Modality and Negation in Event Extraction

Sander Bijl de Vroe*, Liane Guillou*, Miloš Stanojević, Nick McKenna, Mark Steedman
School of Informatics, University of Edinburgh
{sbdv, liane.guillou, m.stanojevic, nick.mckenna}@ed.ac.uk
steedman@inf.ed.ac.uk

Abstract

Language provides speakers with a rich system of modality for expressing thoughts about events, without being committed to their actual occurrence. Modality is commonly used in the political news domain, where both actual and possible courses of events are discussed. NLP systems struggle with these semantic phenomena, often incorrectly extracting events which did not happen, which can lead to issues in downstream applications. We present an open-domain, lexicon-based event extraction system that captures various types of modality. This information is valuable for Question Answering, Knowledge Graph construction and Fact-checking tasks, and our evaluation shows that the system is sufficiently strong to be used in downstream applications.

1 Introduction

Linguistic modality is frequently used in natural language to express uncertainty with respect to events and states. Downstream NLP tasks that depend on knowing whether an event actually occurred, such as Knowledge Graph construction, Fact-checking, Question Answering and Entailment Graph construction, can benefit from understanding modality. Such information is crucial in the medical domain, for instance, where it facilitates more accurate Information Extraction and search for radiology reports (Wu et al., 2011; Peng et al., 2018). Similarly, if we pose a question in the socio-political domain, such as *Did the protesters attack the police?*, our answer will be different depending on the evidence that the system has observed: *Protesters attacked the police* [yes] or *Protesters are unlikely to have attacked the police* [uncertain][1].

These challenges are exacerbated by the prevalence of the phenomenon. In a multi-domain uncertainty corpus (Szarvas et al., 2012), sentences containing uncertainty cues are significantly more common in newswire text (18%) compared to encyclopedic text (13%). Modality is also frequently observed in editorials (Bonyadi, 2011). We show that within the news genre, modality is common in the politics and sports domains, where experts often make predictions and state their opinions on the possible outcomes of events such as elections or sports matches, and analyse alternative outcomes where situations unfold differently.

We present MONTEE[2], an open-domain system for **Mo**dality and **N**egation **T**agging in **E**vent **E**xtraction. Tagging these phenomena allows us to distinguish between events that took place (e.g. *Protesters attacked the police*), those that did not take place (*Had protesters attacked the police...*), or are uncertain at the time that a document is written (*Protesters may have attacked the police*).

The extracted relations include a predicate and one or two arguments, for example: **Protesters**-*attack*-**police** (from the sentence *Protesters attacked the police*). The predicates are analysed according to the following semantic phenomena: negation, lexical negation, modal operators, conditionality, counterfactuality and propositional attitude. See Table 1 for examples of each category.

We contribute a lexicon of words and phrases that trigger modality, a parser that extracts and tags open-domain event relations for modality (along with an intrinsic evaluation), and a corpus study focusing on the politics domain of a large corpus of news text.

---

*The first two authors contributed equally to this work
[1] Assuming trustworthy source text

[2] https://gitlab.com/lianeg/montee

Category	Example
∅	Protesters attacked the police
Negation	Protesters did **not** attack the police
Lexical negation	Protesters **refrained** from attacking the police
Modal operator	Protesters **may** have attacked the police
Conditional	**If** protesters attack the police...
Counterfactual	**Had** protesters attacked the police...
Propositional attitude	Journalists **said** that protesters attacked the police

Table 1: Modality and negation categories

2 Background

2.1 Semantic Phenomena

Modality: In this work, the focus is on any kind of modality indicating uncertainty, including modal verbs, conditionals, propositional attitudes, and negation. We see modality primarily as a signal for determining whether or not the event in question actually occurred, so that downstream applications can take this into account. We begin by discussing the typical, more specific category of modal operators.

Linguistic modality communicates a speaker's *attitude* towards the propositional content of their utterance. Formally, modality has been defined in terms of quantification over possible worlds (Kratzer, 2012). Other definitions focus on categorising the speaker's attitude, such as epistemic necessity (*That must be John.*), epistemic possibility (*It might rain tomorrow.*), deontic necessity (*You must go.*), and deontic possibility (*You may enter.*) (Van Der Auwera and Ammann, 2005). Sometimes a lexical trigger of modality is ambiguous between categories; English *may*, for example, is ambiguous between an epistemic possibility reading (*It may rain tomorrow.*) and a deontic possibility reading (*You may enter.*)

These definitions have brought about a variety of annotation schemes in practice. Prabhakaran et al. (2012) propose five classes of modality: *ability*, *effort*, *intention*, *success*, and *want*, and train a classifier on crowd-sourced annotated data. Baker et al. (2010) extend the number of modality classes to include *requirement*, *permission*, and *belief*, and combine these with negation. Peñas et al. (2011) take a coarser, epistemic approach, asking whether events are *asserted*, *negated*, or *speculated*, and Saurı et al. (2006) enrich the TimeML specification language with yet other categories (e.g. evidentiality, conditionality).

