Paths for Uncertainty: Exploring the Intricacies of Uncertainty Identification for News

Chrysoula Zerva, Sophia Ananiadou
National Centre for Text Mining
School of Computer Science, The University of Manchester, United Kingdom
{chrysoula.zerva, sophia.ananiadou}@manchester.ac.uk

Abstract

Currently, news articles are produced, shared and consumed at an extremely rapid rate. Although their quantity is increasing, at the same time, their quality and trustworthiness is becoming fuzzier. Hence, it is important not only to automate information extraction but also to quantify the certainty of this information. Automated identification of expressions that affect certainty has been studied both in the scientific and newswire domains, but performance is considerably higher in tasks focusing on scientific text. We compare the differences in the definition and expression of uncertainty between a scientific domain, i.e., biomedicine, and newswire. We delve into the different aspects that affect the certainty of an extracted event in a news article and examine whether they can be easily identified by techniques already validated in the biomedical domain. Finally, we present a comparison of the syntactic and lexical differences between the the expression of certainty in the biomedical and newswire domains, using two annotated corpora.

1 Introduction

The increasing amount of data readily available in digital form across various domains presents challenges for both researchers and the general public. Although this has greatly improved access to data and dissemination of knowledge, it is becoming increasingly difficult to quickly identify a piece of information that is pertinent to our needs among the vast amounts of data, as well as to assess its certainty and credibility. Advances in information extraction methods and in particular *event* extraction tasks (McClosky et al., 2011; Nguyen et al., 2016; Cao et al., 2016), capture complex information structures to that can capture n-ary relations between entities, and better represent facts and statements made by authors.

While being able to extract rich information in a structured manner is important, not all extracted information is equally trustworthy. It is thus necessary to apply measures of confidence that will allow us to assess the credibility of events mined from different documents. Such measures may take into account different factors affecting our confidence in a specific event, such as the reliability of the source (Lucassen and Schraagen, 2010), the timeliness of the event (Pustejovsky, 2017), the performance of the event extraction tool etc. Along with such "external" factors affecting our trust in the event, another important aspect is how certainty is expressed in the context of the event by the author, since not all information mentioned in text is expressed with equal certainty. Some events are explicitly identified as speculations, as hypothetical situations, as disputed allegations, as conditional facts, and so on. Thus, it is important to complement event extraction methods with identification of such textual phenomena, in order to enrich extracted events with an attribute of certainty.

Identification of textual uncertainty and hedging is a mature research topic, with an emphasis on the scientific domain (Hyland, 1998). Methods to detect certainty and related types of information are widely applied in the field of biomedical text mining to assess the veracity of information, and the problem is approached both in terms of framing certainty and annotating corpora accordingly, and by applying machine learning techniques for the automated identification of uncertain statements and events (Kilicoglu et al., 2017; Malhotra et al., 2013). In the news domain, while machine learning techniques have been used to mine sentiment, subjectivity etc, efforts concerned with (un)certainty identification have focussed mostly on the provision of classification framework for uncertainty (Rubin, 2010) or its combination with polarity to determine event factuality (Sauri and Pustejovsky, 2007). However, there has been less emphasis on applications that focus on automatically recognising uncertainty,

Proceedings of the Workshop on Computational Semantics beyond Events and Roles (SemBEaR-2018), pages 6–20
New Orleans, Louisiana, June 5, 2018.

especially in relation to events. Moreover, early attempts at automated identification of uncertainty cues (weasels) in both the general and biomedical domains showed more than 0.30 difference in F-score between the two domains (0.50 for Wikipedia versus 0.87 for Bio (Tang et al., 2010)), thus illustrating the challenges of uncertainty identification in the general language domain.

Newswire text can prove more problematic in terms of uncertainty identification, since news stories tend to be reported in a subjective manner (Godbole et al., 2007; Vis, 2011) and allow for less strict use of language, while the truth value of reported events greatly depends on the time and context in which an article is written. As uncertainty identification is affected by various textual phenomena which are challenging to contextualise (metaphorical speech, colloquial expressions, etc), methods that identify event uncertainty from context are becoming increasingly crucial. The widespread use of the term "fake news" in recent years highlights the need to distinguish valuable and reliable facts, especially when it comes to automated information extraction. While detection of fake news is an involved process requiring more in depth discourse and stance analysis (Thorne et al., 2017), identifying certainty of extracted events is an important parameter towards the assessment of credibility of such events. The availability of an increasing number of resources annotated with news events and concepts related to uncertainty provide good opportunities to apply and adapt uncertainty identification techniques that are focussed on news articles.

In this work, we present our efforts on adapting uncertainty event extraction techniques developed for biomedical text, to allow them to be applied to newswire text. We use two corpora annotated with events and meta-knowledge (different types of interpretative information within a sentence that can affect an event (Thompson et al., 2011)) to analyse the differences between the two domains and we discuss the challenges that arise. We evaluate a hybrid machine learning approach to the identification of different uncertainty aspects (see Section 3.2.1) and propose ways of improving and customising uncertainty identification for newswire.

2 Related Work

In this section, we provide an overview of related work on uncertainty in both the scientific and newswire domains. We examine different classification frameworks of uncertainty and related concepts, the availability of annotations and existing classification systems used in each field.

The means of conveying uncertainty have long been studied by linguists, using a range of different terminology. Palmer (2001) introduced the term *epistemic modality* to refer to the degree of commitment to the truth of a proposition. The term continues to be used, especially for scientific text (De Waard and Maat, 2012; Vold, 2006) along with other related terms, such as *factuality*, which combines the notions of uncertainty and polarity (Saurí, 2017), *veracity* and *evidentiality* (Cornillie, 2009; Davis et al., 2007). The use of hedge words and their impact on the certainty of statements has also been studied extensively both in the scientific (Morante et al., 2010) and generic domain (Ganter and Strube, 2009). As computational technologies have evolved, there has been an increasing interest in the implications of textual uncertainty and the way it is expressed, resulting in a wide range of classification frameworks and annotation efforts.

In the scientific domain, Light (2004) studied uncertainty in biomedical papers, classifying expressions as denoting high or low certainty. Medlock and Briscoe (2007) further expanded the categorisation to incorporate the cases of admission of lack of knowledge, relays of hypotheses from others, speculative questions and hypotheses (investigation). More recently, Chen (2018) proposed a wider definition of uncertainty that covers phenomena of citation distortion, contradictions and claim inconsistencies, and also presented a method based on word embeddings for expanding a small seed list of cues to generate rich resources for uncertainty identification.

