to over-generate. For example, “other climate” gets wrongly recognised as a variant of “climate”. Another cause of problems is when the incorrect part-of-speech has been identified, causing a verb such as “lead” to be identified as a noun and thus creating spurious terms. Some of these spurious terms can be prevented from matching by adding a larger stop list of words which should not be used as modifiers (such as “other”); however, it is difficult to make hard and fast rules about such things (for example, colours can sometimes be an integral part of a term and sometimes just a descriptive adjective). Some excellent new terms are found by these rules, however, such as “snowmelt flooding” as an extension of “flooding”, and “mean annual rainfall” as an extension of “rainfall”. The rules can be tweaked a little to alter the tradeoff between precision and recall; furthermore, if the module is found to overgenerate too dramatically, it can simply be switched off without affecting the rest of the term extraction process. Errors may also be propagated by incorrect syntactic analysis at the linguistic pre-processing stage, something which is a particular hazard when working with informal text such as tweets [Maynard 2014].

2.1.4. Using word embeddings to enrich the lexicons

We experimented with Word2Vec [Mikholov 2013] to see if we could extend the climate term lexicons with some new terms, based on distribution in a large corpus of environmental tweets. Word2Vec is a set of models providing word embeddings: basically an efficient implementation of the continuous bag-of-words and skip-gram architectures for computing vector representations of words. These models are shallow, two-layer neural networks, that are trained to reconstruct linguistic contexts of words, based on the idea of the *distributional hypothesis* [Harris 1954]. Essentially, the idea is that semantically or syntactically similar words have similar contexts, e.g. the same adjective might precede different terms. Once trained, the models can be used to map each word to a vector of typically several hundred elements, which

represent that word's relationship with other words. For example, one can find the most similar words to a given term, or the most similar sentence to a given sentence.

We first ran a small experiment with word2vec using the Earth Hour 2014 corpus as a training set. This is a set of 10,420 tweets (about 876K words) which were collected during the Earth Hour 2014 campaign and which contain various keywords and hashtags such as #eh2014. After removing the terms which are not nouns, verbs or adjectives (many, towards, since etc.), we end up with a fairly relevant set of similar terms, shown in Table 1 along with their scores (the closer the score is to 1, the higher the similarity). However, many of these terms are already present in our ontologies, or are related somehow to climate change but not directly useful to incorporate into our lexicons (for example, "awareness" is useful when specifically related to a tweet about climate change, but is not useful on its own as a term).

Term	Similarity
impact	0.9163834452629089
future	0.876690149307251
environment	0.8671588897705078)
sustainable	0.8671308159828186
issues	0.8636510372161865
energy consumption	0.8542860746383667
awareness	0.8514355421066284
environmental	0.8428712487220764
climate	0.8401201963424683
change	0.834912896156311
natural	0.8348809480667114
action	0.8322836756706238
reduce	0.8319122195243835
conservation	0.831143856048584
threat	0.8281859159469604
living	0.8184852600097656

Table 1: Top 15 most similar English terms to "climate change"

We therefore retrained on a larger corpus containing tweets from Earth Hour 2014, Earth Hour 2015 and Earth Hour 2016, comprising approximately 210k tweets and about 6 million words. From this, we collected the top 100 most similar terms for a seed list of 20 highly relevant environmental terms, e.g. "environment", "climate", "conservation" etc. and then added any of these new terms that were noun phrases to our lists (because we do not want to include verbs, adjectives etc. as terms) if they did not already exist. These new terms were manually verified to ensure their relevance, giving us an extra 37 terms. However, these 37 extra terms also can also be used for the on-the-fly variant generation, if e.g. longer versions of them are found in the text being processed, so the number of new terms that can be found as a result is actually much greater.

2.1.5. Evaluation

In D2.2.1 we performed an evaluation of the term recognition on 3 different data sets which had been manually annotated: climate corpus, energy corpus and fracking corpus. These 3 gold standard datasets have been made available publicly.[3] We have re-evaluated the improved recognition, and show the results below in Tables 2-4. Numbers in bold indicate improved results: we can see that Recall and F1-measure are significantly improved on all corpora, and for the energy and fracking corpora, Precision is also improved. There are also some small improvements in the figures as a result of amending the gold standard corpus to include multiple annotators rather than single annotation, which means that some errors were corrected (for example, some nested terms were annotated, rather than just annotating the longer (containing) term in this case). We should note also that all three corpora do contain some duplicate or near-duplicate content due to retweets (sometimes the text in the tweet is slightly altered in the retweet, e.g. new hashtags are added). This accounts partly for the difference between the corpora, since some terms typically occur many times. The climate corpus is the least diverse of the three, while the fracking corpus is the most diverse, in terms of content. This is due to the terms which were used for initial collection (the fracking corpus contains tweets about fracking, drilling and the Arctic, while the energy corpus only contains tweets about energy, and the climate corpus only contains tweets that mention climate change explicitly). The F1 scores reflect this diversity: the more diverse the content, the less good the tools are at recognizing the terms.

Climate Corpus	P	R	F1
ClimaTerm v1	**85.87**	53.05	65.58
ClimaTerm v2	81.49	**82.82**	**82.15**

Table 2: Comparison of ClimaTerm versions on the climate corpus

Energy Corpus	P	R	F1
ClimaTerm v1	80.94	36.42	50.23

[3] http://gate.ac.uk/projects/decarbonet

ClimaTerm v2	**88.42**	**70.35**	**78.36**

Table 3: Comparison of ClimaTerm versions on the energy corpus

Fracking Corpus	P	R	F1
ClimaTerm v1	77.64	53.55	63.38
ClimaTerm v2 (corrected)	**79.07**	**67.22**	**72.66**

Table 4: Comparison of ClimaTerm versions on the fracking corpus

2.2. Extracting German terms

The German version of ClimaTerm matches terms in the text with those found in the Gernan versions of REEGLE and DBpedia (we could not access a German version of Gemet). As with the English version, a feature is added giving information about the URI from the ontology, where relevant. The URI is for the English version of the term, because German terms are encoded in the ontology as a variant (alternative linguistic representation) of the preferred English term. Figure 2 shows part of the relevant listing in REEGLE of the term "anthropogenic climate change" with the alternative labels in other languages (German, Spanish, Portuguese and French). The lexicons contain 1795 terms from Reegle, 707 term variants from Reegle, and 38 terms from Dbpedia.

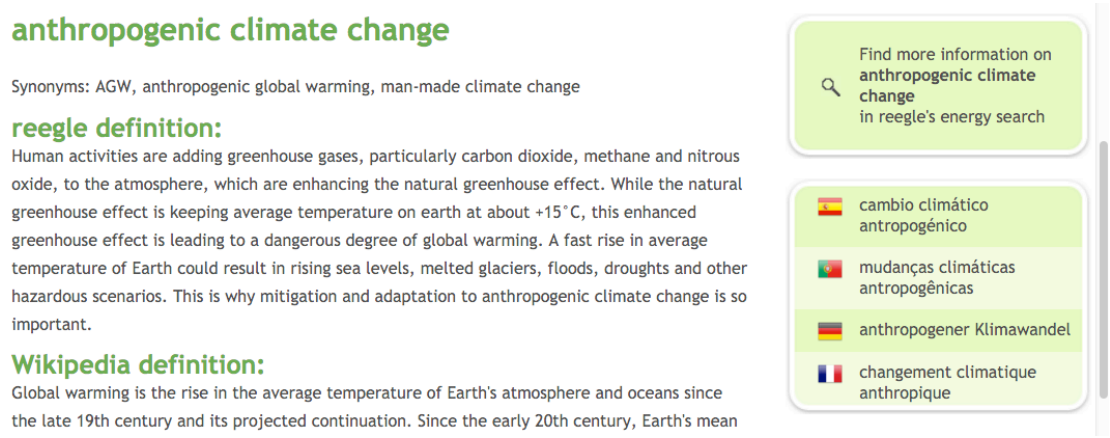

Figure 2: Excerpt from REEGLE showing labels in different languages

The application for annotating German terms is based on a slightly simplified version of the English ClimaTerm tool. It uses the same universal tokeniser and sentence splitter as for the English version, but a German-specific POS tagger based on training models from Stanford CoreNLP[4]. The recognition grammars are also adapted slightly

[4]http://stanfordnlp.github.io/CoreNLP/

from the English ones, due to the use of a different tagset for the German parts-of-speech (TIGER as opposed to the Penn TreeBank[5]) and due to some differences in the way German terms may be formed. Some experiments were carried out with establishing the best form of gazetteer matching, as described below in Section 2.2.2 (case-sensitive or not, approximate string matching, and so on).

2.2.1. Expanding the German lexicon

Since the German versions of the ontologies are a bit sparse (for example, DBpedia has far fewer German terms than English ones for the same categories), we also translated the English terms from DBpedia, Gemet and the manually created lists to German and added these, keeping the URI for the English terms. On a test document from Wikipedia about climate change, the original number of terms found before expansion with translated lists was 42. This increased to 111 after list expansion.

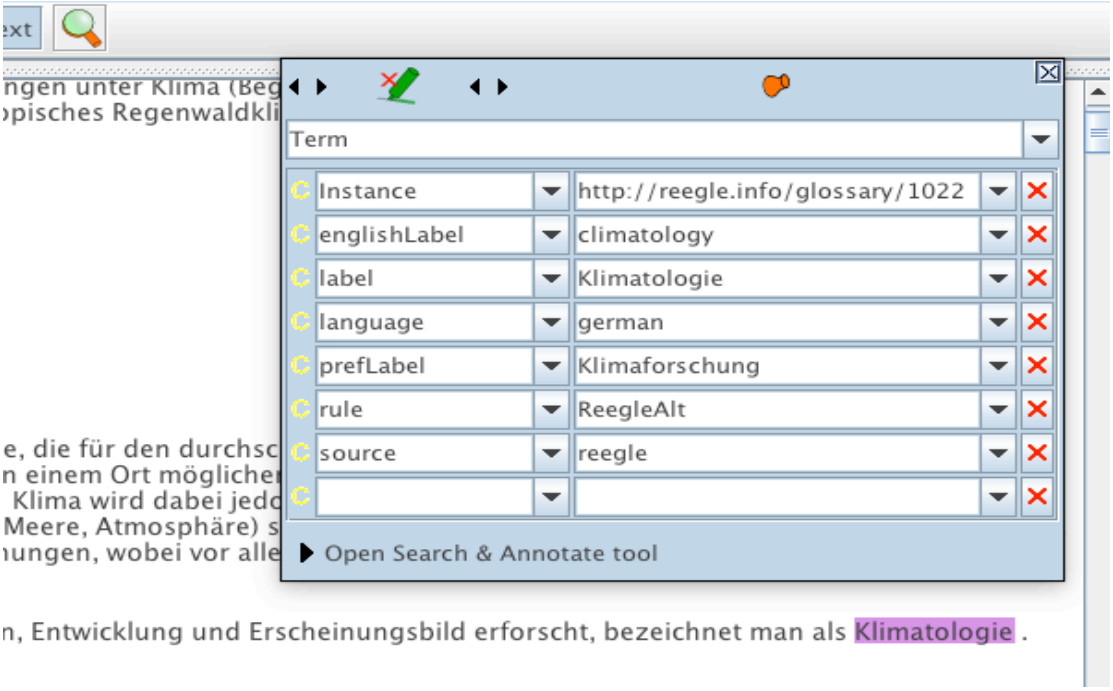

Figure 3: Figure 3 shows the matched German term "Klimatologie" in the text, corresponding to the preferred German term "Klimaforschung" and the English term "climatology".