In English, modality can be expressed in a variety of ways. The modal auxiliaries (e.g. *might, should, can*) are commonly used, but modality can be lexicalised in many other trigger words. Nouns (e.g. *possibility*), adjectives (e.g. *obligatory*), adverbs (e.g. *probably*) and verbs (e.g. *presume that*) can all indicate modality. In the long tail, speakers have access to vastly productive phrases that indicate their attitude. The following examples occurred naturally in the news domain (Zhang and Weld, 2013): *That's how close they were to ...*, *I cannot come up with a scenario that has...*, *That's based on the world wide assumption that....*

Conditionality: A conditional sentence is composed of a subordinate clause (which we will refer to as the antecedent) and a main clause (the consequent). The antecedent and consequent are connected by a conditional conjunction (which in English is often the word *if*), as in the sentence *If they attack there will be war* (Dancygier, 1998). Conditional sentences can have a variety of semantic interpretations, but the most commonly studied, the *hypothetical* conditional, expresses that the consequent (*there will be war*) will hold true if the antecedent (the *attack*) is satisfied (Athanasiadou and Dirven, 1997). For our purposes, the most important part of their semantics is that neither the antecedent nor the consequent are normally entailed by the sentence, so that the speaker is not committed to their truth.

Counterfactuality: In the counterfactual construction a more complicated semantic relation is established between antecedent and consequent, as in the example: *Had they protested, they would be content*. As with modality, this has been formalised more precisely with a possible world semantics (Lewis, 1973; Kratzer, 1981). With a counterfactual, the speaker communicates that in any world similar to the current one, differing only by the proposition in the antecedent, the consequent would hold true (Lewis, 1973). In the above example, if the world is altered by the *protest* in the antecedent, *they would be content* holds true. Again, the crucial semantic information for our work is that neither the antecedent nor the consequent are entailed.

Negation is a semantic category used to change the truth value of a proposition in order to convey that an event, situation or state does not hold (Horn, 1989). It may be expressed explicitly using various means, most notably closed-class function words such as *not, no, never, neither, nor, none*

and *without*, but can also be expressed lexically in open grammatical categories such as nouns (e.g. *impossibility*), verbs (e.g. *decline*, *prevent*), and adjectives (e.g. *unsuccessful*). It may also be expressed implicitly, such as with combinations of certain verb types and tenses (e.g. *The polls were supposed to have closed at midnight*). In this work we consider only explicit cues of negation.

Propositional Attitude and Evidentiality: Propositional attitude allows speakers to indicate the cognitive relations that entities bear to a proposition (McKay and Nelson, 2000). For example, in *Republicans think that Trump has won*, the speaker expresses that *Republicans* hold certain beliefs. In English, such reports are often made using propositional attitude verbs such as *claim, warn* or *believe*. Normally only the entity's thoughts regarding the event are entailed, not the event itself. Propositional attitudes are often used as markers of *evidentiality* in English (Biber and Finegan, 1989). These are important in Question Answering. For example when answering a question using the sentence *The Kremlin says protesters attacked the police* as evidence, mentioning the source (*The Kremlin*) might be particularly important.

2.2 Modality Taggers and Annotated Datasets

A number of approaches have been proposed for the automatic tagging of modality in text. These differ in both the granularity of the classes of modality that the model tags, and the model design.

At the lowest granularity all modality classes are collapsed into a single label. This strategy was employed in the pilot task on modality and negation detection at CLEF 2012, in which participants were asked to automatically label a set of events/states as negated, modal, neither, or both (Morante and Daelemans, 2012). The submitted systems were either purely rule-based (Lana-Serrano et al., 2012; Pakray et al., 2012), or applied rules to the output of a parser (Rosenberg et al., 2012). Modality tagging has also been cast as a supervised learning task (Prabhakaran et al., 2012). Performance of their classifier is reasonably strong on in-domain data (variable across 5 proposed modality classes), but out-of-domain data proves challenging.

Due to the lack of a large, open-domain modality training dataset, we opt for a lexicon-based approach in line with that of Baker et al. (2010). They combine a set of eight modality tags that cap-

ture *factivity* with negation, to denote whether an event/state did or did not happen. They employ two strategies for tagging modal triggers and their targets: 1) string and POS-tag matching between entries in a modality lexicon and the input sentence, 2) a structure-based method which applies rules derived from the lexicon to a flattened dependency tree, inserting tags for modality triggers and targets into the sentence.

Although there is no large, open-domain corpus in which modality is labelled, a number of small datasets exist for specific domains including biomedical text (Thompson et al., 2011), news (Thompson et al., 2017), reviews (Konstantinova et al., 2012), and web-crawled text comprising news, web pages, blogs and Wikipedia (Morante and Daelemans, 2012).

2.3 Event Extraction

Since the introduction of the Open Information Extraction (OIE) task by Banko et al. (2007), a range of open-domain information extraction systems have been proposed for the extraction of relation tuples from text. OIE systems make use of patterns, which may be hand-crafted (Fader et al., 2011; Angeli et al., 2015) or learned through methods such as bootstrapping (Wu and Weld, 2010; Mausam et al., 2012). These patterns may be applied at the sentence level, or to semantically simplified independent clauses identified during a pre-processing step (Del Corro and Gemulla, 2013; Angeli et al., 2015). The majority of systems are restricted to the extraction of binary relations (i.e. relation *triples* consisting of a predicate and two arguments), but systems have also been proposed for the extraction of n-ary relations (Akbik and Löser, 2012; Mesquita et al., 2013). Our system is a form of n-ary event extraction; we extract both binary and unary relations, and relations of higher valencies can be inferred by combining sets of binary relations. A comprehensive survey of OIE systems is provided by Niklaus et al. (2018).