The aforementioned concepts have also been annotated in corpora at different levels of granularity. The BioScope corpus (Vincze et al., 2008), as well the biomedical part of the CoNLL 2010 task (Farkas et al., 2010) contain annotations of speculation and negation cues and their scope within the sentence. The BioNLP Shared Task corpora (Kim et al., 2009, 2011; Nédellec et al., 2013) also contain speculation and negation annotations, marked-up as attributes of events. The GENIA-MK corpus (Thompson et al., 2011) also contains event-level attribute annotations, but covering more meta-knowledge aspects, including *certainty level*, *polarity* and *knowledge type* (see Sec-

tion 3.2.1). Various models for the automated identification of the types of information annotated in the aforementioned corpora have been developed , with the best performing methods using a combination of rules and machine learning approaches. Overall, performance is highest for sentence-based annotations, with recent work reaching an F-score of 0.97 on BioScope (Kilicoglu et al., 2017), while on the event-level annotations of GENIA-MK, the best reported F-score surpasses 0.80 for the 3-level certainty classification problem (Miwa et al., 2012) and 0.88 for the binary problem (Zerva et al., 2017).

Bridging definitions of uncertainty across different domains, Szarvas (2012) proposes a hierarchical categorisation which distinguishes between two main classes: hypothetical and epistemic uncertainty. Vincze (2013), attempts a different categorization, looking at discourse-level uncertainty and related phenomena as they appear in text in the generic domain (Wikipedia). They identify three different types of uncertainty; weasels (relevant but insufficiently specified arguments), hedges and peacocks (exaggerated, subjective statements).

On work dealing with newspaper articles, subjectivity is identified as a further phenomenon (along with hedging and speculation) that is inextricably related to the expression of uncertainty (Rubin, 2007; Morante and Daelemans, 2009). Moreover, Rubin (2010) proposes a four-dimensional classification of certainty, also pointing out the aspect of timeliness and focus (abstract versus factual information). Their proposed annotation schema was applied to a small corpus of 82 documents. In terms of further resources, FactBank (Saurí and Pustejovsky, 2009) is a small corpus consisting of texts from the newswire domain annotated with events, accompanied with their factuality value (a combination of certainty level and polarity) judged from the viewpoint of their sources. The MPQA corpus (Cardie et al., 2003) elaborates on the issue of subjectivity and combines it with polarity markers to classify different opinions. The ACE 2005 corpus (Walker et al., 2006) contains events from news texts that are annotated with meta-knowledge attributes, among which *modality* and *genericity*. Subsequently, the meta-knowledge annotations were extended to include among others the aspect of *subjectivity* (see Section 3.2.1). More recently, there has been significant work in assessing factuality and credibility of news articles, as part of the fake-news challenge (FNC-I) that focusses on detection of stance.

In comparison to the scientific domain, there have been relatively fewer attempts to automatically identify uncertainty in news text, apart from the classification of particular aspects that embody uncertainty, such as subjectivity (Wilson, 2008). The most significant work is the wikipedia related task of CoNLL 2010, which concerned weasel cue detection. The best performing systems at the time compared poorly to the results in the biomedical field but more recently Jean (2016) proposed a probabilistic model that achieved an F-score of 55.7, showing a promising degree of improvement. Even more encouragingly, there have recently been important efforts on the classification of factuality values based on FactBank and related factuality corpora (UW, MEANTIME), showing great improvements in their predictions (Stanovsky et al., 2017; Lee et al., 2015) compared to earlier attempts (Prabhakaran et al., 2010). Such efforts motivate our interest in studying the detection of uncertainty in the newswire domain.

3 Methods

In this section, we provide a definition of the problem we aim to tackle, as well as definitions of terms that we use subsequently. We also describe the datasets and resources that we have used, and we present the methods and technical details used for the experiments and analysis in Section 4.

3.1 Event Definition

In both the GENIA-MK and the ACE-MK corpora, the definition of events shares some core properties. An event consists necessarily of one trigger entity and usually one or more participant NEs (arguments) that are linked to the trigger. The trigger entity determines the type of the event, and is usually one word (can be verb, noun or adjective) that describes the event. Similarly, the relation between the trigger and each argument determines argument's role. Examples of events from the two domains are presented in Figure 1.

3.2 Uncertainty Identification Task

As described in the previous section, uncertainty can be interpreted in different ways. In this work, we cast uncertainty identification as the task of identifying textual information (cues) that render the truth of a specific *event* uncertain. Hence,

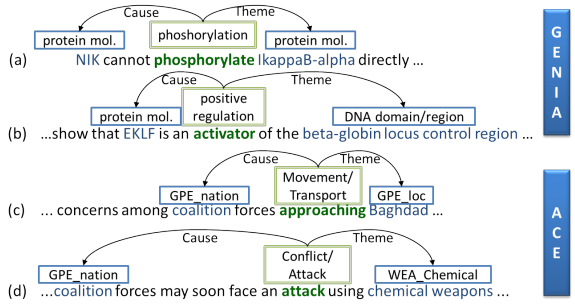

Figure 1: Event examples extracted from GENIA-MK (a-b) and ACE-MK (c-d).

uncertainty is treated as an attribute of an event, rather than an attribute of a sentence or clause. This is because it has been shown that a given unit of text may contain more than one event, each with a potentially different level of uncertainty (Saurí and Pustejovsky, 2009; Thompson et al., 2017). We limit the discovery of uncertainty cues to those occurring in same sentence as the event in question, following the annotations of the two corpora.

We cast uncertainty identification as a binary classification task, where an event can either be *certain* or *uncertain*. Our decision was motivated by the findings of Rubin (2007) who showed that a finer grained classification of uncertainty (5 levels) resulted in unacceptably low levels of inter-annotator agreement.

We treat uncertainty of an event as an attribute that can be affected by various factors (modality, hypothesis, subjectivity etc), that are already annotated in existing corpora. Hence, we want to take advantage of existing corpora annotations, and examine how such annotations relate to uncertainty, either individually or combined. We examine the performance and robustness of automated uncertainty identification method developed in (Zerva et al., 2017) based on different combinations of meta-knowledge dimensions to draw our conclusions, acknowledging that (as discussed in Section 2) for different domains there can be different dimensions affecting uncertainty. In the following section, we describe the datasets as well the meta-knowledge annotations that we consider to be related to uncertainty identification in the biomedical and newswire domains.

3.2.1 Datasets and Uncertainty

We focus our analysis for the newswire domain on the recent annotations of the ACE 2005 corpus (Walker et al., 2006) (English version). The corpus was originally annotated with named entities (NEs), events, as well as some meta-knowledge information and has been subsequently enriched with additional meta-knowledge annotations (Thompson et al., 2017). We refer to the meta-knowledge annotated version of the corpus as ACE-MK [1]. The corpus comprises of 600 news articles originating from various sources, and contains annotations for 5349 events. The ACE-MK meta-knowledge annotation scheme, includes 6 meta-knowledge attributes, of which four (4) were present in the original 2005 annotated corpus and the rest were introduced in the 2017 annotation enrichment effort (the latter are marked with an asterisk in the enumeration that follows). The respective cues for each type were annotated whenever present within a sentence.

1. Subjectivity (*) towards the event by the source. Can be Positive, Negative, Neutral or Multi-valued (two or more sources expressing opposite sentiments for the same event).
2. Source (*), that can be Author, Involved (attributed to a specified source, somehow involved with the event) or Third-Party.
3. Modality, that can have four possible values; Asserted, Speculated, Presupposed(*) and Other
4. Polarity, that can be either Positive or Negative.
5. Tense, that can be Past, Present, Future or Unspecified.
6. Genericity, that can either be Specific (event referring to a specific occurrence) or Generic.