We also experimented with using word2vec for German, using a corpus of approximately 170,000 tweets about climate extracted from the MWCC[6]. The top 15 terms containing the word “Klima” (climate) are shown in Table 5. One outlier is the term “klima navi” which refers to a make of car -- this error is most likely due to the way in which the tweets were collected (no term disambiguation takes place), and thus these tweets about the Klima Navi are irrelevant. This is one of the hazards of data collection -- we cannot always ensure that only relevant tweets are collected

[5] http://www.ims.uni-stuttgart.de/forschung/ressourcen/lexika/TagSets/stts-table.html

[6] http://www.ecoresearch.net/climate/

when terms are ambiguous (in English we also accidentally collected tweets about the band “Arctic Monkeys” until we blocked this term from occurring).

German Term	English Translation
klimakiller	climate killer
klimaschwindel	climate swindle
klimafreundlich	climate-friendly
klima navi	[a make of car]
klimakonferenz	climate conference
historisches klimaversprechen	historical climate promise
klimawandels	climate change
klimaretter	climate saviour
klimawandel stoppen	stop climate change
klimaziel	climate target
klimaanlagen	air conditioning
klimaanlage	air conditioning
klima netz	climate network
klimaschutz	climate protection
automatik klima	automatic climate

Table 5: Top 15 German terms containing “Klima”

Table 6 shows the top 15 terms most similar to the German term “Klimawandel” (climate change). As with the English terms, we can see that these terms are mostly very relevant to the topic, although some of them are only relevant when the context is defined (warming, changes, etc).

Using the trained model, we collected the top 100 most similar terms for a seed list of 20 highly relevant German environmental terms, and then added any new terms comprising noun phrases to our lists. These new terms were manually verified to ensure their relevance. The effect of adding these new terms is shown in Table 8.

German term	English translaton
erwärmung	warming

hitzewelle	heatwave
gletscher	glacier
ipcc	IPCC
globale erwärmung	global warming
menschheit	mankind
hitzewellen	heatwaves
klima	climate
antarktis	Antartica
global	global
flucht	escape
klimakatastrophe	climate disaster
cop21	cop21
cop20	cop20
änderungen	changes

Table 6: Top 15 most similar words to Klimawandel (climate change)

2.2.2. Evaluation

Our initial evaluation consisted of taking a set of 500 climate-related tweets extracted from the MWCC and which were in German, and running the German version of ClimaTerm over them. They were then evaluated manually for accuracy. Out of 155 terms, 140 were correct (in the context of climate-related tweets), i.e. scoring a Precision of 90.32%. The errors were typically due to some abbreviations in the translated lexical entries, and to some terms which were not deemed relevant (e.g. “bus”, “risk management”), even though they were present in REEGLE or Gemet. The lexicons have been manually post-edited to remove such terms.

We carried out a second, more precise, evaluation on a set of 200 tweets randomly extracted from a larger set of tweets about climate change acquired from the MWCC, mentioned above. These 200 tweets were manually annotated by an environmental expert who was also a native German speaker. A CrowdFlower annotation task was set up using the Crowdsourcing plugin for GATE [Bontcheva 2014], where the task was to annotate any mention of a term in each tweet. The results from the automated German system were compared with the gold standard annotations, and are shown in Table 7. ClimaTerm v1 shows the results for the basic German system; while ClimaTerm v2 shows the results after the addition of the results from Word2Vec. As expected, we can see that the Precision of the basic system (v1) is good but Recall is

rather lacking. In the improved version, Precision drops slightly but this is more than compensated for by the huge increase in Recall, as shown by the improvement in F-measure.

Climate Tweets	P	R	F1
ClimaTerm v1	**82.76**	32.43	46.60
ClimaTerm v2	77.17	**77.17**	**77.17**

Table 7: Evaluation of German ClimaTerm versions 1 and 2

We also experimented with various different forms of gazetteer matching, such as one based on the Levenshtein edit distance [Levenshtein 1966], in order to deal with possible spelling and grammatical variants. However, this was largely unsuccessful, due to the fact that it over-generated too many unconnected terms with similar spellings.

Making the term lists case-insensitive brought some improvements over case-sensitive lists. However, retaining only terms for which our environmental expert was highly confident about, damaged Recall considerably. The effects of the different variations can be seen in Table 8: the best results came from adding Word2Vec terms and the variants (using a German version of the on-the-fly methodology described for English), and making the gazetteers case-insensitive.

Climate Tweets	P	R	F1
Case sensitive	**80.83**	35.41	48.99
Case-insensitive	74.34	40.94	52.80
Levenshtein	61.54	28.99	39.41
Case-insensitive+Word2Vec	73.08	41.30	52.78
Case-insensitive+Word2Vec (high confidence)	73.30	39.86	51.63
Case-insensitive+Word2Vec+variants	77.17	**77.17**	**77.17**

Table 8: Effect of different gazetteers on evaluation

2.3. Software availability

As with version 1, demos of the term extraction applications (one for English and one for German) are available at http://services.gate.ac.uk/decarbonet/.

In addition, a new web service has been made available via the GATE Cloud platform (https://cloud.gate.ac.uk). This service provides both a simple REST API which accepts an HTTPS POST request containing a document and returns standoff annotations with term and URI information in JSON or XML format, and a batch

processing system that can run the same applications over large batches of data including collections of tweets in the standard Twitter streaming format.

The web service also includes the extraction of linguistic events described in Section 3, and the opinions and emotions described in D2.3.1. An updated version will be made available towards the end of the project reflecting the improvements to the opinion mining, which will be part of D2.3.2

3. Extraction of linguistic events

This task involves the extraction of linguistic events. As mentioned in the Introduction, this work is required particularly in WP4 for the case studies about the Earth Hour and COP21 events: (1) to provide linguistic criteria for the automatic identification of different stages of behaviour towards climate change based on the users' social media contributions; and (2) in experiments to automatically categorise users into behavioural stages using the analysis produced by the tools. D4.2.1 explains in more detail how various linguistic modalities can be correlated with the behaviour cycle stages. For example, deliberative questions (e.g. "*should we turn off the lights?*") are strongly associated with stage 1 (Desirable), while conditionals (e.g. "*If you turn off the lights, you will save energy*") are often linked with stage 2 (Enable Context), and imperatives and jussives (e.g. "*Turn off your lights!*") with stage 3 (Can do).

To this end, we have developed a tool which annotates text with features describing climate-related terms, sentiment, and linguistic events (modalities). The first version of this tool was described in D4.2.1. Here, we describe the extraction of linguistic modalities, which has extended the preliminary work described there. Because the tools do not easily lend themselves to be run as a web service, especially over a large dataset of csv files (the format in which the documents are extracted from the Media Watch for Climate Change (MWCC) tools), we adapted our GCP (Gate Cloud Processor) standalone processing tools in order to enable this functionality. This meant that for WP4, project partners are able to run the analysis tools from the command line over their datasets. The term recognition and sentiment analysis components are also included in this tool. More information about this is given also in D4.2.1.

3.1. Types of Linguistic Event

As described in D4.2.1, a preliminary manual analysis over a number of tweets was carried out in order to define which linguistic types might be useful to identify, and which, either individually or in correlation with other features (sentiment, emotion, presence of URLs etc) might enable correlation of tweets with the behaviour cycle stages of their authors. Furthermore, knowing whether the author of a tweet is a person or organisation is particularly useful here. We developed a separate tool to disambiguate twitter username mentions, since it is not clear whether a twitter handle refers to a person, organisation, location or even something else. This work is described in Section 4.

The kinds of linguistic events we aim to recognise can be broken down into 4 main types: questions, conditionals, directives and other (general) sentences. We identified these types based on the manual analysis of tweets and on what we believed we could

identify automatically with a good degree of competence. Each type can have several features associated with it: for example, imperatives (orders) can be positive or negative; questions can be direct or indirect; and the presence of first or second person pronouns can also be critical (compare "*I should turn off the lights*" with "*You should turn off the lights*"). These all follow standard linguistic theories. Table 4 shows the different kinds of events and features we recognise, with some examples. Note that for some events, the features are not mutually exclusive; for example, a wh-question can be either direct or indirect. Similarly, "*should we turn off the lights?*" can be categorised both as a directive question and as a deliberative directive.

We can clearly see how some of these linguistic modalities correlate with the behaviour model. Table 9 gives some examples of these correlations. The work on attributing correlations is carried out in WP4, so we do not report in detail on it here.

Behavioural Stage	**Linguistic Patterns**
Desirability	Questions (*how can I? / what should I?)*
Enabling context	Conditional sentences (*if you do [..] then [...]*)
Can do	Orders and suggestions (*I/we/you should/must*...)
Buzz	1st person + present tense *(I am doing / we are doing)*
Invitation	vocative *(Friends, guys - Join me / tell us / with me*)

Table 9: Examples of correlation between Behavioural Stage and Linguistic Patterns

3.1.1. Questions

There are two ways of asking most questions: either **directly** or **indirectly**. In general, an indirect question is considered more polite. Compare for example "*Could you turn the lights off?*" with "*I wonder if you could turn the lights off?*". Indirect questions also have a higher chance of being rhetorical, e.g. "*I wonder if it is better to turn the lights off.*"

A **directive** question, according to our categorisation, is one which phrases a directive (see Section 3.1.2) in the form of a question. Whereas a directive would be an instruction to do something, e.g. "*You should turn off the lights*", a directive question turns this order into a query about obligation: "*Should you turn off the lights?*"

An **invitation** is a more polite form of directive, where the hearer is invited to do something, e.g. "*Will you turn off the lights?*"

A **wh-question** is an information-seeking question, to which the answer should not be yes or no (unless the responder is being deliberately obtuse). It comprises one of the following question words: *who, what, why, when, how, where, which, whom, whose.*

A **general** question is simply a catch-all category for any other type of question which is not a directive or wh-question.