3 Event Extraction System Overview

Whilst many event extraction systems have been developed, none capture the wide range of modality phenomena introduced in Section 2.1. For example, neither OpenIE nor OLLIE extract unary relations. They also fail to adequately handle all of the phenomena we are interested in, in particular counterfactuals and lexical negation. (See Sec-

Johnson | doubts | that | Labour | will | win | the | election

$$\frac{N}{NP}^{TC} \quad \overline{(S[dcl]\backslash NP)/S[em]} \quad \overline{S[em]/S[dcl]} \quad \frac{N}{NP}^{TC} \quad \overline{(S[dcl]\backslash NP)/(S[b]\backslash NP)} \quad \overline{(S[b]\backslash NP)/NP} \quad \overline{NP/N} \quad \overline{N}$$

$$\frac{NP}{}{}^>$$
$$\frac{S[b]\backslash NP}{}{}^>$$
$$\frac{S[dcl]\backslash NP}{}{}^<$$
$$\frac{S[dcl]}{}{}^>$$
$$\frac{S[em]}{}{}^>$$
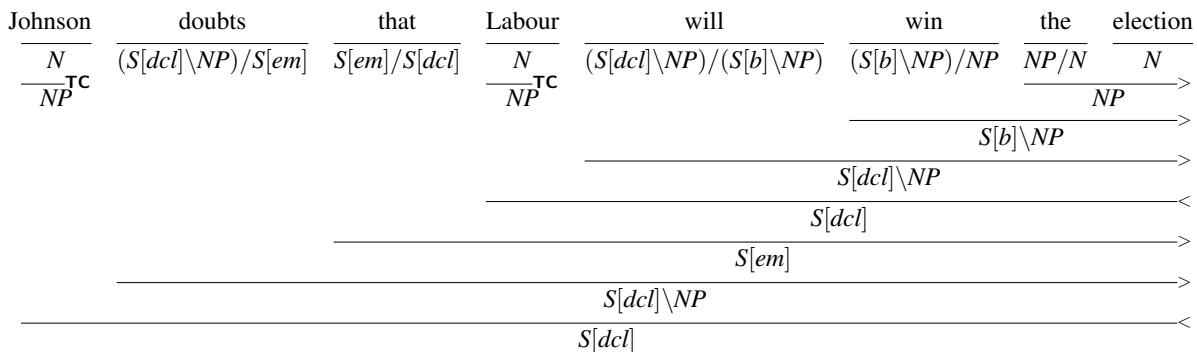$$\frac{S[dcl]\backslash NP}{}{}^<$$
$$S[dcl]$$

Figure 1: CCG parse tree for *Johnson doubts that Labour will win the election*
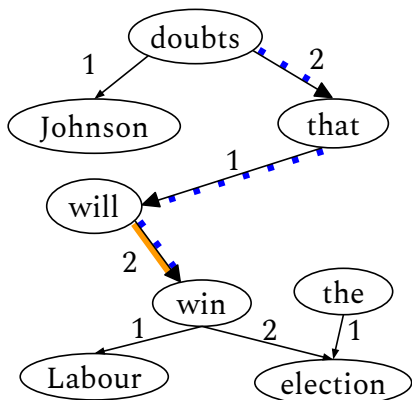

Figure 2: CCG dependency graph for *Johnson doubts that Labour will win the election*; marked paths from *doubts* (blue, dotted) and *will* (orange, solid) to *win*.

tion 6 for a comparison of our system with OpenIE and OLLIE.) We therefore construct our own event extraction system.

Our system takes as input a text document, and for each sentence outputs a set of event relations. An event relation tuple consists of a predicate and either one, or two arguments (e.g. **(The) protest**-*ended*, **Angela Merkel**-*addressed*-**NPD protesters**). We use a pipeline approach similar to that described by Hosseini et al. (2018), which allows us to extract open-domain relations.

Each sentence in the document is parsed using the *RotatingCCG* parser (Stanojević and Steedman, 2019) over which we construct a CCG dependency graph using a method similar to the one proposed by Clark et al. (2002). (See Figure 2 for an example of a dependency graph and Figure 1 for the CCG parse tree from which it was extracted.) CCG dependency graphs are more expressive than standard dependency trees because they can encode long-range dependencies, coordination and reentrancies. We traverse the dependency graph, starting from verb and preposition nodes, until an

argument node is reached. The traversed nodes, which are used to form the predicate strings, may include (non-auxiliary) verbs, verb particles, adjectives, and prepositions. The CCG argument slot position, corresponding to the grammatical case of the argument (e.g. 1 for nominative, 2 for accusative), is appended to the predicate.

Our focus is on the extraction of binary and unary relations. Binary relations may be extracted from dependency paths between two entities. Extraction of unary relations, which have only one such endpoint, poses a harder challenge (Szpektor and Dagan, 2008) – we must decide whether they are truly a unary relation, or form part of a binary relation. Therefore linguistic knowledge must be carefully applied to extract meaningful unary relations. We extract unary relations for the following cases: verbs with a single argument including intransitives (*bombs exploded*) and passivised transitives (*protests were held*), and copular constructions (*Greta Thunberg is a climate activist*).

In addition to binary and unary relations we also extract n-ary relations which combine two binary relations via prepositional attachment. These are of the form: **arg1**-*predicate*-**arg2**-*preposition*-**arg3**, and are constructed by combining the two binary relations **arg1**-*predicate*-**arg2** and **arg2**-*preposition*-**arg3**. For example **Protesters**-*marched on*-**Parliament Square** and **Parliament Square**-*in*-**London** combine to form the new relation **Protesters**-*marched on Parliament Square in*-**London** (from the sentence: *Protesters marched on Parliament Square in London*).

Passive predicates are mapped to active ones. Modifiers such as *managed to* as in the example *Boris Johnson managed to secure a Brexit deal* are also included in the predicate. As these may be rather sparse, we provide the option to also extract the relation without the modifier.