As discussed in Section 2, various concepts, such as modality, subjectivity, genericity and timeliness have been linked to uncertainty in the newswire domain. In fact, most of the aforementioned event attributes annotated in ACE-MK could affect event certainty. In this work, we focus on the dimensions of *Modality*, *Genericity* and *Subjectivity*. (Saurí and Pustejovsky, 2009) Considering these three different attributes as well as their combination as uncertainty indicators, we generate four different test-sets, each corresponding to a different uncertainty definition:

[1]The ACE-MK corpus annotations and guidelines are available at http://www.nactem.ac.uk/ace-mk/ .

1. M: uncertainty corresponds only to *Modality*, and only *Asserted* events are equivalent to *Certain*. Based on descriptions in (Baker et al., 2014; Szarvas et al., 2012).

2. G: uncertainty corresponds only to *Genericity*, and only *Specific* events are equivalent to *Certain*. We thus claim that that generic, more vague events lack certainty, inspired by the distinction between abstract and specific statements in (Rubin, 2010).

3. S: uncertainty corresponds only to *Subjectivity*, and only *Neutral* events are equivalent to *Certain*. Based on (Wiebe and Riloff, 2005) which has shown that positive or negative bias can affect the certainty of an event. *Multi-valued* instances are treated as *Uncertain* since contradictory assertions have also been linked to uncertainty (Alamri, 2016).

4. MGS : uncertainty corresponds to the union of the above; only an event that is *Asserted*, *Neutral* and *Specific* is considered *Certain*.

In both corpora, the annotations of all meta-knowledge dimensions are on the event level (the values of each event annotated separately). The evidence, if it can be attributed to one or more words in the same sentence as the event, is annotated as a cue, for the dimension annotated, and linked to the event(s) that it affects. In Figure 2 (a-b) we demonstrate one example from each corpus where the cue affects only one of the events in a sentence. While in both corpora for most dimensions investigated the cues are word sequences different than the trigger of the event, for Subjectivity, we have cases where the trigger is also acting like a Subjectivity cue. This is because based on the definition of Subjectivity for ACE-MK, biased attitude expressed in text denotes subjectivity (including expressions of intention, command, fear, hope, condemn etc). Example (c) in Figure 2 demonstrates such a case.

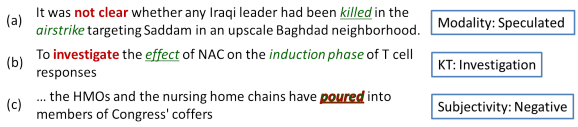

Figure 2: Examples of cue annotations. Cues are in bold-red while events in green-italic. Events that are affected by the highlighted cue are underlined in each sentence.

We train and test separate classifiers for each case and discuss their performance and the implication on the predictability of uncertainty.

We should note that *Polarity* has been identified as a dimension that is orthogonal to uncertainty (Saurí and Pustejovsky, 2009) and thus we choose not to include it in our investigation, although both corpora contain such annotations. In future work, we would like to further investigate the combination of certainty and polarity and maybe expand our analysis on the FactBank corpus. It would also be interesting, as future work, to expand our experiments and investigate whether *Tense* could also be used to account for the timeliness aspect, or whether *Source* could help to identify weaselling phenomena, thus expanding the coverage of uncertainty. For an efficient accounting of these two dimensions in future work, we would like to include additional resources such as timeliness or citation analysis components.

Apart from comparing performance among the different uncertainty-related definitions described above, we compare our results for ACE-MK with those obtained for a biomedical corpus, GENIA-MK (Kim et al., 2003; Thompson et al., 2011), for binary uncertainty identification using the same hybrid method, as reported in (Zerva et al., 2017).

The GENIA-MK corpus consists of 1000 abstracts extracted from PubMed and annotated with 36,858 events[2]. It has also been annotated with meta-knowledge attributes for each event, and the respective cues. The meta-knowledge attributes for each event include *Certainty Level* (L1, L2, L3), *Polarity* (Positive, Negative), *Manner* (High, Low and Neutral), *Source* (Current, Other) and *Knowledge Type* (Investigation, Observation, Analysis, Method, Fact, Other). Of those, *Certainty Level* L1 and L2 as well as *Knowledge type* of Investigation were treated as uncertainty indicators (denoting an event as *Uncertain*).

3.3 Machine Learning Approach

For the experiments described in Section 4.1 we use a hybrid machine learning approach to classify ACE-MK events as *Certain* or *Uncertain*. We use a Random Forest (RF) classifier (Liaw et al., 2002) and a range of semantic, lexical, syntactic and dependency features. The majority of the lexical features are related to the cue and its sur-

[2]The GENIA-MK annotations are available at: http://www.nactem.ac.uk/meta-knowledge/ .

face and grammatical properties, while syntactic and dependency features are related to the syntactic dependencies between the cue and the event. Features also include dependency-based rules that capture one and two-hop paths between the cue and an event trigger. Finally, there is an additional set of features related to the semantics of the event itself (event type, arguments). A more detailed description and examples of the features can be found in Appendix A.

The full processing of ACE-MK corpus, including other NLP tasks such as sentence splitting, tokenisation etc, was performed using Argo platform, a web-based, graphical workbench that facilitates the construction and execution of modular text mining workflows (Batista-Navarro et al., 2017). For the implementation of the RF classifier, dedicated components were implemented using the WEKA API (Frank et al., 2004). We used 10-fold cross-validation to evaluate and compare the performance of different generated models. Since some of the features are sentence and/or document based, we avoided the automated 10-fold cross validation of the WEKA API, and instead modified the random fold generation so that no document would be split over several folds, thus ensuring the models were not biased or overfitted to specific documents.

3.4 WordNet-based Analysis

In order to interpret the differences in the performance of our models between the GENIA-MK and the ACE-MK, we compared the lexical and semantic properties of the cues in each corpus. For this purpose, we used WordNet (Miller, 1995) version 3.0 to examine the synsets and relations between uncertainty cues, the generated word graphs and the distributions of cues per synset. To process cues against information contained within WordNet, the JWI API (Finlayson, 2014) was used.

In order to study the links between cues, we consider WordNet as a multi-graph where each word is a node, and all potential relations between two words constitute an edge. The types of relations are used as edge attributes. To generate the graph from each corpus, we start with the lemmatised cues and iteratively expand the graph using a set of available relations between words as well as synsets until there are no other nodes to visit. We use all relations available in WordNet between synsets and words, but we exclude expansion for some senses that are semantically irrelevant to all potential cues, as described in Appendix B.

The analysis and visualisation of the graphs was performed using Gephi (Bastian et al., 2009).