Type	Features	Example
Question	direct	Shall we turn off the lights?
	indirect	I wonder if you will turn off the lights.
	directive	Should we turn off the lights?
	wh-question	Why should we turn off the lights?
	invitation	Will you turn off the lights?
	general	Is it best to turn off the lights?
Directive	obligative	You must turn off the lights.
	negative obligative	You must not turn off the lights.
	imperative	Turn off the lights!
	prohibitive (negative imperative)	Do not turn off the lights!
	jussive	Go me!
	deliberative	Should we turn off the lights?
	indirect deliberative	I wonder if we should turn off the lights.
Conditionals	type 0	If you turn off the light, you save energy.
	type 1	If you turn off the light, you will save energy
	type 2	If you turned off the lights, you would save energy
	type 3	If you had turned off the lights, you would have saved energy

Table 10: Categories of Linguistic Event

3.1.2. Directives

A directive sentence is basically some kind of order or instruction. They were described by Searle [Searle, 1975] as one of five basic speech acts (along with assertives, commissives, expressives, and declaratives). According to Searle, directives are speech acts that require the hearer to take a particular action, such a requests, commands, and advice.

An **obligative** signifies necessity (in the mind of the speaker) and typically involves modal verbs such as *should* and *must*, e.g. "*You must turn off the lights*." A **negative obligative** is the same thing negated, e.g. "*You must not turn off the lights*." Note the particular case of obligative expressions such as "have to" and "ought to", which in English cannot be negated semantically by adding a negative (something which frequently trips up non-native English speakers). For example, while "*You have to turn off the lights*" and "*You must turn off the lights*" are synonymous, "*You do not have to turn off the lights*" is not synonymous with "*You must not turn off the lights*".

An **imperative** sentence comprises a verb in the imperative mood, and in English is typically signalled by a command with an implicit rather than explicit second-person subject (you), such as "*Turn off the lights*". A **prohibitive** sentence is simply a negated imperative, e.g. "*Do not turn off the lights*". A j**ussive** sentence, which is quite a rare construction, is one where the imperative has an implicit first-person subject (I/we), such as "*Go me!*".

Direct and indirect **deliberatives** have already been described above, as directive questions. Direct deliberatives are direct questions, e.g. "*Should I turn off the lights?*", while indirect deliberatives are indirect questions, e.g. "*I wonder if we should turn off the lights*". We include them here for the sake of completeness, and because they can be labelled both as kinds of directive and as kinds of question.

3.1.3. Conditionals

Conditional sentences involve real or hypothetical situations and their consequences, and incorporate verbs in the conditional mood. Complete conditional sentences contain a conditional clause and the consequence, such as "if...then" constructions.

There are 4 types of conditional construction, incorporating different verb tenses. The verb tense combination depends on whether the speaker thinks the result is real, probable, or unreal (only exists in the imagination). Table 11 shows the situations denoted by each type, and the verb tense combinations for the condition (if-clause) and result (main clause) of each. Any of the types can also involve negated verbs in either or both clauses. The square brackets in the examples help to show the timing of when the condition should take place. Note that in some European languages, the condition in a type 1 construction is denoted by a future tense (if you *will* turn off the lights), since the condition should take place in the future.

Type	Situation	Condition	Result	Example
0	present real	present	present	*If you turn off the lights, you save energy.*
1	future real	present	future	*If you turn off the lights [tonight], you will save energy*

2	present or future imaginary	simple past	present conditional	*If you turned off the lights [tonight], you would save energy*
3	past imaginary	past perfect	perfect conditional	*If you had turned off the lights [yesterday], you would have saved energy*

Table 11: Types of conditional sentence

3.1.4. Use of first/second person

Finally, we also categorise tweets according to whether they use the first or second person, which may also bear correlation with different stages of engagement (for example, encouraging others or talking about oneself). Users who post social media updates mostly relating to themselves are known as me-formers, while users who post updates which are mostly information-sharing or directed at other people are known as informers. In one study [Naaman 2010], 80% of regular Twitter users were found to be me-formers, talking about themselves and often sharing emotions. Informers who share information, on the other hand, typically have larger social networks and are more interactive with their followers -- this has important implications for engagement and behavioural studies (see WP4).

3.2. Automatic recognition of linguistic modalities

The linguistic modalities described above are all recognised and annotated (with appropriate features) automatically using GATE tools. Linguistic pre-processing tools are first run on the documents, and then a knowledge-based approach is used to categorise them.

Linguistic pre-processing consists of the standard pipeline of tokenisation, sentence splitting, and POS-tagging, plus verb phrase chunking. Verb phrase chunkers delimit verbs, which may consist of a single word such as "*bought*" or a more complex group comprising modals, infinitives and so on (for example "*might have bought*". They may even include negative elements such as "*might not have bought"* or "*didn't buy*".

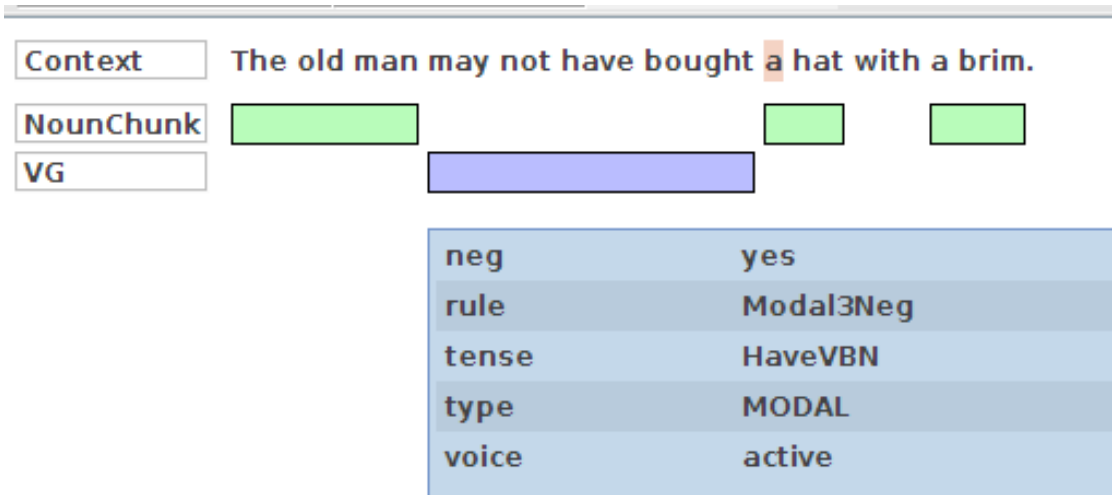

Figure 4: Annotation of Noun Phrase and Verb Phrase chunks in GATE

In this task, we use the GATE Verb Phrase chunker which is written in JAPE, GATE's rule-writing language, and is based on grammar rules for English [Cobuild 1999] [Azar 1989]. It contains rules for the identification of non-recursive verb groups, covering finite ("*is investigating"*), non-finite ("*to investigate*"), participles ("*investigated*"), and special verb constructs ("*is going to investigate"*). All the forms may include adverbials and negatives. One advantage of this tool is that it explicitly marks negation in verbs ("*don't")*. The verb chunker is critical for this task, which requires the identification of modal verbs ("*can*", "*could*", "*might*"), negative forms of verbs ("*can't*", "*shouldn't*"), verb tenses (future, past, conditional etc) and moods (active, passive, subjunctive etc). The rules make use of POS tags as well as some specific strings (e.g. words such as "*might*" are used to identify modals). An example of a sentence divided into noun phrase (denoted by NounChunk and with green highlighting) and verb phrase chunks (denoted by VG and with blue highlightling) is shown in Figure 4. The blue box shows also the features generated by the verb phrase chunker: negation, tense, type and voice (plus the rule used to generate the annotation, simply for debugging purposes).

The various linguistic events (conditionals, questions and directives) are then annotated using a set of hand-crafted grammar rules based on the POS tags, verb chunks, and presence of certain groups of words (e.g. verbs of commanding). For example, a rule to find an indirect deliberative (e.g. "*I don't know if I should turn off the light")* involves finding two clauses in a sentence, where the second one must be an "if" clause in a conditional, followed by a verb of commanding ("*should*" and then another action ("*turn off the light*"). These rules are deliberately quite underspecified for maximum recall: in this case we do not care what the first clause contains, as long as it does not already match a different rule in our set. If necessary, the rules can later be made more specific if we find they overgenerate.

In general, the rules perform well, especially conditionals and questions. However, the rules are generally only as good as the linguistic pre-processing components on which they rely, and these can fail at times, especially on tweets which do not contain perfect grammar or spelling. In particular, imperatives are not always correctly recognised by our tools, due to some inaccuracies of the Verb Phrase Chunker caused by ambiguity. Future work will look at a fuller evaluation and on tightening up the accuracy.

3.3. Summary

In summary, the task of linguistic event recognition has focused on the identification of various linguistic types relevant for tasks such as mapping users to their level of engagement according to the various theories proposed in the project. The tools we have developed provide linguistic criteria which help to identify the different stages of behaviour towards climate change, as used in WP4 for the case studies about the Earth Hour and COP21 events. They have also been used in experiments to automatically categorise users into behavioural stages. A number of linguistic types have been identified, and results of the experimental work in WP4 using these are promising.

4. Actor recognition in tweets

The subtask of actor recognition in tweets comprises the identification and characterisation of people tweeting, i.e. twitter usernames. We call this twitter username mention classification / disambiguation. The task is important because there could be big differences between tweets from personal accounts and tweets emanating from organisations, in particular with respect to things like the identification of engagement (we would expect organisations such as WWF already to be very engaged with environmental topics, for example).