Lemma	Category	POS-tag	Strength
succeed	MOD	VB	4
shall	MOD	MD	3
conceivably	MOD	RB	2
impossible	MOD	JJ	0
as long as	COND	RB	2
concede	ATT_SAY	VB	4
reckon	ATT_THINK	VB	2

Table 2: Example lexicon entries

Algorithm 1 Tagging Modal Events

```
 1: procedure TAGMODALEVENTS(sentence s, events e, lex-
    icon l)
 2:     G, event_nodes ← CCG_dep_parse(s, e)
 3:     trigger_nodes ← [ ]
 4:     for n in G do
 5:         if check_lexicon(n,l) or check_cf(n,G) then
 6:             trigger_nodes.add(n)
 7:         end if
 8:     end for
 9:     for e_n in event_nodes do
10:         for t_n in trigger_nodes do
11:             if path_between(e_n, t_n) then
12:                 e_n ← update(e_n,t_n.tag)
13:             end if
14:         end for
15:         e_n.tag ← tag_precedence(e_n)
16:         event_nodes.update(e_n)
17:     end for
18:     return event_nodes
19: end procedure
```

Arguments are classified as either a Named Entity (extracted by the CoreNLP (Manning et al., 2014) Named Entity recogniser), or a general entity (all other nouns and noun phrases). Arguments are mapped to types by linking to their Freebase (Bollacker et al., 2008) IDs using AIDA-Light (Nguyen et al., 2014), and subsequently mapping these IDs to their fine-grained FIGER types (Ling and Weld, 2012). For example, *Angela Merkel* would be mapped to *person/politician* and *NPD* (Nationaldemokratische Partei Deutschland) to *government/political_party*. The type system may be leveraged to identify events belonging to specific domains, for example, to identify and track political events such as elections, debates, protests etc. and the entities involved.

4 Lexicon

Since many of the phenomena we capture involve lexical trigger items, we opt for a lexicon-based approach. Triggers identified using the lexicon can then be linked to event nodes in the CCG dependency graph. Entries in the lexicon cover modality, lexical negation, propositional attitude, and conditionality, with counterfactuality handled separately. Each entry contains the lemma, the categories that it covers, the POS-tag and an estimate of the epistemic strength that the word would normally indicate. A few examples are included in Table 2.

Our lexicon is constructed by pooling together various lexical resources. The majority of the entries derive from the modality lexicon presented by Baker et al. (2010), who use it for a similar rule-based tagging approach. Their lexicon contains just under a thousand instances, but includes multiple forms for each verb inflection. Using only infinitival forms, we add approximately 200 of the modal entries to our own lexicon.

For modelling propositional attitude, we include a list of reporting verbs found in Fay (1990). This added roughly another 120 phrases to the resource.

The new entries were separated by attitudes expressed through speech (tag ATT_SAY, e.g. *say, state*) and attitudes of thought (tag ATT_THINK, e.g. *suspect, assume*).

More phrases expressing uncertainty are found in a data set of news domain sentences describing conflicting events, such as a *win* and a *loss* (Guillou et al., 2020). Such sentences often contained descriptions of events that didn't actually happen. Yet more related words were found by generating each entry's WordNet synonyms and antonyms (Miller, 1995). We filtered and annotated these manually to obtain just under another 200 phrases, and added these to the lexicon. We also took inspiration from Somasundaran et al. (2007), especially for conditionals. In aggregate, this work resulted in a resource of 530 phrases.

We also annotated each phrase with a modal category. Our lexicon contains the categories *deontic*, *intention* and *desire*, and for the remaining phrases lists a indication of epistemic strength, with values 4 (*definitely*), 3 (*probably*), 2 (*possibly*), 1 (*probably not*) and 0 (*definitely not*). The latter correspond to lexical negation. The epistemic strength values were manually annotated by the authors, and are proposed as a means to collect subsets of events, such as all events marked as *probable* or higher. This phenomenon deserves more attention in future research however, as it is highly contextualised. For example, *could win the lottery* should deserve a different annotation to *could have breakfast*.

5 Modality Parser

We use the CCG-based event extraction system (Section 3) and the expanded modality lexicon (Section 4) in tandem to assign modal categories to events. The procedure is described in Algorithm 1. The focus of the tagger is to identify the bulk of uncertain events: we prioritise recall over precision, so that we can expect events without a tag to have actually happened.

The event extractor produces a CCG dependency graph $\mathcal{G}$ that contains a node n for each word in the sentence (line 2 of the algorithm). We then decide which of these nodes is a trigger (lines 4-7). For modality, negation, lexical negation, propositional attitude and conditionals, we tag these nodes if the node's lemma is present in the lexicon (*check_lexicon* function, line 5). The loop in the algorithm covers the simple case of single token modal triggers (such as *possible*), and can be extended to multi token triggers (e.g. *shoot for*)[3].

Counterfactual nodes are identified separately. The *check_cf* function (line 5) finds instances of the token "had" that are assigned one of two indicative CCG supertags: $(((S\backslash NP)\backslash(S\backslash NP))/(S[pt]\backslash NP))/NP$ or $((S/S)/(S[pt]\backslash NP))/NP$. For example in the sentence *The protesters would have been arrested, had they attacked the police*, the token "had" would be assigned the CCG supertag $(((S\backslash NP)\backslash(S\backslash NP))/(S[pt]\backslash NP))/NP$ and is therefore recognised as an instance of counterfactual had. Additionally, any instance of "if" that governs an instance of "had", is labelled as counterfactual. Upon realising that even this common counterfactual pattern was rare in the corpus, we decided not to engineer further counterfactual patterns.