4 Results and Discussion

4.1 Automated Classification of Uncertainty

As a first step, we used the set of cues extracted from GENIA-MK for the generation of all features in the cue and dependency related feature sets. We then trained and evaluated the performance of the trained models on each of the test sets of the ACE-MK corpus, as shown in the top three rows of Table 1. The results show that the classifier trained with GENIA-MK cues does not achieve particularly high performance for any of the three cases of uncertainty, or for their combination. We subsequently proceeded to replace the GENIA-MK cues with the ones extracted from the ACE-MK corpus, and repeated the experiments, as shown in the bottom three rows Table 1.

When using ACE-MK cues, F-score increases significantly ($p < 0.01$) for all different test sets. This is mostly due to the consistent improvement in recall for all test sets (in terms of precision, it is only the case of *Modality* that the ACE-MK cues outperform the GENIA-MK cues). This result confirms the domain dependence of uncertainty expressions and stresses the need of domain specific approaches, to achieve higher performance.

	M	G	S	MGS	Cues
Precision	0.53	0.27	0.40	0.61	
Recall	0.55	0.62	0.46	0.69	GEN
F-score	0.54	0.38	0.34	0.65	
Precision	0.57	0.26	0.40	**0.69**	
Recall	0.69	0.67	0.63	**0.74**	ACE
F-score	0.62	0.37	0.49	**0.71**	

Table 1: Performance of uncertainty identification on each uncertainty test-case using GENIA-MK (GEN) and ACE-MK (ACE) cues.

	GENIA-MK cues	ACE-MK cues
Precision	0.94	0.82
Recall	0.83	0.86
F-score	0.88	0.84

Table 2: Performance for uncertainty identification on GENIA-MK corpus using different cues.

More interestingly however, we notice that even when using ACE-MK cues, the performance we obtain is significantly lower compared to the performance obtained when the same method is applied to the GENIA-MK corpus. Indeed we see in Table 2 that on GENIA-MK even when using cues extracted from ACE-MK, performance is significantly higher for all metrics (Zerva et al., 2017).

Genericity seems to be the hardest attribute to distinguish, especially in terms of precision. This can be explained through an examination of the training data, which reveals that there are very few Generic event instances that are linked to a Genericity cue. Thus, while there is a sufficient number of training instances for Generic events (1132 Generic versus 4217 Specific) strong feature vectors can only be produced for a few of them. The classifier also seems to be having difficulties in predicting *Subjectivity*, but for different reasons. Looking more closely at the results for *Subjectivity*, we discovered that one issue relates to Multi-valued test cases, which are particularly complex since they often involve the existence of more than one Subjectivity cue linked with the event, and at the same time they are significantly under-sampled (18 instances). Moreover, Subjectivity cues seem to involve more nouns and longer, often colloquial expressions compared to other dimensions.

Further enhancement of the machine learning approach and feature engineering could try to address such issues, in order to better identify *Subjectivity* and *Genericity* dimensions. A possible future direction would be to enhance current vectors methods that can account for positive or negative bias of nouns, or other methods borrowed by work on subjectivity. Coupled with a training corpus containing more positive instances, such methods could help drawing further conclusions.

In the last column of Table 1 we present the performance of the models trained on the combined dimensions. By combining the meta-knowledge dimensions into one uncertainty identification task, we can see that we get improved performance, compared to the individual tasks. This provides an indication that relationships exist between these different dimensions in the context of detecting uncertainty. Still, as mentioned earlier, we notice that for all possible combinations, performance is lower compared to results reported for biomedical corpora using the same machine learning approach, even when we use cues extracted from the same corpus. This difference in score, even in the case of Modality, much like the one seen in the work of (Tang et al., 2010) for the CoNLL datasets, provides motivation to look more closely into the differences between the means of expressing uncertainty in the two different domains. In the next section, we attempt to interpret this difference in performance, explore why the cue and dependency based features used might be less effective for the newswire domain, and what could be done to remedy this.

4.2 Comparison of the Properties of Uncertainty Cues Between Corpora

4.2.1 Dependency-based Comparison

As mentioned in Section 3.2.1 the machine learning classifiers used in this work, are heavily dependent on features related to the dependencies between potential uncertainty cues and the triggers of events. For the extraction of dependency paths we use a dependency parser in order to extract the dependency relations for each sentence of the corpus. The Enju dependency parser (Miyao et al., 2008) was used for both corpora, with models trained on biomedical and newswire data for GENIA-MK and ACE-MK respectively.

We then treat the dependencies as a directed graph and examine the shortest paths between annotated cues and event triggers as shown in the example of Figure 4. In case of multi-word cues or multi-word events we consider the shortest possible path between any word of the cue and any word of the trigger. The comparison of dependency path lengths for the two corpora can be seen in Figure 3.

It is clear from the distribution that the dependency paths for the GENIA-MK corpus (gray-

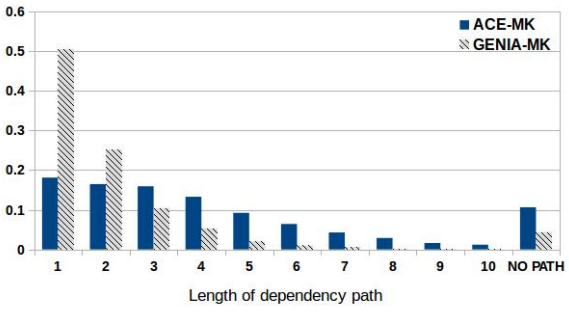

Figure 3: Histogram of length distribution for shortest dependency paths between uncertainty cues and triggers for ACE-MK and GENIA-MK.

striped bars) follows a long-tail pattern, with more than 50% of the cues being directly linked to the trigger and more than 85% being at a distance of three or less dependency links. On the contrary for ACE-MK corpus we have a more evenly spread distribution of dependency paths, since to contain 85% of the cases we need to reach dependency paths of length 7. Looking at the last bar of the histogram, which accounts for paths longer than ten (10) hops or non-existent paths, we note that the percentage of such cases is double for ACE-MK compared to GENIA-MK.

This difference in the dependency path distribution, could explain why features based on dependency paths as well as dependency rules are not as efficient for newswire documents. Indeed, analysis of feature informativeness (using Mutual Information measures (Battiti, 1994)) for the two corpora further supports these observations. In the 30 top scoring features for GENIA-MK, 19 are dependency features (14 of them dependency rules) versus only 5 dependency features for ACE-MK (and only 1 dependency rule). These observations reveal a potential higher complexity in the sentence syntax and language structure in newswire texts as opposed to scientific texts. For example, in ACE-MK we observe more occurrences of event triggers being nouns that are not close to the main verb (and surrounding modals) and of cues indicating uncertainty (especially Subjectivity) found in a different sub-phrase than the event (see Figure 3). There are also some wrongly structured sentences where the dependency paths are distorted due to problematic syntax.