Named Entity Recognition and Classification (NERC) is the task of identifying and classifying occurrences of entities (e.g. places, people, locations) in text. NERC in tweets differs from other domains and genres in that references to entities occur not just within the tweet text (e.g. "I went to Paris with @myBestFriend"), but also in the form of user name mentions. Since such mentions are preceded by an @ symbol, they are trivial to identify. Previous NERC work has made the simplifying assumption, however, that they are also trivial to classify, as they always refer to persons [Ritter 2011, Plank 2014]. While this was true in the early days of Twitter, there are now many user accounts of organisations (@CNN), locations (@OXO_Tower), and products (@iPhone), which motivated this research in automatic @mention classification.

An additional challenge comes from the lack of suitable datasets. Current human-annotated Twitter NERC corpora either do not classify @mentions [Ritter 2011], anonymise them [Cano Basave 2013], or are small-sized and noisy [Finin 2010].

To address these issues, we have developed both an annotated corpus and a tool for @mention identification. The corpus introduces a sizeable, publicly available[7] new dataset of crowd-classified username mentions in tweets, while the tool presents a learning-based approach that automatically classifies @mentions into one of three types (person, location, or organisation).

While the task of username mention classification is similar to the standard NERC task, @mentions differ from other occurrences of named entities in tweets, because they are monosemous, i.e. a given mention always links to the same user profile. These user profiles provide an additional rich context (complementary to tweet texts), which can help with @mention classification. To the best of our knowledge, however,

[7] https://gate.ac.uk/projects/decarbonet/

existing Twitter NERC methods have ignored this information, despite it being present in the JSON of each tweet.

Our experiments also demonstrate that state-of-the-art NERC methods do not perform well on @mention classification, since @mentions do not contain white spaces delimiting token boundaries, and tend to be used socially to tag and direct messages in a way which often does not conform to conventional syntactic and grammatical patterns.

We have therefore developed a dedicated Random Forest [Breiman 2001] @mention classification approach, which utilises a wide range of user profile features. It outperforms significantly a number of state-of-the-art Twitter and general purpose NERC systems on the @mention classification task. Our experimental results also show that although the context surrounding the @mention helps with its type classification, user profile information is even more essential.

4.1. Related Work

Ritter et al. [Ritter 2011] take a pipeline approach performing first tokenisation and POS tagging before using topic models to find named entities. Liu [Liu 2011] propose a gradient-descent graph-based method for doing joint text normalisation and recognition. [Plank 2014] investigate how distant supervision improves Twitter NER performance. In particular, they project reliable NER tags from web pages onto tweets that contain links to those pages, in order to create additional training data. Brown clusters and word vectors have also been shown to improve Twitter NER performance [Cherry 2015]. Our experiments also use a Twitter-adapted version of the state-of-the-art Stanford CRF-based NERC system [Finkel 2005], which we trained on our new @mention classification dataset. All these approaches, however, have failed to address @mention classification in a principled way.

With respect to training and evaluation datasets, existing Twitter NERC corpora are not well suited for @mention classification. The widely used Ritter corpus [Ritter 2011] considers @mentions trivial to annotate, since they are always denoted as people. Likewise, the UMBC NE annotated tweets [Finin 2010] contain very few annotated username mentions, all of which are marked as person. The same is true of the expert reannotated version of this dataset [Fromreide 2014]. Other Twitter NERC datasets have anonymised all @mentions by replacing them with @USER.

4.2. The @Mention Dataset

Our aim was to classify the mentioned usernames from a collection of 3000 tweets as belonging to either person, location, organisation or other. Though users could be classified without a tweet or mention, classifying @mentions within naturally occurring tweet texts, rather than user profiles in isolation, ensures that our work stays broadly compatible with the Twitter NERC task and associated data sets. After removing spam and retweets, we were left with 2,274 complete tweets in our data. We expanded our corpus by also including the authors of these posts, for each of whom we retrieved one mention via the Twitter search API. This process added a further 659 tweets, though we could not retrieve a suitable tweet for many of the authors.

In total, 3,141 usernames and their associated tweets were collected and put forward for classification via crowdsourcing. The task guidelines and interface design were

piloted first, to ensure clarity and ease of use, and that guidelines reflected difficult cases through examples. The crowdworkers were shown detailed user profile information, including the Twitter username, full name, profile text and profile picture. The classification classes were 'Person', 'Location', 'Organisation', 'Other or spam' and 'Unknown'.

The 3,141 usernames were classified by 3 crowdworkers each, with a Kappa score of 0.5138. The judgements were collected from reliable crowdworkers from native English speaking countries, recruited via the CrowdFlower crowdsourcing platform[8]. After majority vote adjudication, there were 190 usernames on which the respective 3 crowdworkers disagreed with one another.

These 190 @mentions were then re-annotated independently by three researchers each and again adjudicated based on majority. After this second iteration, there were still 41 usernames with no agreement, so they were excluded from the gold standard. All usernames annotated as unknown were also excluded from the final dataset. The dropped users were accounts with very limited profile data, profiles which were not in English, and spam accounts.

Class	Frequency	Proportion
Person	2180	70.32%
Organisation	704	22.71%
Unknown	123	3.97%
Other	73	2.35%
Location	20	0.65%

Table 12: Frequency of entity types in the initial crowdsourced data

The distribution of entity types in the resulting dataset is shown in Table 12. Unfortunately, there were only 20 Twitter accounts classified as locations. To balance the dataset, additional location accounts were added by crawling the FourSquare social network[9] for venues which have Twitter accounts, and assuming they were all locations. A total of 796 additional location usernames were added to the corpus automatically, as well as 796 corresponding tweets which mention these (one tweet per location username).

Table 13 shows the entity class statistics in the final dataset, after splitting the corpus into fixed training, development and testing portions.

Data set	Organisation	Location	Person
Training	487	608	1382
Development	155	102	444
Testing	145	106	460

Table 13: Frequency of entity types in the final gold standard data.

In order to gain additional insight into the dataset, the distributions of follower and friend counts for the authors were computed (see Figure 5). Both of these counts are

[8] http://www.crowdflower.com/

[9] https://foursquare.com/

normally distributed when log scaled, with a large range included in the data. The mean count of followers for the person class was very high (1,363,904), but also had a very high standard deviation (σ^2 = 7165803), compared to organisations, which had a lower mean but also lower variance in number of followers (938,370, σ^2 = 2931483). Although there are some celebrity Twitter users in our data, many of the very popular accounts belonged to organisations, such as news outlets.

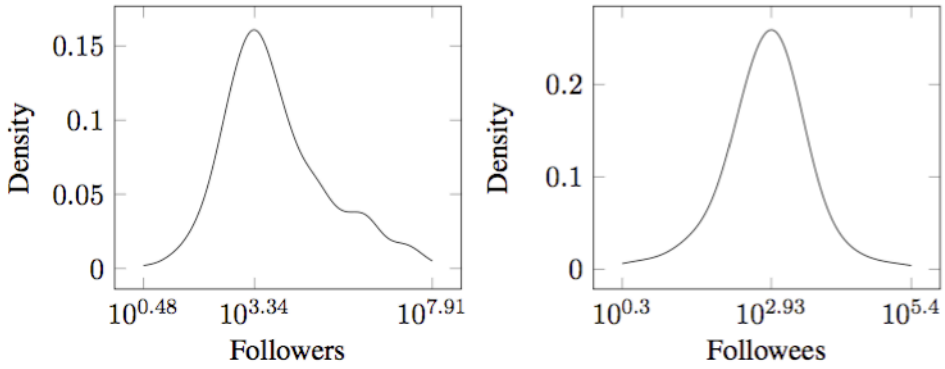

Figure 5: Distribution of followers and followees of authors in the development set

We also investigated the kinds of Twitter users included in the dataset, based on the most frequent terms in the user descriptions, excluding stopwords other than pronouns. As can be seen in Table 13, user profiles belonging to people tend to mention occupations, such as writers, company directors and media employees. Some of the frequent terms also suggest non-professional Twitter users, such as 'love', 'life' and 'fan'. The most frequent terms from locations indicate hospitality venues ('beer', 'restaurant', 'bar', 'food') and music venues ('music', 'live'). The terms for organisations are indicative of media ('breaking', 'comment', 'features', 'bbc', 'news'), but also contain often the word 'official'.

4.3. Methods

This section describes the features, baselines, and machine learning methods used in the user mention classification experiments. The software, gazetteers and models required to reproduce this work are available through GitHub[10].

4.3.1. Baselines

The first baseline is *majority class*, which classifies every @mention as a person, following assumptions made in prior Twitter NERC research [Ritter 2011].

The second set of baselines comprises Naive Bayes classifiers: one (NB profile) uses as features the unigrams of the user's profile text and the second one (NB context) uses a 3-token context window surrounding the @mention in the tweet text. We also tested the Naive Bayes classifier with the full set of features used by the Random Forest classifier (see Section 4.3.2).

[10]https://github.com/GateNLP

Person	Location	Organisation
author	we	news
journalist	food	twitter
love	restaurant	official
all	beer	independent
editor	bar	features
fan	great	support
life	located	we
media	wine	follow
social	de	appeal
writer	el	comment
own	en	christmas
views	offer	all
live	live	account
now	music	bbc
world	downtown	uk
out	venue	more
member	american	world
about	place	every
director	delicious	breaking
also	menu	tweets

Table 14: Most frequent description terms by class.

The third baseline was Ritter et al's Twitter NERC system. It was trained and evaluated on a dataset which deliberately excluded mentions, but in this paper we apply it to @mention classification, using the tweet text as a context. The system uses more entity types than the person, location and organisation classes used here, so we applied the mapping introduced by [Derczynski 2015]. The publicly available, pre-trained Twitter NERC models were used[11].

The final baseline is the general purpose Stanford CRF-based NER system [Finkel 2005], which is not specifically designed for Twitter NERC, but is easily retrainable to such datasets. Therefore, for fair comparison we trained a model on the training part of our corpus, instead of using the general purpose news-trained NERC model.

The latter two state-of-the-art approaches were chosen because they are publicly available; have been shown to perform well on tweets [Derczynski 2015], and could easily be applied to @mention classification. Other Twitter NER systems, such as TwitIE [Bontcheva 2013], were excluded as they could not be adapted easily to this classification task. We did use TwitIE, however, for feature generation (see Section 4.3.3).