We can then decide whether an event node should be tagged, by checking whether there is a path in the dependency graph from the trigger nodes to the event node (lines 9-12). Figure 2 illustrates the intuition behind walking the dependency graph. The graph shows a path from both *doubt* and *will* to *win*. This works because the existence of a path between a trigger node and an event node corresponds to the trigger node taking syntactic scope over the event node. The semantic phenomena we handle all rely heavily on this syntactic process (for example negation, see McKenna and Steedman (2020)).

A single event node may be connected to multiple triggers, so we choose the final tag on line 15. Since our primary concern is whether the event happened, we do not combine tags and instead assign a single tag based on the following order of precedence: MOD, ATT_SAY, ATT_THINK, COND, COUNT, LNEG, NEG. The negation categories need to be ordered last because an event that is negated and modal is still uncertain (e.g. *might not play* shouldn't result in NEG_play), but the ordering is otherwise arbitrary.

6 Comparison with Existing Event Extraction Systems

We highlight the capabilities of our system on five example sentences, comparing with two existing event extraction systems: OpenIE (Angeli et al., 2015) and OLLIE (Mausam et al., 2012). Note that this is not intended as a conclusive evaluation of systems, but rather as a high-level overview of the phenomena captured by each of the systems. See Table 3 for a comparison of the relations extracted by MONTEE, OpenIE and OLLIE. The examples are all naturally occurring sentences from the news domain, obtained by a web search targeted to the modality categories discussed in this paper. To enable a fair comparison, we focus on the extraction of binary relations, as neither OpenIE nor OLLIE was designed to extract unary relations.

Whilst Stanford OpenIE (Angeli et al., 2015), OLLIE (Mausam et al., 2012), and OLLIE's predecessor REVERB (Fader et al., 2011) may be used to extract binary relations for events, they do not explicitly mark events for modality or negation. Stanford OpenIE (Angeli et al., 2015) typically includes modals as part of the predicate (for example: (Protesters; may have attacked; police)), but ignores the other categories of linguistic modality described in Section 5. In particular it does not extract relations for sentences involving negation or propositional attitude, omits lexical negations, and is easily confused by sentences involving conditionals or counterfactuals.

OLLIE (Mausam et al., 2012) handles the phenomena in more detail. It identifies conditionals by detecting markers such as "if" and "when", and labels the *enabling condition* for extracted relations that are governed by a conditional[4]. It typically includes modals and negation as part of the predicate,

---

[3]We implement this as a recursive loop over a Trie data structure.

[4]The labelling of *conditional* is not applied in the first example in Table 3 as no relation is extracted for the consequent.

MoNTEE	OpenIE	OLLIE
The guerrillas are ready to talk with the Soviets, if Moscow is willing.		
MOD_(guerrillas; talk; Soviets)	(guerrillas; are; ready)	(Moscow; is; willing)
COND_(Moscow; be willing)	(guerrillas; talk with; Soviets)	
	(guerrillas; talk; if Moscow is willing)	
	(guerrillas; talk; willing)	
	(Moscow; is; if Moscow is willing)	
	(Moscow; is; willing)	
Had Trump won the election, Cummings would still be in Downing Street.		
COUNT_(Trump; win; election)	(Trump; Had Trump won; election)	(Trump; Had won; the election)
MOD_(Cummings; be in; D.St.)	(Cummings; would; would still be in D.St.)	(Cummings; would still be in; D.St.)
Protesters did not attack the Police.		
NEG_(Protesters; attack; police)	∅	(Protesters; did not attack; the police)
Parliament failed to investigate the Kremlin.		
(Parliament; failed to investigate; Kremlin)	(Parliament; investigate; Kremlin)	(Parliament; failed to investigate; the Kremlin)
LNEG_(Parliament.; investigate; Kremlin)		(Parliament; to investigate; the Kremlin)
Ed Miliband says the government betrayed Yorkshire.		
ATT_SAY_(government; betray; Yorkshire)	∅	(the government; betrayed; Yorkshire)
(Ed-Miliband; say)		[attrib=Ed Miliband says]

Table 3: Comparison of MoNTEE with OpenIE and OLLIE

and captures propositional attitude in its handling of attribution (e.g. *Ed Miliband says...*). Like OpenIE, OLLIE is not designed to handle counterfactuals. In terms of lexical negations, OLLIE extracts the predicate both with and without the negation cue, which is undesirable if the downstream NLP application needs to be able to distinguish between events that took place and those that did not.

7 Evaluating System Performance

In the absence of a pre-existing open-domain evaluation dataset that closely matches the task we are interested in, we conduct an intrinsic evaluation of our modality-aware event extraction system. We measure performance on a set of 100 extracted event relations with manually annotated labels denoting the degree of certainty (happened, didn't happen, uncertain). An event relation consists of a predicate plus argument pair (e.g. (Protesters; attack; police)). Note that we exclude both OLLIE and OpenIE from this evaluation as neither system is designed to handle the complete set of modality or negation phenomena we are interested in (c.f. Section 6).

We filtered the articles in the NewsSpike corpus (Zhang and Weld, 2013) to obtain those where at least 20% of the event relations are tagged (to guarantee a reasonably dense distribution of modality). We then randomly selected five articles and processed them using our system to extract event relations. From these articles we selected 100 event relations[5]. At the sentence-level we ensured that we include only one event relation for each predicate node in the dependency graph, since all event relations with the same predicate node will be assigned the same modality.