This difference may occur as a result of the

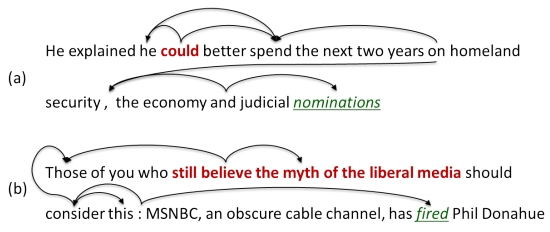

Figure 4: Dependency paths between cue (red-bold) and trigger (green-underlined) for ACE-MK. Arrows denote the edges of the dependency graph that participate in the shortest path between cue and trigger. In (a) *could* is a Modality cue, influencing a *Personel_nominates* event. In (b) we have a phrase that is annotated as a Subjectivity cue and the event is *Personnel_end_position*.

greater freedom of expression in news articles as opposed to scientific texts, where language and syntax follow stricter rules, and formal expressions are preferred to colloquial ones. Although it has be shown that even in scientific text, many statements are far from factual assertions, we can expect phenomena of vagueness, weaselling, hedging and speculating to be much more prevalent in news articles compared to scientific ones. It should though be noted that this difference might be further aggravated by the fact that GENIA-MK consists of abstracts, where requirements for precise language are even stricter.

4.2.2 Lexical Comparison

It seems that it is not only in syntax that the two corpora and respective domains differ. By focussing on the lexical and semantic properties of the cue lists in each case, we also found a set of differences at this level. A simple initial observation concerns the differences between the lengths of cues, in terms of the number of words, between the two domains. We can see in Figure 5 that in GENIA-MK, with the exception of some very lengthy outliers, most of the cues are one or two word expressions. In contrast, ACE-MK contains more lengthy uncertainty expressions, including various colloquial expressions, weasels etc.

We also examined the semantic properties of the two cue-lists and generated two WordNet graphs for each corpus as described in Section 3.4. Apart from the sense limitation mentioned before, there was no further attempt to disambiguate cues that belonged to more than one synset. Instead, all possible synsets for each word were added to the graph ending, resulting in a total of 781 synsets covered by the cues for GENIA-MK, compared to 1444 synsets for ACE-MK. Thus the cues in ACE-

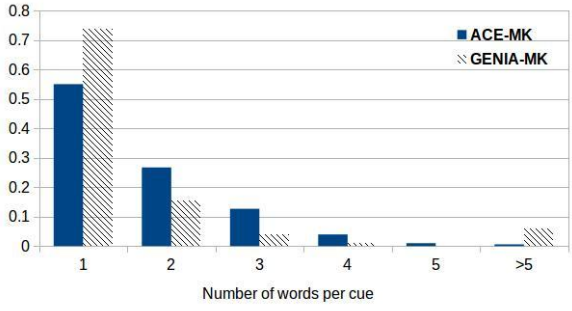

Figure 5: Histogram of words per cue distribution for ACE-MK and GENIA-MK corpora.

MK seems to have a far broader semantic coverage, which means much greater lexical variability and harder to predict cues. To generate the graphs, we use the words in the cue list as seed nodes and then expand them to include all 1-hop neighbors and corresponding edges for each cue. We end up with a graph of 4293 nodes for GENIA-MK and 6123 nodes for ACE-MK.

Looking at the connectivity properties of the two graphs and the number of fully connected components (sub-graphs), we notice that the GENIA-MK graph has only two fully connected sub graphs, versus fifteen (15) for ACE-MK. The difference in sub graphs is another indication supporting the difference in semantic range for the two corpora, although it should be noted that for both corpora 85% of the nodes is contained in the largest sub graph.

We then proceeded to carry out modularity based community detection for the two graphs (Newman, 2006) in order to identify and visualise patterns in the senses of each graph. We focussed on the first 10 largest communities (size calculated on the basis of node count) and their central nodes. To identify central nodes, we ranked nodes using three different centrality measures: betweenness, closeness (Brandes, 2001) and eccentricity (Hage and Harary, 1995) and then used the intersection of the top ranked nodes for each measure. We provide the visualisation of the graphs in Appendix C. As expected, in both graphs the communities are semantically related, and it is easy to see that in some communities the central nodes are related to uncertainty (likelihood, probability etc). Some of the communities evolve around similar concepts, such as ability, probability, communication and investigation, although the concepts are expressed using different terms.

It is important to note that using only 1-hop expansion of the original cues gathered from the two corpora, we were able to generate a graph with semantically meaningful communities. Hence, it would be interesting to further explore the use of WordNet and other semantic graphs as an unsupervised way to expand cue lists and use them on previously unseen data. This could prove particularly useful for domains lacking annotated resources.

5 Conclusion

In this paper we have analysed uncertainty identification in the newswire domain and compared it with the scientific (biomedical) domain both in terms of uncertainty definition and performance of methods. We have explored different meta-knowledge aspects available in newswire corpora, in terms of their relation to uncertainty and the feasibility of their automated identification in text.

We have shown that it is possible to transfer methods similar to the ones employed in the biomedical domain for the automated identification of uncertain events in the news text. However we found that regardless of whether detecting uncertainty is restricted to individual dimensions, or they are treated as a combined task, the performance is significantly lower than the performance obtained by applying the same methods to biomedical articles. To try to understand reasons for this difference, we have analysed the syntactic and lexical properties of textual uncertainty in the newswire domain, and have discovered a number of factors that render the task of uncertainty identification more difficult to tackle in newswire documents. Our analysis has highlighted the role of longer dependencies between cues and events as one of the main issues that complicate the task in newswire articles, along with lengthy cues with increased semantic variability.

We consider this work a promising first step towards a more detailed and fine-tuned approach to uncertainty identification in the newswire domain. As future work, we aim to take advantage of our findings regarding the syntactic and lexical properties that were highlighted above, in order to build more robust classifiers. Moreover, we would like to expand our analysis of uncertainty in the newswire domain using word-embeddings and potentially expand the uncertainty definition in a similar fashion to (Chen et al., 2018). To support this goal, we also intend to experiment with further corpora in the newswire domain.

Efficient uncertainty identification will provide a useful tool for a more meaningful and semantically interpretable information extraction.

Acknowledgments

We would like to thank Mr. Paul Thompson for his invaluable comments that greatly improved the manuscript. This work has been supported by the Engineering and Physical Sciences Research Council [Grant: EP/1038099/1 (CDT)]; and the Biotechnology and Biological Sciences Research Council [Grants: BB/M006891/1 (EMPATHY)].

References

Abdulaziz Alamri. 2016. *The Detection of Contradictory Claims in Biomedical Abstracts*. Ph.D. thesis, University of Sheffield.

Kathryn Baker, Michael Bloodgood, Bonnie J Dorr, Nathaniel W Filardo, Lori Levin, and Christine Piatko. 2014. A modality lexicon and its use in automatic tagging. *arXiv preprint arXiv:1410.4868*.

Mathieu Bastian, Sebastien Heymann, Mathieu Jacomy, et al. 2009. Gephi: an open source software for exploring and manipulating networks. *Icwsm*, 8:361–362.