[11] https://github.com/aritter/twitter_nlp

4.3.2. The Random Forest Mention Classifier

In addition to the Naive Bayes classifiers described above, we trained a Random Forest (RF) classifier [Breiman 2001] using contextual features derived from the tweet text and metadata and text-based features, derived from the Twitter user profiles. RFs are well suited to the @mention classification task, since some of the features are not independent, and also vary in type (binary, nominal and real values). An additional motivation is that the number of training examples is relatively small, so we need a classifier that can learn over a complex decision space with limited overfitting.

4.3.3. Features

The following features were introduced based on the user profile information and incorporated as part of the Random Forest learning algorithm.

Account Metadata Features

The following numeric and binary features are based on user account metadata:

- Twitter verified account (true/false) (verified);
- average number of posts per day since account creation (posts-per-day);
- total number of posts (total-posts);
- number of followers of this account (followers);
- number of followees (i.e. accounts followed by the given user, followees).

These metadata features attempt to capture the popularity and activity level of the given user. The inclusion of both follower and followee counts was motivated by prior work, which observed that a high follower count coupled with a low followee count indicates that the account is using Twitter as a broadcast medium, rather than for social purposes; in other words this is more indicative of organisations and locations. Such accounts also tend to produce a high number of posts-per-day.

Display Name Features

Edit Distance with Username: The Levenshtein edit distance [Levenshtein 1966] between the user's display name (e.g. Donald J. Trump) and their username (e.g. @realDonaldTrump) is used as a numeric feature (name-edit-distance). It reflects the observation that organisations and locations tend to choose similar usernames and full names, with more variation for people.

Entity Type Features: The display name (e.g. Donald J. Trump) is analysed with the TwitIE NERC system and three numeric features are generated, indicating whether the display name is categorised as a person, location, or organisation respectively (e.g. twitie-org-fullname).

Features based on User Profile Text

The user profile text is used as the basis for generating numerous features, as detailed next.

Dictionary-based Profile Text Features: As seen in Table 13, occupations and job titles are some of the most indicative words in user profile texts of people and can thus serve as a high precision indicator for the person class. We use a list of job titles and occupations, which we merged from the WordNet-based ones in the TwitIE NERC system and those in the Wikipedia list of occupations[12].

Similarly, singular and plural first person pronouns, and unigrams and bigrams from disclaimer texts (e.g. personal, my own, views, representative, employer, official) feature prominently. For organisations, important indicators are words such as club, society, school. We also looked up names of organisations, persons, and locations against the extensive lists of the TwitIE rule-based entity recogniser. Two count sets of these features are produced – one from the text of the user description (lookup-per-desc) and the other from the profile full name (ie lookup-per-fullname).

Lastly, since user accounts belonging to organisations and locations tweet in an official capacity, they tend to use more syntactically and grammatically formal language, unlike personal users who often use abbreviations, informal terms, and misspelled words. Since parsing tweets is error prone and slow, we used the count and proportion of out-of-vocabulary words[13] within the user profile text as a proxy, giving the features oov-count and oov-ratio.

DBpedia-based Candidate Class Features: DBpedia [Bizer 2009] is a large, open-domain database of entities and terms (all uniquely identified via URIs), which was created automatically from the infoboxes and other structured Wikipedia data. Since many celebrities, notable people, organisations, and locations tend to have both Wikipedia pages and Twitter accounts, we use a high-precision, low recall DBpedia lookup, to identify a potential matching URI for the given Twitter username.

In more detail, the user's self-declared full name from their Twitter profile is looked up against the string properties of DBpedia URIs (namely, dbpedia:name, rdf:label, and foaf:name) and all candidate URIs are collected for subsequent disambiguation. In addition, we also match the user's website URL in their Twitter profile (if available), against the website URLs present in DBpedia (i.e. against the values of dbpedia:url, dbpedia:website, and foaf:homepage properties).

If at the end of this matching process there is a single matched URI, that entity and its type[14] are selected. Where multiple results are found, the DBpedia abstracts are retrieved and the token overlap with the user's Twitter profile text is used to determine which is the best matching candidate DBpedia URI.

Once a matching DBpedia candidate is identified, its class (the value of its rdf:type property) is mapped to either location, person, organisation, or other, based on the

[12]https://en.wikipedia.org/wiki/Category:Lists_of_occupations

[13]Measured against the JASpell dictionary: http://jaspell.sourceforge.net/

[14]In DBpedia each entity with an URI is an instance of a class and three of the top level ones are direct equivalents to the person, location, and organisation types we use for @mention classification.

DBpedia class ontology. If no matching DBpedia candidate is found (particularly common for personal accounts), then the target type is considered null. We use this DBpedia candidate class information as features (e.g. dbpedia-per).

WordNet distance scores: The Lin semantic similarity score [Lin 1998] is calculated between each term in the user profile and the WordNet senses "organization#n#1", "location#n#1", and "person#n#1". For each user, the maximum, mean and sum over all terms in the description, compared to the target set is calculated. Information content was calculated for this metric using our own training data. Comparing the terms in the description and name semantically with the appropriate root terms using WordNet allows for matching of terms that are not in the gazetteer.

Entity Type Features: Similar to the display name, the text of the user profile is analysed with the TwitIE NERC system, and three numeric features are generated, one each for the number of persons, locations, and organisations mentioned within the profile text (e.g. twitie-org-desc).

4.3.4. Tweet-based Features

The text of the tweet within which the @mention appears is used to derive a number of **contextual features**. The most traditional ones are the unigrams of the three preceding and following tokens, which are used by the baselines (See Section 4.3.1).

The Random Forest classifier uses instead a smaller set of positional features, derived from the tweet text. Firstly, we added a feature whether the @mention appears at the start of the tweet (context-start). Four other binary features reflect the presence of the following high-frequency preceding tokens: *RT*, *to*, *with*, and *at*. These were derived from the training part of our corpus, based on frequency counts.

4.4. Results

Evaluation results are reported on the held out development[15] and testing subsets, shown in Tables 15 and 16. Precision, recall and F1 are calculated using the metrics provided in the Python SciKitLearn package[16]. Accuracy is calculated by micro-averaging across all samples in the evaluation set. Where Random Forests were used, the number of trees was fixed at 1000, allowing good generalisability. Using greater numbers of trees did not yield performace improvement.

As can be seen in Table 14, the retrained Stanford CRF performs best amongst all baselines. It also outperforms the Random Forest mention classifier when it uses only a subset of features.

Method	P	R	F	P	R	F	P	R	F	Accuracy
Always Person	0.63	**1.00**	0.78	0.00	0.00	0.00	0.00	0.00	0.00	0.63
Stanford CRF	0.83	0.89	0.85	0.68	0.48	0.56	0.71	0.34	0.46	0.75
Ritter	0.67	0.31	0.42	0.23	0.18	0.20	0.32	0.15	0.20	0.34

[15]We did not use the development portion for parameter tuning or training.

[16]http://scikit-learn.org/

Bayes (Profile)	0.87	0.65	0.74	0.56	0.65	0.61	0.41	0.66	0.50	0.65
Bayes (Context)	0.77	0.16	0.27	0.29	0.33	0.31	0.26	**0.82**	0.39	0.33
Bayes (Features)	**0.92**	0.46	0.61	0.25	**0.92**	0.39	0.72	0.46	0.56	0.53
RF (Metrics Features)	0.81	0.86	0.84	0.52	0.67	0.59	0.71	0.45	0.55	0.74
RF (Dictionaries)	0.73	0.98	0.84	0.42	0.10	0.16	0.75	0.39	0.52	0.72
RF (TwitIE Features)	0.68	0.94	0.79	0.40	0.19	0.26	0.54	0.14	0.22	0.65
RF (WordNet)	0.73	0.88	0.80	0.40	0.28	0.33	0.60	0.35	0.45	0.68
RF (Mention Context)	0.68	0.99	0.80	**0.91**	0.10	0.18	0.68	0.17	0.28	0.68
All Features	0.89	0.96	**0.93**	0.67	0.73	**0.69**	**0.87**	0.63	**0.73**	**0.85**

Table 15: Results on the development dataset

The weakness of the Ritter tagger compared to Stanford CRF is due to the former being trained on a corpus where no mentions were classified by type. Similar to Stanford CRF, we expect that stronger performance would likely be achieved by retraining the Ritter models; however, the system distribution at present does not provide information on retraining. The reasons for the poor performance of the Naive Bayes classifiers on the training set results is discussed below.

Since some of the features detailed in Section 4.3.3 require additional resources (e.g. WordNet, DBpedia) and computation overhead (e.g. running TwitIE on the user profile texts), we investigated the impact on performance if only a subset is used by the Random Forest classifier. Later in this section we also report on the importance of individual features.

The last six rows of Table 15 show that there is a very significant increase in precision, recall, and accuracy scores when the complete set of features is used, which demonstrates that the Random Forest classifier is learning effectively from these diverse types of information. In addition, we observe that some feature sets demonstrate very high precision or recall for particular classes. For instance, dictionary-based features yield high precision for both person and organisation, and extremely high recall for person (0.98), but were not useful in discriminating locations. Likewise, the contextual features were useful for discovering person @mentions (0.80 F1) and yielded very high precision for location @mentions, thanks to the 'at' context term.

The strong performance of user account metadata features (e.g. followers, account age, post frequency) demonstrates that user profile metadata is beneficial. In particular, using these features on their own yields classification performance comparable to the Stanford NERC model, which is using much richer contextual features.

Table 16 shows the most important features used by the Random Forest learning model, according to the average reduction for each feature in the Gini impurity score [Breiman 2001]. The most important features in the model are also the simplest, including presence of the user's full name in a list of person names, and the volume of their posts. The semantic relatedness features were also prioritised by the model, demonstrating that the text of the description itself adds important additional information. The context-start feature is the only one that is derived from the tweet itself, though the other context features are comparatively far more sparse, and as such they cannot split the data set sufficiently to be prioritised for inclusion by the Random Forest model.

Lastly, the results on the held out test data (Table 17) demonstrate the generalisability of the Random Forest @mention classifier and the baselines. RF with all features outperforms all baselines, including the retrained Stanford CRF model. It also has the highest F1 scores for all three classes. This clear improvement over the baselines demonstrates the advantage of using user profile data, in addition to the tweet text.