The set of 100 event relations was manually annotated by two of the authors of this paper, one native English speaker and one fluent speaker. For each event relation, we asked the annotators to answer the question *Does the text entail that the event definitely happens?* using the following labels: the event happened (2), is uncertain (1), didn't happen (0). Inter-annotator agreement over the set of 100 event relations was measured using Cohen's Kappa (Cohen, 1960). The agreement score was 0.77, indicating *substantial agreement*, and the annotations differed for only 16 examples. Following the initial annotation task, the two annotators resolved the disagreements, which resulted in the gold standard test set.

To evaluate our system, we mapped system-assigned modal and negation tags to the set of certainty labels, with LNEG and NEG tags mapped to 0 (didn't happen), empty tags mapped to 2 (happened), and all other tags mapped to 1 (uncertain). In Table 4 we report the micro- and macro-averaged precision, recall and F1 scores. As the number of event relations per modality tag category is too small for a meaningful error analysis over types,

---

[5]We excluded those event relations for which the predicate contains only a preposition as these have little meaning unless they form part of a high-order n-ary relation.

	Precision	Recall	F1
Micro-average	0.81	0.81	0.81
Macro-average	0.72	0.88	0.76

Table 4: Intrinsic evaluation results

	General	Politics	Sports
Articles	532,651	58,521	196,098
Sentences	20,683,584	2,280,312	8,056,704
Relations	96,774,467	11,265,585	37,936,677
Distribution of tags (percentage of all relations)			
∅	77.83	74.78	78.75
Tag	22.17	25.22	21.25
Distribution of *types* of tag (percentage of tagged relations)			
Modal	64.59	66.04	65.10
ATT_say	21.54	21.28	19.94
ATT_think	2.22	1.72	2.32
Conditional	4.03	4.09	3.99
Counterfactual	0.17	0.19	0.19
Negation	6.86	6.00	7.79
Lexical Negation	0.58	0.67	0.67

Table 5: Relation tagging summary by news domain

we provide aggregated scores. The distribution of certainty labels is also uneven, with few negations marked in the gold standard. We therefore take the micro-averaged F1 score of 0.81 to be the definitive result.

We performed an error analysis of the 17 errors made by our system on the test set of 100 event relations. Parsing was a common issue, with five errors attributed to general parsing mistakes, and five errors due to missing dependency links between reporting verbs and events in quoted text (e.g. *"Police were attacked", they said*). Two mistakes were due to human error, as the annotators also missed these reporting verbs in longer sentences. Then, three errors arose from issues with the lexicon. Two of these stemmed from lack of coverage: our lexicon does not handle temporal displacement, as in *We won't act **until** the white house gives more information*. The other was caused by incorrect application of a lexical entry, which would need to be disambiguated using context. Finally, two errors could also have been avoided by handling linguistic *aspect*, as in *they began the process to....* Future research could thus focus on expanding the lexicon by these final categories of displacement, and taking context into account when linking a word to the lexicon.

8 Corpus Analysis

We conducted a corpus analysis of extracted relations over the NewsSpike corpus (Zhang and Weld, 2013). NewsSpike contains approximately 540K multi-source news articles (approximately 20M sentences) collected over a period of six weeks. We report on the distributions of tagged phenomena over the set of binary relations[6] extracted from news articles in the complete corpus (general domain), and for the subsets of articles related to the politics and sports domains.

The NewsSpike corpus does not include topic or domain information in the article-level metadata. Therefore to identify articles belonging to the politics and sports domains we leveraged the named

---

[6]The corpus study of unary relations is left for future work

entity linker AIDA-Light (Nguyen et al., 2014) and the FIGER type system (Ling and Weld, 2012). We first identified the set of fine-grained FIGER types related to each sub-domain, and then obtained the set of entities belonging to each type. Next we used the output of AIDA-Light to identify the set of articles for which more than 40% of the entities found by the linker belonged to the politics domain, with at least two political entities. We repeated this process for the sports domain, with a lowered threshold of 25%, as the sports topic is less likely to overlap with other topics.

The distribution of relation tags over the general, politics, and sports domains is shown in Table 5. For the politics domain just over 25% of the extracted relations are tagged by the modality parser, which is more than for the sports or general domains. In particular, modals are more prevalent. This suggests that whilst it is important to identify modality in the general news domain, it is particularly important in the politics domain.

The top ten most frequent trigger words found in the general domain are: the propositional attitude trigger *say*, the modal triggers *will*, *would*, *can*, *could*, *may*, *should*, *want* and *have to*, and the conditional trigger *if*. The same top ten are also observed for the politics domain (with different frequencies), and for the sports domain the propositional attitude trigger *think* replaces *want*. The similarity of these lists is perhaps not surprising as all three domains belong to the news genre.

9 Future Work

An obvious limitation of our approach is that it does not take into account the context in which events and trigger words occur. Modality is a context-dependent phenomenon, so using the sentential context would improve accuracy. For example, the word *unbelievable* is ambiguous between an *unlikely* and an *amazing, and happened* reading. Relatedly, our concept of epistemic strength is highly context-sensitive, and requires further development. A promising avenue is to develop a pre-training procedure for a modality-aware contextualised language model (Devlin et al., 2019; Zhou et al., 2020). We plan to use our modal lexicon to identify sentences with modality triggers. We will then gather human annotations of the certainty that each event happened, and use this annotated data to train a modality-aware language model able to classify event uncertainty. Such a system might eventually even tackle the long-tail of modal examples mentioned in Section 2.1.

We will also investigate the application of zero shot and few shot learning to the problem of detecting modality and negation. This could provide a way to leverage a large pre-trained language model together with a small annotated corpus.