Riza Batista-Navarro, Nhung TH Nguyen, Axel J Soto, William Ulate, and Sophia Ananiadou. 2017. Argo as a platform for integrating distinct biodiversity analytics tools into workflows for building graph databases. *Proceedings of TDWG*, 1:e20067.

Roberto Battiti. 1994. Using mutual information for selecting features in supervised neural net learning. *IEEE Transactions on neural networks*, 5(4):537–550.

Ulrik Brandes. 2001. A faster algorithm for betweenness centrality. *Journal of mathematical sociology*, 25(2):163–177.

Kai Cao, Xiang Li, and Ralph Grishman. 2016. Leveraging dependency regularization for event extraction. In *FLAIRS Conference*, pages 20–25.

Claire Cardie, Janyce Wiebe, Theresa Wilson, and Diane J Litman. 2003. Combining low-level and summary representations of opinions for multi-perspective question answering. In *New directions in question answering*, pages 20–27.

Chaomei Chen, Min Song, and Go Eun Heo. 2018. A scalable and adaptive method for finding semantically equivalent cue words of uncertainty. *Journal of Informetrics*, 12(1):158–180.

Bert Cornillie. 2009. Evidentiality and epistemic modality: On the close relationship between two different categories. *Functions of language*, 16(1):44–62.

Christopher Davis, Christopher Potts, and Margaret Speas. 2007. The pragmatic values of evidential sentences. In *Semantics and Linguistic Theory*, volume 17, pages 71–88.

Anita De Waard and Henk Pander Maat. 2012. Epistemic modality and knowledge attribution in scientific discourse: A taxonomy of types and overview of features. In *Proceedings of the Workshop on Detecting Structure in Scholarly Discourse*, pages 47–55. Association for Computational Linguistics.

Richárd Farkas, Veronika Vincze, György Móra, János Csirik, and György Szarvas. 2010. The conll-2010 shared task: learning to detect hedges and their scope in natural language text. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning—Shared Task*, pages 1–12. Association for Computational Linguistics.

Mark Finlayson. 2014. Java libraries for accessing the princeton wordnet: Comparison and evaluation. In *Proceedings of the Seventh Global Wordnet Conference*, pages 78–85.

FNC-I. Fake news challenge. http://www.fakenewschallenge.org.

Eibe Frank, Mark Hall, Len Trigg, Geoffrey Holmes, and Ian H Witten. 2004. Data mining in bioinformatics using weka. *Bioinformatics*, 20(15):2479–2481.

Viola Ganter and Michael Strube. 2009. Finding hedges by chasing weasels: Hedge detection using wikipedia tags and shallow linguistic features. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 173–176. Association for Computational Linguistics.

Namrata Godbole, Manja Srinivasaiah, and Steven Skiena. 2007. Large-scale sentiment analysis for news and blogs. *Icwsm*, 7(21):219–222.

Per Hage and Frank Harary. 1995. Eccentricity and centrality in networks. *Social networks*, 17(1):57–63.

Ken Hyland. 1998. *Hedging in scientific research articles*, volume 54. John Benjamins Publishing.

Pierre-Antoine Jean, Sébastien Harispe, Sylvie Ranwez, Patrice Bellot, and Jacky Montmain. 2016. Uncertainty detection in natural language: A probabilistic model. In *Proceedings of the 6th International Conference on Web Intelligence, Mining and Semantics*, page 10. ACM.

Halil Kilicoglu, Graciela Rosemblat, and Thomas C Rindflesch. 2017. Assigning factuality values to semantic relations extracted from biomedical research literature. *PloS one*, 12(7):e0179926.

J-D Kim, Tomoko Ohta, Yuka Tateisi, and Junichi Tsujii. 2003. Genia corpusa semantically annotated corpus for bio-textmining. *Bioinformatics*, 19(suppl_1):i180–i182.

Jin-Dong Kim, Tomoko Ohta, Sampo Pyysalo, Yoshinobu Kano, and Jun'ichi Tsujii. 2009. Overview of bionlp'09 shared task on event extraction. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing: Shared Task*, pages 1–9. Association for Computational Linguistics.

Jin-Dong Kim, Sampo Pyysalo, Tomoko Ohta, Robert Bossy, Ngan Nguyen, and Jun'ichi Tsujii. 2011. Overview of bionlp shared task 2011. In *Proceedings of the BioNLP shared task 2011 workshop*, pages 1–6. Association for Computational Linguistics.

Kenton Lee, Yoav Artzi, Yejin Choi, and Luke Zettlemoyer. 2015. Event detection and factuality assessment with non-expert supervision. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1643–1648.

Andy Liaw, Matthew Wiener, et al. 2002. Classification and regression by randomforest. *R news*, 2(3):18–22.

Marc Light, Xin Ying Qiu, and Padmini Srinivasan. 2004. The language of bioscience: Facts, speculations, and statements in between. In *HLT-NAACL 2004 Workshop: Linking Biological Literature, Ontologies and Databases*.

Teun Lucassen and Jan Maarten Schraagen. 2010. Trust in wikipedia: how users trust information from an unknown source. In *Proceedings of the 4th workshop on Information credibility*, pages 19–26. ACM.

Ashutosh Malhotra, Erfan Younesi, Harsha Gurulingappa, and Martin Hofmann-Apitius. 2013. hypothesisfinder:a strategy for the detection of speculative statements in scientific text. *PLoS computational biology*, 9(7):e1003117.

David McClosky, Mihai Surdeanu, and Christopher D Manning. 2011. Event extraction as dependency parsing. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 1626–1635. Association for Computational Linguistics.

Ben Medlock and Ted Briscoe. 2007. Weakly supervised learning for hedge classification in scientific literature. In *Proceedings of the 45th annual meeting of the association of computational linguistics*, pages 992–999.

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

Makoto Miwa, Paul Thompson, John McNaught, Douglas B Kell, and Sophia Ananiadou. 2012. Extracting semantically enriched events from biomedical literature. *BMC bioinformatics*, 13(1):108.

Yusuke Miyao, Rune Sætre, Kenji Sagae, Takuya Matsuzaki, and Jun'ichi Tsujii. 2008. Task-oriented evaluation of syntactic parsers and their representations. *Proceedings of ACL-08: HLT*, pages 46–54.

Roser Morante and Walter Daelemans. 2009. Learning the scope of hedge cues in biomedical texts. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*, BioNLP '09, pages 28–36, Stroudsburg, PA, USA. Association for Computational Linguistics.

Roser Morante, Vincent Van Asch, and Walter Daelemans. 2010. Memory-based resolution of insentence scopes of hedge cues. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning—Shared Task*, pages 40–47. Association for Computational Linguistics.

Claire Nédellec, Robert Bossy, Jin-Dong Kim, Jung-Jae Kim, Tomoko Ohta, Sampo Pyysalo, and Pierre Zweigenbaum. 2013. Overview of bionlp shared task 2013. In *Proceedings of the BioNLP Shared Task 2013 Workshop*, pages 1–7.

Mark EJ Newman. 2006. Modularity and community structure in networks. *Proceedings of the national academy of sciences*, 103(23):8577–8582.