Feature	**Score**
lookup-per-fullname	0.115
posts-per-day	0.100
total-posts	0.096
wordnet-org-max	0.056
wordnet-per-sum	0.055
wordnet-org-sum	0.053
wordnet-org-mean	0.050
wordnet-per-mean	0.048
wordnet-per-max	0.047
oov-count	0.045
name-edit-distance	0.045
context-start	0.038
oov-ratio	0.035
verified	0.020
lookup-org-fullname	0.018

Table 16: Feature importance scores (RF model)

Method	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**	**Accuracy**
Always Person	0.65	**1.00**	0.79	0.00	0.00	0.00	0.00	0.00	0.00	0.65
Stanford	0.79	0.81	0.80	**0.65**	0.37	0.47	**0.77**	0.30	0.44	0.70
Ritter	0.70	0.35	0.47	0.29	0.17	0.21	0.37	0.18	0.24	0.38
Bayes	0.85	0.57	0.69	0.46	0.62	0.53	0.32	0.56	0.40	0.58

(Profile)										
Bayes (Context)	0.79	0.19	0.31	0.31	0.27	0.29	0.23	**0.80**	0.36	0.33
Bayes (Features)	**0.92**	0.44	0.60	0.23	**0.88**	0.37	0.62	0.40	0.49	0.50
RF (all features)	0.88	0.95	**0.91**	0.64	0.62	**0.63**	0.75	0.57	**0.65**	**0.82**

Table 17: Results on the held-out testing data

The first two Naive Bayes models gave worse performance than the majority baseline. This is due to the the very high number of unigram features, which was higher than the number of training examples. This made the models susceptible both to noise in the training data, and to overfitting. The third Naive Bayes model (trained on the same features as the Random Forest classifier) also failed to outperform the most common class baseline. However, it achieved high precision for the person class, and high recall from the location class, suggesting that many person instances were being incorrectly classified as locations. The proportion of locations in the training set (24.5%) is higher than that in the development set (14.6%). Additionally, the data as a whole is skewed towards the person class, so classification errors could stem from the learning of incorrect prior probabilities, a problem which did not affect Random Forest learning.

	loc	org	per
loc	66	15	25
org	26	83	36
per	11	13	43

Table 18: Confusion matrix for RF (all features)

The confusion matrix (Table 18) shows the Random Forest classifier often mistook organisations for persons. Given that the data is unbalanced with person being the majority class, this is to be expected. In addition, none of the feature subsets yielded particularly high recall for organisations. Naive Bayes with context features, however, consistently achieves high recall on organisations, so future work will investigate the potential of combining the outputs of the two classifiers.

4.5. Summary

This section has focused on the problem of classifying usernames mentioned in tweets as belonging to persons, locations, organisations, or other entities – a task largely similar to named entity recognition and classification. The first research contribution of the work is in introducing a new public, crowdsourced dataset for development and evaluation of @mention classification methods. The distribution of classes in this dataset demonstrates that users cannot be trivially assumed to be all persons. The

second research contribution is the development of a tool for username identification, and a set of experiments comparing it with relevant baselines.

The best performing Random Forest model outperformed all state-of-the-art NERC baselines on the @mention classification task. It uses features derived not only from the @mention context within the tweet text, but also from the metadata and additional textual information in the Twitter profile belonging to this username. Since profile information is already included as standard within the JSON of each tweet, this does not impose additional data gathering overheads.

Through comparison to the best performing Stanford CRF model based purely on the tweet text surrounding the @mention, we have shown that effective @mention classification can be improved significantly through using the additional information from the associated user profile.

Future work will extend the entity classes into more fine grained sub-classes, such as celebrity, politician, company or NGO. We will also continue to enlarge the gold standard dataset. Additional experiments will be carried out using unsupervised learning and dimensionality reduction based on a larger, unlabelled collection of user profiles. We also plan to experiment with using also recent tweets authored by the user to improve @mention classification accuracy. However, this will come at the cost of needing to retrieve additional tweets and thus increase computation times.

5. Recognyze: a service for named entity recognition

5.1. Software availability

As described in D2.2.1, the Recognyze web service is available at http://triple-store.ai.wu.ac.at/. Given a text input, the Recognyze service returns a set of named entities, together with their start and end positions within the input text. Under the hood, Recognyze makes use of open data portals such as DBpedia and GeoNames for its queries, returning predefined subsets (property-wise) of respective entities. Note that service usage is limited to 100 requests per day (max. 1MB data transfer per request).

When querying, a search profile to search within must be provided. A search profile describes a domain from the real world; currently the following set of domains exists:

{en,de}.organization.ng

Organizations in English and German, taken from DBpedia. Returns type.

{en,de}.people.ng

Person names in English and German, taken from DBpedia. Returns type.

{en,de,fr}.geo.50000.ng

Geolocations (cities, countries) with a population larger than 50000, taken from GeoNames. Returns type.

Passing multiple profiles at once is also supported by the API.

The REST interface can easily be accessed via the open source webLyzard API at https://github.com/weblyzard/weblyzard_api. For more information on the API, please consult the documentation at http://weblyzard-api.readthedocs.org/en/latest/.

Recognyze returns a JSON list object of all entities found. For each entity found, the service returns the entity type, the associated search profile (see above), the entity's occurrences within the given text (start, end, sentence, surface form), the confidence of the correctness of the entity, the public key where the entity links to (e.g. http://sws.geonames.org/4990729), as well as extra properties where available.

5.2. Evaluation of Recognyze

Recognyze was evaluated on several corpora using an evaluation framework designed to handle most of the current evaluation formats (NIF, TAC kbp, csv, etc.) and provide data about which entities various tools are not able to extract well (missed entities, hard to disambiguate entities, etc). In addition to custom scripts and converters between various formats, it is also possible to use third-party (e.g. Gerbil) tools together with the corpora and tools included in this framework. The evaluation framework is described in (Brașoveanu et. al., 2016) and is depicted in Figure 6.

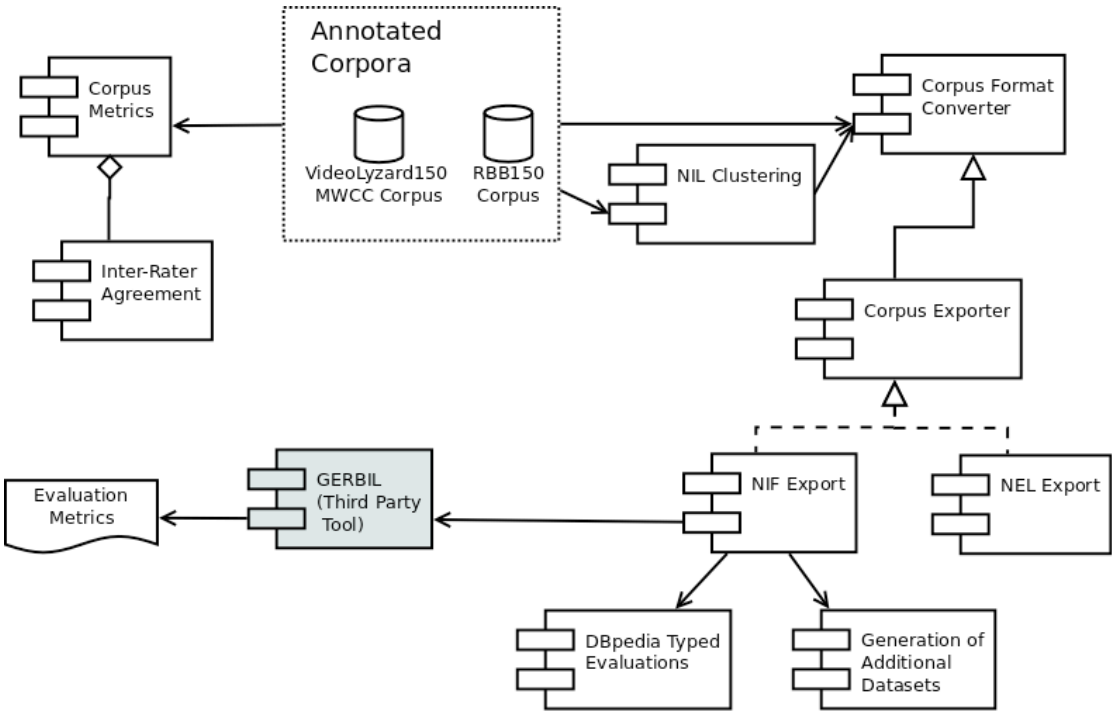

Figure 6: Evaluation framework for Recognyze. Adapted from [Brasoveanu et al., 2016].

For the current evaluation we have selected two corpora:

RBB150[17] [Brasoveanu et al., 2016] corpus contains 150 annotated texts in German from domains like climate change (floods, rising temperatures, etc) politics (local elections, anniversaries) and sports (tennis, soccer). While there are several other

[17] RBB150 corpus is available at: https://github.com/linkedtv/videocorpus

German corpora for testing Named Entity Linking (NEL) (e.g. [Tiedemann 2010], [Guzman et al., 2013]), most of them were created for educational purposes and not extracted from real news media. This corpus being extracted from local news media contains localized geographical information (street names, neighborhoods, highways, etc.), person names that are not necessarily famous enough to be included in Wikipedia or in the large Knowledge Bases, local branches of national or international organizations, and events that are important to the local community. Also abbreviations tend to be used more often in television content than in the news media (news articles, blog posts, etc). The 150 transcripts received from RBB were manually annotated by human annotators following an annotation guideline similar to those from TAC KBP and NEEL challenges[18]. The quality of the agreement was judged by the main developers of the corpus and several quality metrics were computed (agreement, NIL clustering, etc.)[Brasoveanu et al., 2016]. The corpus was then made available in various formats for evaluation purposes: NIF, csv, etc. The DBpedia version used for annotation was German DBpedia 2015.

VideoLyzard150 MWCC[19] corpus contains 150 documents selected from youTube news and videos about Climate Change in English from domains like climate change (hurricanes, floods, etc.), politics (elections, incidents), and green tech (electric cars, green energy, etc.). This corpus contains international news, therefore most of the entities are present in English DBpedia. However, since the content is not regional, the surface forms of the entities generally point to the most widely known entity bearing the respective name as opposed to the RBB150 corpus. For example, a mention of Google will be linked to the Mountain View entity and not to Google Germany. The corpus contains fewer abbreviations, as some of the videos came from independent publishers as opposed to television broadcasters. The corpus was created through the same process as the RBB150 corpus, being manually annotated by human annotators, the results being judged by the developers of the corpus.