Our system was developed for English, but work is already underway to develop event extraction systems for other languages including German and Chinese. Extending to other languages would allow us to apply our methods to multilingual and crosslingual NLP tasks. Finally, most CCG parsers, including the one used in this work, are trained on English CCGbank (Hockenmaier and Steedman, 2007). This makes them perform well on news text, but accuracy suffers on out-of-domain sentences, primarily those involving questions. The results could be improved by retraining the parser on the CCG annotated questions dataset (Rimell and Clark, 2008; Yoshikawa et al., 2019), allowing us to apply our system to the task of open-domain Question Answering in an extrinsic evaluation.

10 Conclusion

We have presented MONTEE, a modality-aware event extraction system that can distinguish between events that took place, did not take place, and for which there is a degree of uncertainty. Being able to make such distinctions is crucial for many downstream NLP applications, including Knowledge Graph construction and Question Answering.

Our parser performs strongly on an intrinsic evaluation of examples from the politics domain and our corpus analysis supports our claim that modality is an important phenomenon to handle in this domain.

References

Alan Akbik and Alexander Löser. 2012. KrakeN: N-ary facts in open information extraction. In *Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction (AKBC-WEKEX)*, pages 52–56, Montréal, Canada. Association for Computational Linguistics.

Gabor Angeli, Melvin Jose Johnson Premkumar, and Christopher D. Manning. 2015. Leveraging linguistic structure for open domain information extraction. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 344–354, Beijing, China. Association for Computational Linguistics.

Angeliki Athanasiadou and René Dirven. 1997. Conditionality, hypotheticality, counterfactuality. *Amsterdam Studies in the Theory and History of Linguistic Science Series 4*, pages 61–96.

Kathryn Baker, Michael Bloodgood, Bonnie Dorr, Nathaniel W. Filardo, Lori Levin, and Christine Piatko. 2010. A modality lexicon and its use in automatic tagging. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Michele Banko, Michael J. Cafarella, Stephen Soderland, Matt Broadhead, and Oren Etzioni. 2007. Open information extraction from the web. In *Proceedings of the 20th International Joint Conference on Artifical Intelligence*, IJCAI'07, page 2670–2676, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Douglas Biber and Edward Finegan. 1989. Styles of stance in english: Lexical and grammatical marking of evidentiality and affect. *Text-interdisciplinary journal for the study of discourse*, 9(1):93–124.

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: A collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*, SIGMOD '08, page 1247–1250, New York, NY, USA. Association for Computing Machinery.

Alireza Bonyadi. 2011. Linguistic manifestations of modality in newspaper. *International Journal of Linguistics*, 3(1):E30.

Stephen Clark, Julia Hockenmaier, and Mark Steedman. 2002. Building deep dependency structures using a wide-coverage CCG parser. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 327–334, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Barbara Dancygier. 1998. *Conditionals and Prediction: Time, Knowledge and Causation in Conditional Constructions*, volume 87 of *Cambridge Studies in Linguistics*. Cambridge University Press.

Luciano Del Corro and Rainer Gemulla. 2013. Clausie: Clause-based open information extraction. In *Proceedings of the 22nd International Conference on World Wide Web*, WWW '13, page 355–366, New York, NY, USA. Association for Computing Machinery.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Anthony Fader, Stephen Soderland, and Oren Etzioni. 2011. Identifying relations for open information extraction. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1535–1545, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Richard Fay, editor. 1990. *Collins Cobuild English Grammar*. Collins, United Kingdom.

Liane Guillou, Sander Bijl de Vroe, Mohammad Javad Hosseini, Mark Johnson, and Mark Steedman. 2020. Incorporating temporal information in entailment graph mining. In *Proceedings of the Graph-based Methods for Natural Language Processing (TextGraphs)*, pages 60–71.

Julia Hockenmaier and Mark Steedman. 2007. CCG-bank: A corpus of CCG derivations and dependency structures extracted from the Penn Treebank. *Computational Linguistics*, 33(3):355–396.

Laurence Horn. 1989. *A Natural History of Negation*. University of Chicago Press.

Mohammad Javad Hosseini, Nathanael Chambers, Siva Reddy, Xavier R. Holt, Shay B. Cohen, Mark Johnson, and Mark Steedman. 2018. Learning typed entailment graphs with global soft constraints. *Transactions of the Association for Computational Linguistics*, 6:703–717.

Natalia Konstantinova, Sheila C.M. de Sousa, Noa P. Cruz, Manuel J. Maña, Maite Taboada, and Ruslan Mitkov. 2012. A review corpus annotated for negation, speculation and their scope. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 3190–3195, Istanbul, Turkey. European Language Resources Association (ELRA).

Angelika Kratzer. 1981. Partition and revision: The semantics of counterfactuals. *Journal of Philosophical Logic*, 10(2):201–216.

Angelika Kratzer. 2012. *Modals and conditionals: New and revised perspectives*, volume 36. Oxford University Press.

Sara Lana-Serrano, Daniel Sánchez-Cisneros, Paloma Martínez Fernández, Antonio Moreno-Sandoval, and Leonardo Campillos Llanos. 2012. An approach for detecting modality and negation in texts by using rule-based techniques. In *CLEF 2012 Evaluation Labs and Workshop, Online Working Notes, Rome, Italy, September 17-20, 2012*, volume 1178 of *CEUR Workshop Proceedings*. CEUR-WS.org.

David Lewis. 1973. Counterfactuals and comparative possibility. *Journal of Philosophical Logic*, pages 418–446.

Xiao Ling and Daniel S. Weld. 2012. Fine-grained entity recognition. In *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence*, AAAI'12, page 94–100. AAAI Press.

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

Mausam, Michael Schmitz, Stephen Soderland, Robert Bart, and Oren Etzioni. 2012. Open language learning for information extraction. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 523–534, Jeju Island, Korea. Association for Computational Linguistics.