Thien Huu Nguyen, Kyunghyun Cho, and Ralph Grishman. 2016. Joint event extraction via recurrent neural networks. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 300–309.

Frank Robert Palmer. 2001. *Mood and modality*. Cambridge University Press.

Vinodkumar Prabhakaran, Owen Rambow, and Mona Diab. 2010. Automatic committed belief tagging. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 1014–1022. Association for Computational Linguistics.

James Pustejovsky. 2017. Iso-timeml and the annotation of temporal information. In *Handbook of Linguistic Annotation*, pages 941–968. Springer.

Victoria L Rubin. 2007. Stating with certainty or stating with doubt: Intercoder reliability results for manual annotation of epistemically modalized statements. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers*, pages 141–144. Association for Computational Linguistics.

Victoria L Rubin. 2010. Epistemic modality: From uncertainty to certainty in the context of information seeking as interactions with texts. *Information Processing & Management*, 46(5):533–540.

Roser Saurí. 2017. Building factbank or how to annotate event factuality one step at a time. In *Handbook of Linguistic Annotation*, pages 905–939. Springer.

Roser Sauri and James Pustejovsky. 2007. Determining modality and factuality for text entailment. In *Semantic Computing, 2007. ICSC 2007. International Conference on*, pages 509–516. IEEE.

Roser Saurí and James Pustejovsky. 2009. Factbank: a corpus annotated with event factuality. *Language resources and evaluation*, 43(3):227.

Gabriel Stanovsky, Judith Eckle-Kohler, Yevgeniy Puzikov, Ido Dagan, and Iryna Gurevych. 2017. Integrating deep linguistic features in factuality prediction over unified datasets. In *Proceedings of the*

55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), volume 2, pages 352–357.

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.

Buzhou Tang, Xiaolong Wang, Xuan Wang, Bo Yuan, and Shixi Fan. 2010. A cascade method for detecting hedges and their scope in natural language text. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning—Shared Task*, pages 13–17. Association for Computational Linguistics.

Paul Thompson, Raheel Nawaz, John McNaught, and Sophia Ananiadou. 2011. Enriching a biomedical event corpus with meta-knowledge annotation. *BMC bioinformatics*, 12(1):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.

James Thorne, Mingjie Chen, Giorgos Myrianthous, Jiashu Pu, Xiaoxuan Wang, and Andreas Vlachos. 2017. Fake news stance detection using stacked ensemble of classifiers. In *Proceedings of the 2017 EMNLP Workshop: Natural Language Processing meets Journalism*, pages 80–83.

Veronika Vincze. 2013. Weasels, hedges and peacocks: Discourse-level uncertainty in wikipedia articles.

Veronika Vincze, György Szarvas, Richárd Farkas, György Móra, and János Csirik. 2008. The bioscope corpus: biomedical texts annotated for uncertainty, negation and their scopes. *BMC bioinformatics*, 9(11):S9.

Kirsten Vis. 2011. Subjectivity in news discourse: A corpus linguistic analysis of informalization.

Eva Thue Vold. 2006. Epistemic modality markers in research articles: a cross-linguistic and cross-disciplinary study. *International Journal of Applied Linguistics*, 16(1):61–87.

Christopher Walker, Stephanie Strassel, Julie Medero, and Kazuaki Maeda. 2006. Ace 2005 multilingual training corpus. *Linguistic Data Consortium, Philadelphia*, 57.

Janyce Wiebe and Ellen Riloff. 2005. Creating subjective and objective sentence classifiers from unannotated texts. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 486–497. Springer.

Theresa Ann Wilson. 2008. *Fine-grained subjectivity and sentiment analysis: recognizing the intensity, polarity, and attitudes of private states*. University of Pittsburgh.

Chrysoula Zerva, Riza Batista-Navarro, Philip Day, and Sophia Ananiadou. 2017. Using uncertainty to link and rank evidence from biomedical literature for model curation. *Bioinformatics*, 33(23):3784–3792.

A Appendix A: Machine Learning Features

The features presented in Table 3 are used in the models that were generated for all the experiments presented in Section 4 of the main part of the article. The table presents the main feature categories that are extracted for each event (columns 1 and 2), providing a brief description (column 3) and feature type (column 3).

We should note that for the GENIA-MK corpus, features analysis showed that the contribution of lexical features for cues is overshadowed by the dependency rule features, that capture a combination of surface for and dependencies. On the contrary, such features are more informative for the case of ACE-MK uncertainty classification, since as we have shown in the main document, dependency paths are often longer in ACE-MK rendering the dependency rules inefficient in capturing such relations. Moreover, lexical features for events score very high in terms of informativeness in ACE-MK and quite low in GENIA-MK. This could be attributed to the more uniform type of events in GENIA-MK.

We note that for constituency features, command of a word *a* over a word *b*, signifies that in the syntactic tree *a* is the head of a branch that contains *b*. In both corpora, constituency features scored very high in terms of informativeness.

For dependency path rules, features capture the dependency path as a chain of words (lemmatised[3]) and the type of dependency edges between them. For the experiments presented in this work (Section 4.1 of the main part of the article), rules spanning up to 2 consecutive edges were used (1-hop and 2-hop rules). In Figure 6 we present an example of rule extraction from a sentence. The sentence contains one *Modality* cue (would stipulate) and one *Subjectivity* cue (hates). All the paths between any word of each cue and the the event trigger (war) is extracted based on the dependencies (shown above the sentence). Subsequently, all paths that have length equal or shorter than 2 are converted to rules, as shown below the sentence.

[3]Stanford lemmatiser from the CoreNLP toolkit and Enju parser were used for lemmatisation in all features that required lemmas and/or surface forms of words.

In the case of 2-hop rules, the lemma of the word between the cue and the event trigger in the path, is also captured as part of the rule (as shown in the Modality rule of Figure 6).

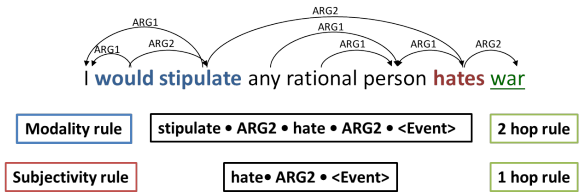

Figure 6: Example of dependency based rule extraction for a phrase extracted from ACE-MK.

B Appendix B: WordNet Senses

When using Wordnet for the graph generation we excluded some of the lexicographic sense groups that are available in WordNet, since they were judged to be too distant to uncertainty expressions (eg referring to specific objects etc). The choice was guided by the description of each sense, in order to avoid senses that do not relate to any of the dimensions of uncertainty described in the main document. By thus excluding senses related to concepts such as food, countries, activities etc we achieve reduced complexity, size and processing time of the resulting graphs. Nevertheless, inclusion of such senses could be interesting to consider in future experiments to see if they can better account for metaphors and colloquial expressions. Alternatively, graphs generated by word embedding approaches could be studied and compared against the WordNet ones.