In both cases the evaluation was performed without taking into consideration NIL entities and only the *top DBpedia types* were considered for each type of entity (dbo:Person, dbo:Organisation, dbo:Place) in order to avoid confusion. By extending each type to all related types (e.g. by taking into account foaf:Person, schema.org/Person, yago:Person classes, in addition to dbo:Person), the results might differ (especially due to the fact that the coverage of the top DBpedia types is not as good as one would expect – in some cases even up to 1 million entities are not assigned to the right top type). Currently there is not enough available information about what types are included in the lexicons or builds of each tool, therefore we considered that restricting the evaluation to the best known types for these entities would be best. Both corpora were annotated w.r.t. the DBpedia 2015-04 Knowledge Base (German and English), therefore additional entities that might be available in the latest build (DBpedia 2015-10) were not included (this was due to the fact that the 2015-10 build was published after the gold standards were compiled).

[18]The annotation guideline used for RBB150 corpus is also available on GitHub: https://github.com/linkedtv/videocorpus/blob/master/rbb150/guideline/annotation-guide.pdf

[19] This corpora was extracted from MWCC video indexes. It will be made available at the same address like RBB150 corpus once the paper that describes its annotation process is published.

It has to be noted that, regardless of how the evaluation is performed w.r.t to the *evaluation scripts* (with Gerbil [Usbeck et al., 2015], with our scripts, and with other scripts), *DBpedia version*, *type coverage* or *language coverage*, most of the tools have not published their best settings, therefore we only considered the tools that had their best settings advertised through their public endpoints or publications. For the RBB150 corpus, a choice was made to select only those annotation tools that return German results through their online endpoints or whose results had good conversion scores (close to 100%) from English to German DBpedia. For the VideoLyzard150 MWCC corpus, English DBpedia links were used.

The evaluation draws upon the following three named entity linking systems: Babelfy [Moro et al., 2014], Spotlight [Daiber et al., 2013], and Recognyze [Weichselbraun et al., 2015]. Gerbil [Usbeck et. al., 2015] includes more annotation services (annotators), but it does not provide us with the possibility to access the evaluation results or create typed evaluations (i.e. different evaluation for each type of entity: Person, Organization, Location).

DBpedia Spotlight is well-known within the Semantic Web and NLP communities for being one of the first tools to use DBpedia and offer semantic approaches to the named entity recognition and disambiguation problems. It was built around a vector space model and is available through a public endpoint.

Babelfy was one of the first graph disambiguation tools that worked in a multilingual setting and it was built around the idea of word sense disambiguation. It offers a free webservice with a limited number of requests and the possibility to evaluate it for research purposes.

Recognyze was built using a lexicon-based NLP approach and later updated to include a wide-array of disambiguation methods.

As can be seen from the results, no tool managed to correctly assign more than half of the entities of the three main entity types (Person, Organisation, Location) on the RBB150 corpus (as predicted due to the fact that the content was regional), while on the second corpus each tool managed to extract more than half of the entities for at least one of the types. Recognyze Location results were not included as the tool used Geonames instead of DBpedia and due to the lack of links between many of the entities from the two Knowledge Bases, the results were very hard to compare.

Corpus	Type	Tool	P	R	F1
RBB150 (German) Regional content	Person	Babelfy	0.61	0.40	0.48
		Recognyze	**0.64**	**0.40**	**0.49**
		Spotlight	0.25	0.35	0.29
	Organization	Babelfy	0.43	0.39	0.23
		Recognyze	0.26	0.16	0.20
		Spotlight	**0.32**	**0.29**	**0.30**
	Location	Babelfy	0.45	0.24	0.31
		Spotlight	**0.31**	**0.42**	**0.36**
VideoLyzard150 MWCC (English) International content	Person	Babelfy	0.61	0.39	0.48
		Recognyze	**0.84**	**0.39**	**0.54**
		Spotlight	0.52	0.54	0.53
	Organization	Babelfy	0.40	0.24	0.30

		Recognyze	0.25	0.21	0.23
		Spotlight	**0.54**	**0.37**	**0.44**
	Location	Babelfy	0.70	0.50	0.59
		Spotlight	**0.67**	**0.67**	**0.67**

Table 19. Evaluation results per type: Person, Location and Organization

As opposed to evaluations that use the full datasets (with or without NILs), evaluations of performance on single types provide better insight into the tool's performance. It has to be noted that it is not uncommon to see differences of several percent after running the experiments again several days later with the same annotation tools due to the fact that some tools use machine learning (therefore current results should be taken as reflecting state-of-the-art in early March 2016). As outlined in the results, all the top tools for a particular entity type are relatively close in terms of F1 measure, although the differences between the types (Person, Organization and Location) are quite significant. Tools that draw upon advanced disambiguation techniques (Babelfy and Recognyze) tend to show higher precision than recall values. These results underline that NEL is a very dynamic field, where most of the evaluated tools outperform their competitors in at least one of the evaluations. As it can be seen, Recognyze consistently wins the Person category, but tends to perform worse than or on a par with Babelfy for the Organization category. We have identified and we are currently in the process of fixing most of the issues that caused this difference in performance for the two profiles: lack of abbreviations, multiple annotations of the same entity due to the different name variants or URIs present in the Knowledge Base (e.g. Google, Inc. and Google), etc. Some of these changes are described in the next section.

5.3. **Ongoing Work**

The evaluations were performed using the current production version of Recognyze. While not included in this evaluation, we have added new features in the next version of Recognyze (some of them based on improvements suggested by the evaluation results). We predict that the changes that would impact the performance the most are the following: extending type coverage to include the types related to the top types; special abbreviation handlers that take abbreviations from abstracts and from various DBpedia fields (especially for people and organizations); new rules for highly ambiguous entities (especially for organizations); support for recent annotation formats; and new profiles (e.g. working towards a DBpedia location profile, etc.).

It has to be noted that by far the feature that impacts the performance the most is the extension of type coverage. As can be seen in Table 19, we were able to add almost 1 million (more than 900,000) additional persons, and more than 150,000 organizations to our Recognyze builds for the next version. These were not taken into account during the current evaluations due to the fact that it is currently not clear how these additional types are used (if they are used) by each tool.

Entity Type	**DBpedia Types**	**Count of DBpedia entities**
Person	dbo:Person	1842134
	schema.org/Person	1730055

	yago:Person100007846	724089
	foaf:Person	2753520
Organization	dbo:Organisation	86466
	schema.org/Organization	86457
	yago:Organization108008335	245192
	yago:Organization101136519	92

Table 20. Number of entities contained in DBpedia if type coverage is extended to multiple classes (with overlapping).

We have also implemented several abbreviation handlers that extract abbreviations from different DBpedia fields (dbo:abstract, dbo:wikiPageRedirects, dbp:acronyms, dbp:abbreviations, Abbreviation107091587, etc.) and used these in the disambiguation process. This is another feature that is hard to compare across multiple tools, as most of them only detect the abbreviations that are very famous and also available as such (e.g., the name itself is available as a resource name in DBpedia builds, as is often the case with names of sports organizations).

Shifting the timeline beyond the next release, we plan to also include multiple new profiles, including events, even though we are still working towards the Knowledge Base builds that will make some of these available (e.g., while DBpedia offers an Event dataset, it mostly contains changes added daily instead of lots of references to current or past events, therefore we created a separate dataset that extracts events directly from Wikipedia pages).

5.4. Summary

Based on the results of the current evaluations, we can also conclude that the German language still poses some challenges for the current generation of annotation tools. The results of the evaluation provided us not only with a clear understanding of our current performance, but also with data regarding which entities are hard to disambiguate. The two corpora used for the evaluations contained many examples from the environmental domain (e.g. Climate Change, ecology, green technology, politics); therefore the evaluations provided us with some insights on what needed to be fixed in order to improve our environmental extraction pipeline. We have categorized these problematic entities (e.g. abbreviations – NCAR, IPO; cross-category entities – Lawton Chiles could be a former Senator, but also a short name for many organizations created by him), and already started working on improvements to fix them in the next version of Recognyze. Future evaluations will also be focused on getting all associated types for the top types, while including more tools, different types of content (social media, videos, news media, etc.) and multiple approaches towards the NEL process (lexicon-based, graph-based, machine learning, etc.).

6. Conclusions and further work

In this deliverable, we have described the second version of the tools we have developed for environmental information extraction. This includes tools to perform entity disambiguation, recognition of environmental terms, and extraction of actors and events. We have made a number of improvements on the first version of the tools:

improving performance, adding a German version of the term recognition tool, and adding new tools for the event and actor recognition, and a number of evaluations have been carried out. The tools are being actively used in the project, in particular in the use cases in WP4 for analysing user engagement around various social media campaigns (Earth Hour and COP21), and in correlating online behaviour with stages of user engagement according to behavioural theories [Fernandez 2016]. The tools have been made available both as web services and as a standalone processing tool that can easily be run from the command line over large datasets by project partners. The evaluation datasets have also been made freely available for public use.

B. List of Tables

Table 1: Evaluation of different term sets on climate corpus 9
Table 2: Spurious terms identified in the corpus 10
Table 3: Comparison of ANNIE and TwitIE as pre-processor on the climate corpus. 10
Table 4: Evaluation against high- and medium-ranked terms in climate corpus..... 11
Table 5: High, medium and low-ranked terms extracted by TermRaider 12
Table 6: Terms unique to each application 12
Table 7: Results of term extraction in the energy corpus 12
Table 8: Results of term extraction in the fracking corpus 13

C. List of Abbreviations

Abbreviation	Explanation
CA	Consortium agreement
DoW	Decription of work, i.e. GA - Annex I
EC	European commission
GA	Grant agreement
IP	Intellectual property
IPR	Intellectual property rights
PC	Project coordinator
PMB	Project management board
SC	Scientific Coordinator
PO	Project officer
PSB	Project steering board
DM	Data Manager
AB	Advisory board
WP	Work package

D. References

[Azar 1989] S. Azar. Understanding and Using English Grammar. Prentice Hall Regents, 1989.

[Baldwin 2915] Timothy Baldwin, Young-Bum Kim, Marie Catherine de Marneffe, Alan Ritter, Bo Han, and Wei Xu. 2015. Shared Tasks of the 2015 Workshop on Noisy User-generated Text: Twitter Lexical Normalization and Named Entity Recognition. In Wei Xu, Bo Han, and Alan Ritter, editors, Proceedings of the Workshop on Noisy User-generated Text, pages 126–135, Beijing, China. Association for Computational Linguistics.

[Bizer 2009] Christian Bizer, Jens Lehmann, Georgi Kobilarov, Soren Auer, Christian Becker, Richard Cyganiak, and Sebastian Hellmann. 2009. DBpedia – a crystallization point for the web of data. Journal of Web Semantics: Science, Services and Agents on the World Wide Web, 7:154–165.