Thomas McKay and Michael Nelson. 2000. Propositional attitude reports.

Nick McKenna and Mark Steedman. 2020. Learning negation scope from syntactic structure. In *Proceedings of the Ninth Joint Conference on Lexical and Computational Semantics*, pages 137–142.

Filipe Mesquita, Jordan Schmidek, and Denilson Barbosa. 2013. Effectiveness and efficiency of open relation extraction. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 447–457, Seattle, Washington, USA. Association for Computational Linguistics.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Roser Morante and Walter Daelemans. 2012. Annotating modality and negation for a machine reading evaluation. In *CLEF 2012 Evaluation Labs and Workshop, Online Working Notes, Rome, Italy, September 17-20, 2012*, volume 1178 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Dat Ba Nguyen, Johannes Hoffart, Martin Theobald, and Gerhard Weikum. 2014. Aida-light: High-throughput named-entity disambiguation. *Workshop on Linked Data on the Web*, 1184:1–10.

Christina Niklaus, Matthias Cetto, André Freitas, and Siegfried Handschuh. 2018. A survey on open information extraction. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3866–3878, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Partha Pakray, Pinaki Bhaskar, Somnath Banerjee, Sivaji Bandyopadhyay, and Alexander F. Gelbukh. 2012. An automatic system for modality and negation detection. In *CLEF 2012 Evaluation Labs and Workshop, Online Working Notes, Rome, Italy, September 17-20, 2012*, volume 1178 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Anselmo Peñas, Eduard H Hovy, Pamela Forner, Álvaro Rodrigo, Richard FE Sutcliffe, Corina Forascu, and Caroline Sporleder. 2011. Overview of qa4mre at clef 2011: Question answering for machine reading evaluation. In *CLEF (Notebook Papers/Labs/Workshop)*, pages 1–20. Citeseer.

Yifan Peng, Xiaosong Wang, Le Lu, Mohammadhadi Bagheri, Ronald Summers, and Zhiyong Lu. 2018. Negbio: a high-performance tool for negation and uncertainty detection in radiology reports. *AMIA Summits on Translational Science Proceedings*, 2018:188.

Vinodkumar Prabhakaran, Michael Bloodgood, Mona Diab, Bonnie Dorr, Lori Levin, Christine D. Piatko, Owen Rambow, and Benjamin Van Durme. 2012. Statistical modality tagging from rule-based annotations and crowdsourcing. In *Proceedings of the Workshop on Extra-Propositional Aspects of Meaning in Computational Linguistics*, pages 57–64, Jeju,

Republic of Korea. Association for Computational Linguistics.

Laura Rimell and Stephen Clark. 2008. Adapting a lexicalized-grammar parser to contrasting domains. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 475–484, Honolulu, Hawaii. Association for Computational Linguistics.

Sabine Rosenberg, Halil Kilicoglu, and Sabine Bergler. 2012. CLaC Labs: Processing modality and negation. working notes for QA4MRE pilot task at CLEF 2012. In *CLEF (Online Working Notes/Labs/Workshop)*, volume 1178 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Roser Saurı, Marc Verhagen, and James Pustejovsky. 2006. Annotating and recognizing event modality in text. In *Proceedings of 19th International FLAIRS Conference*.

Swapna Somasundaran, Josef Ruppenhofer, and Janyce Wiebe. 2007. Detecting arguing and sentiment in meetings. In *Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue*, pages 26–34.

Miloš Stanojević and Mark Steedman. 2019. CCG parsing algorithm with incremental tree rotation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 228–239, Minneapolis, Minnesota. Association for Computational Linguistics.

György Szarvas, Veronika Vincze, Richárd Farkas, György Móra, and Iryna Gurevych. 2012. Cross-genre and cross-domain detection of semantic uncertainty. *Computational Linguistics*, 38(2):335–367.

Idan Szpektor and Ido Dagan. 2008. Learning entailment rules for unary templates. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 849–856, Manchester, UK. Coling 2008 Organizing Committee.

P. Thompson, R. Nawaz, J. McNaught, and S. Ananiadou. 2011. Enriching a biomedical event corpus with meta-knowledge annotation. *BMC Bioinformatics*, 12:393.

Paul Thompson, Raheel Nawaz, John Mcnaught, and Sophia Ananiadou. 2017. Enriching news events with meta-knowledge information. *Language Resources and Evaluation*, 51(2):409–438.

Johan Van Der Auwera and Andreas Ammann. 2005. Overlap between situational and epistemic modal marking. *World atlas of language structures*, pages 310–313.

Andrew S Wu, Bao H Do, Jinsuh Kim, and Daniel L Rubin. 2011. Evaluation of negation and uncertainty detection and its impact on precision and recall in search. *Journal of digital imaging*, 24(2):234–242.

Fei Wu and Daniel S. Weld. 2010. Open information extraction using Wikipedia. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 118–127, Uppsala, Sweden. Association for Computational Linguistics.

Masashi Yoshikawa, Hiroshi Noji, Koji Mineshima, and Daisuke Bekki. 2019. Automatic generation of high quality CCGbanks for parser domain adaptation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 129–139, Florence, Italy. Association for Computational Linguistics.

Congle Zhang and Daniel S Weld. 2013. Harvesting parallel news streams to generate paraphrases of event relations. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1776–1786.

Ben Zhou, Qiang Ning, Daniel Khashabi, and Dan Roth. 2020. Temporal common sense acquisition with minimal supervision. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7579–7589, Online. Association for Computational Linguistics.