[Bontcheva et al. 2013] K. Bontcheva, L. Derczynski, A. Funk, M.A. Greenwood, D. Maynard, N. Aswani. TwitIE: An Open-Source Information Extraction Pipeline for Microblog Text. Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2013).

[Bontcheva 2014] K. Bontcheva, I. Roberts, L. Derczynski, D. Rout. The GATE Crowdsourcing Plugin: Crowdsourcing Annotated Corpora Made Easy. Proceedings of the meeting of the European chapter of the Association for Computation Linguistics (EACL).

Leo Breiman. 2001. Random forests. Machine Learn- 798 ing, 45(1):5–32. A.E. Cano Basave, A. Varga, M. Rowe, M. Stankovic, and A.-S. Dadzie. 2013. Making sense of microposts (#msm2013) concept extraction challenge. CEUR Workshop Proceedings, 1019:1–15.

[Brasoveanu 2016] Brasoveanu, A.M.P., Nixon, L.J.B., Weichselbraun, A. and Scharl, A. (2016) A Regional News Corpora for Contextualized Entity Discovery and Linking, LREC 2016, ELRA.

[Cherry 2015] Colin Cherry and Hongyu Guo. 2015. The Unreasonable Effectiveness of Word Representations for Twitter Named Entity Recognition. In Rada Mihalcea, Joyce Chai, and Anoop Sarkar, editors, Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 735–745, Denver, Colorado. Association for Computational Linguistics.

[Cobuild 1999] Collins Cobuild, editor. English Grammar. Harper Collins, 1999.

[Cunningham et al. 2002] H. Cunningham, D. Maynard, K. Bontcheva, V. Tablan. GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02). Philadelphia, July 2002.

[Cunningham et al. 2011] H. Cunningham, V. Tablan, I. Roberts, M. A. Greenwood, & N. Aswani (2011). Information extraction and semantic annotation for multi-

paradigm information management. In Current Challenges in Patent Information Retrieval(pp. 307-327). Springer Berlin Heidelberg.

[Daiber 2013] Daiber, J., Jakob, M., Hokamp, C., and Mendes, P. N. (2013). Improving Efficiency and Accuracy in Multilingual Entity Extraction. In Marta Sabou, et al., editors, I-SEMANTICS 2013 - 9th International Conference on Semantic Systems, ISEM '13, Graz, Austria, September4-6, 2013, pages 121–124. ACM.

[Derczynski 2015] Leon Derczynski, Diana Maynard, Giuseppe Rizzo, Marieke van Erp, Genevieve Gorrell, Raphael Troncy, and Kalina Bontcheva. 2015. Analysis of named entity recognition and linking for tweets. Information Processing and Management, 51:32–49.

[Fernandez 2016] Miriam Fernandez, Harith Alani, Lara Piccolo, Christoph Meili, Diana Maynard and Meia Wippoo. Talking Climate Change via Social Media: Communication, Engagement and Behaviour. Proc. of WebSci, May 22-25 2016, Hannover, Germany.

[Finin 2010] T. Finin, W. Murnane, A. Karandikar, N. Keller, J. Martineau, and M. Dredze. 2010. Annotating named entities in Twitter data with crowdsourcing. In Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, pages 80–88.

[Finkel 2005] J.R. Finkel, T. Grenager, and C. Manning. 2005. Incorporating non-local information into information extraction systems by Gibbs sampling. In Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics, pages 363–370. Association for Computational Linguistics.

[Fromreide 2014] Hege Fromreide, Dirk Hovy, and Anders Søgaard. Crowdsourcing and annotating NER for Twitter #drift. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncio´n Moreno, Jan Odijk, and Stelios Piperidis, editors, Proceedings of LREC, pages 2544–2547. European Language Resources Association, 2014.

[Harris 1954] Harris, Z. (1954). "Distributional structure". Word 10 23): 146–162.

[Hellmann 2013] Hellmann, S., Lehmann, J., Auer, S., and Brummer, M. (2013). Integrating NLP using Linked Data. In Harith Alani, et al., editors, The Semantic Web - ISWC 2013- 12th International Semantic Web Conference, Sydney, NSW, Australia, October 21-25, 2013, Proceedings, PartII, volume 8219 of Lecture Notes in Computer Science, pages 98–113. Springer.

[Levenshtein 1966] I. Levenshtein. 1966. Binary Codes capable of cor- recting deletions, insertions and reversals. Soviet Phys. Dokl., 10:707–710.

[Lin 1998] Dekang Lin. 1998. An information-theoretic definition of similarity. In Proceedings of the Fifteenth International Conference on Machine Learning, ICML '98, pages 296–304, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

[Liu 2011] X. Liu, S. Zhang, F. Wei, and M. Zhou. 2011. Recognizing named entities in tweets. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, pages 359–367.

[Maynard 2014] Diana Maynard. Challenges in Analysing Social Media. In Adrian Duşa, Dietrich Nelle, Günter Stock and Gert G. Wagner (eds.) (2014): Facing the Future: European Research Infrastructures for the Humanities and Social Sciences. SCIVERO Verlag, Berlin, 2014.

[Mikholov 2013] Mikolov, Tomas; Sutskever, Ilya; Chen, Kai; Corrado, Greg S.; Dean, Jeff (2013). Distributed representations of words and phrases and their compositionality. Advances in Neural Information Processing Systems.

[Moro 2014] Moro, A., Raganato, A., and Navigli, R. (2014). Entity Linking Meets Word Sense Disambiguation: A Unified Approach. Transactions of the Association for Computational Linguistics, 2:231–244. ACL.

[Naaman 2010] Mor Naaman, Jeffrey Boase, and Chih-Hui Lai. 2010. Is it really about me?: message content in social awareness streams. In Proceedings of the 2010 ACM conference on Computer supported cooperative work, CSCW '10, pages 189–192, New York, NY, USA. ACM.

[Plank 2014] Barbara Plank, Dirk Hovy, Ryan McDonald, and Anders Søgaard. 2014. Adapting taggers to Twitter with not-so-distant supervision. In Junichi Tsujii and Jan Hajic, editors, Proceedings of COLING: Technical Papers, pages 1783–1792, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

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

[Rowe 2015] Matthew Rowe, Milan Stankovic, and Aba-Sah Dadzie. 2015. #Microposts2015 – 5th Workshop on 'Making Sense of Microposts': Big things come in small packages. In Proceedings of the 24th International Conference on World Wide Web Companion, pages 1551–1552.

[Scharl et al. 2013] Arno Scharl, Alexander Hubmann-Haidvogel, Albert Weichselbraun, Heinz-Peter Lang, Marta Sabou (2013). Media Watch on Climate Change -- Visual Analytics for Aggregating and Managing Environmental Knowledge from Online Sources, 46th Hawaii International Conference on System Sciences, pp. 955-964.

[Scharl et al. 2014] Scharl, A., Kamolov, R., Fischl, D., Rafelsberger, W. and Jones, A. (2014). Visualizing Contextual Information in Aggregated Web Content Repositories. 9th Latin American Web Congress (LA-WEB 2014). Ouro Preto, Brazil: Forthcoming.

[Searle 1975] Searle, John R. (1975), "A Taxonomy of Illocutionary Acts", in: Günderson, K. (ed.), Language, Mind, and Knowledge, (Minneapolis Studies in the Philosophy of Science, vol. 7), University of Minneapolis Press, p. 344-369.

[Usbeck 2015] Usbeck, R., Roder, M., Ngomo, A. N., Baron, C., Both, A., Brummer, M., Ceccarelli, D., Cornolti, M., Cherix,D., Eickmann, B., Ferragina, P., Lemke, C., Moro, A., Navigli, R., Piccinno, F., Rizzo, G., Sack, H., Speck, R., Troncy, R., Waitelonis, J., and Wesemann, L. (2015). GERBIL: General Entity Annotator Benchmarking Framework. In Aldo Gangemi, et al., editors, Proceedings

of the 24th International Conference on World Wide Web, WWW 2015, Florence, Italy, May 18-22, 2015, pages 1133–1143. ACM.

[Weichselbraun et al. 2014] Weichselbraun, A., Streiff, D. and Scharl, A. (2014). Linked Enterprise Data for Fine Grained Named Entity Linking and Web Intelligence. 4th International Conference on Web Intelligence, Mining and Semantics (WIMS-2014). Thessaloniki, Greece: ACM Press.

[Weichselbraun 2015] Weichselbraun, A., Streiff, D., and Scharl, A. (2015). Consolidating Heterogeneous Enterprise Data for Named Entity Linking and Web Intelligence. International Journal on Artificial Intelligence Tools, 24(2):1–31.

DecarboNet Consortium

The Open University
Walton Hall
Milton Keynes MK7 6AA
United Kingdom
Tel: +44 1908652907
Fax: +44 1908653169
Contact person: Jane Whild
E-mail: h.alani@open.ac.uk

Waag Society
Piet Heinkade 181A
1019HC Amsterdam
The Netherlands
Tel: +31 20 557 98 14
Fax: +31 20 557 98 80
Contact person: Tom Demeyer
E-mail: tom@waag.org

MODUL University Vienna
Am Kahlenberg 1
1190 Wien
Austria
Tel: +43 1320 3555 500
Fax: +43 1320 3555 903
Contact person: Arno Scharl
E-mail: scharl@modul.ac.at

WWF Schweiz
Hohlstrasse 110
8004 Zürich
Switzerland
+41 442972344
Contact person: Christoph Meili
E-mail: Christoph.Meili@wwf.ch

University of Sheffield
Department of Computer Science
Regent Court, 211 Portobello St.
Sheffield S1 4DP
United Kingdom
Tel: +44 114 222 1930
Fax: +44 114 222 1810
Contact person: Kalina Bontcheva
E-mail: K.Bontcheva@dcs.shef.ac.uk

Green Energy Options
Main Street, 3 St Mary's Crt
Hardwick CB23 7QS
United Kingdom
+44 1223850210
+44 1223 850 211
Contact person: Simon Anderson
E-mail: simon@greenenergyoptions.co.uk

Wirtschaftsuniversität Wien
Welthandelsplatz 1
1020 Wien
Austria
Tel: +43 31336 4756
Fax: +43 31336 774
Contact person: Kurt Hornik
E-mail: kurt.hornik@wu.ac.at