We list the inclusion/exclusion decision for each of the senses in the Table 4, along with the description of the lexicographer file according to WordNet documentation (https://wordnet.princeton.edu/documentation/lexnames5wn).

Cat.	Sub-cat.	Feature	Output
Event	Lexical	Event-trigger surface form	Nom.
		POS tags	Nom.
	Semantic	Event type	Nom.
		Argument type	Nom.
		Argument role	Nom.
	Complexity	Complex/simple	Bin.
Cue	Lexical	Existence of cue	Bin.
		Cue surface form	Nom.
		POS tag of the cue	Nom.
Event & Cue	Relative position	#words between cue and event trigger	Num.
		Position of cue on the left/right of the event trigger	Bin
	Dependency	Direct dependency between cue and trigger	Bin.
		Shortest dependency path length	Num.
		Existence of dependency path rule (see example)	Bin.
		Dependency path rule (see example)	Nom.
	Constituency (syntactic)	Command of cue over trigger	Bin.
		Command of cue over arguments	Bin.

Table 3: Features used for uncertainty identification with the RF classifier. The output column shows the type of the generated feature; *Nom.* denotes nominal features, *Bin.* denotes binary features and *Num.* denotes numeric features.

#	Name	Description	Incl/Excl
0	adj.all	all adjective clusters	Included
1	adj.pert	relational adjectives (pertainyms)	Included
2	adv.all	all adverbs	Included
3	noun.Tops	unique beginner for nouns	Excluded
4	noun.act	nouns denoting acts or actions	Included
5	noun.animal	nouns denoting animals	Excluded
6	noun.artifact	nouns denoting man-made objects	Excluded
7	noun.attribute	nouns denoting attributes of people and objects	Included
8	noun.body	nouns denoting body parts	Excluded
9	noun.cognition	nouns denoting cognitive processes and contents	Included
10	noun.communication	nouns denoting communicative processes and contents	Included
11	noun.event	nouns denoting natural events	Excluded
12	noun.feeling	nouns denoting feelings and emotions	Included
13	noun.food	nouns denoting foods and drinks	Excluded
14	noun.group	nouns denoting groupings of people or objects	Excluded
15	noun.location	nouns denoting spatial position	Excluded
16	noun.motive	nouns denoting goals	Included
17	noun.object	nouns denoting natural objects (not man-made)	Excluded
18	noun.person	nouns denoting people	Excluded
19	noun.phenomenon	nouns denoting natural phenomena	Excluded
20	noun.plant	nouns denoting plants	Excluded
21	noun.possession	nouns denoting possession and transfer of possession	Included
22	noun.process	nouns denoting natural processes	Included
23	noun.quantity	nouns denoting quantities and units of measure	Included
24	noun.relation	nouns denoting relations between people or things or ideas	Included
25	noun.shape	nouns denoting two and three dimensional shapes	Excluded
26	noun.state	nouns denoting stable states of affairs	Included
27	noun.substance	nouns denoting substances	Excluded
28	noun.time	nouns denoting time and temporal relations	Included
29	verb.body	verbs of grooming, dressing and bodily care	Excluded
30	verb.change	verbs of size, temperature change, intensifying, etc.	Included
31	verb.cognition	verbs of thinking, judging, analyzing, doubting	Included
32	verb.communication	verbs of telling, asking, ordering, singing	Included
33	verb.competition	verbs of fighting, athletic activities	Included
34	verb.consumption	verbs of eating and drinking	Excluded
35	verb.contact	verbs of touching, hitting, tying, digging	Excluded
36	verb.creation	verbs of sewing, baking, painting, performing	Excluded
37	verb.emotion	verbs of feeling	Included
38	verb.motion	verbs of walking, flying, swimming	Excluded
39	verb.perception	verbs of seeing, hearing, feeling	Included
40	verb.possession	verbs of buying, selling, owning	Excluded
41	verb.social	verbs of political and social activities and events	Included
42	verb.stative	verbs of being, having, spatial relations	Included
43	verb.weather	verbs of raining, snowing, thawing, thundering	Excluded
44	adj.ppl	participial adjectives	Included

Table 4: WordNet sense description and eligibility for graph generation.

C Appendix C: WordNet Graphs

We present below the ACE-MK and GENIA-MK graphs that are described in Section 4.2.2 of the main part of the article. Different colors signify different communities as identified by community detection based on the modularity index of nodes. We visualise only the ten largest (in terms of the participating nodes) communities). We also visualise the top scoring words (regarded as representatives of each community) for the combination of Closeness, Betweenness and Eccentricity metrics.

In Figure 7 we observe the graph for the ACE-MK corpus while in Figure 8 the one for GENIA-MK.

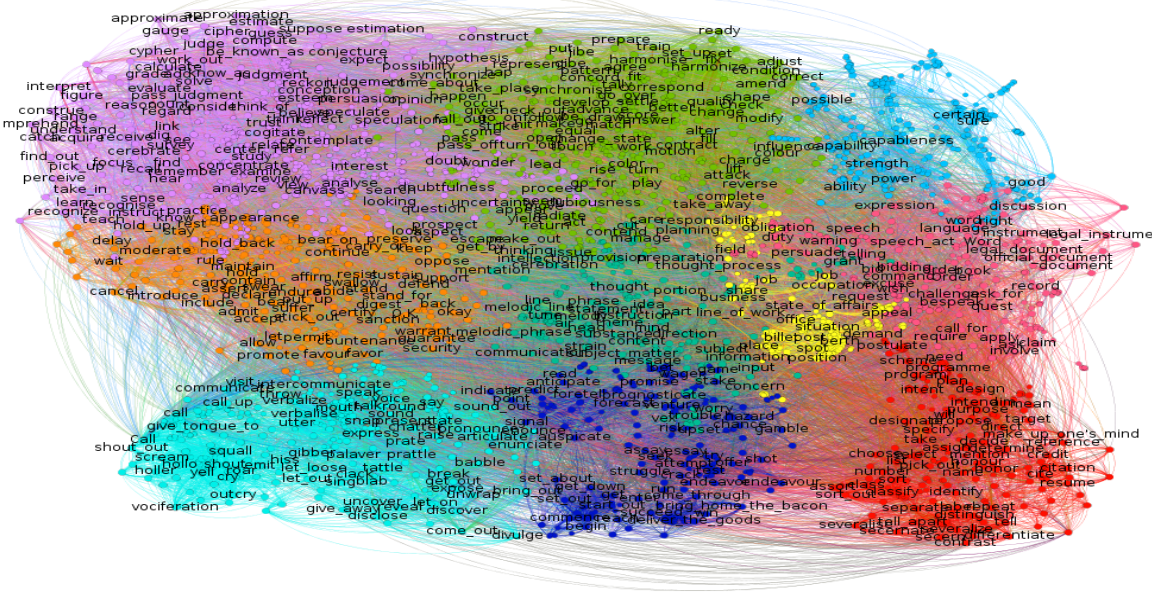

Figure 7: Generated word graph based on WordNet relations for ACE-MK cues.

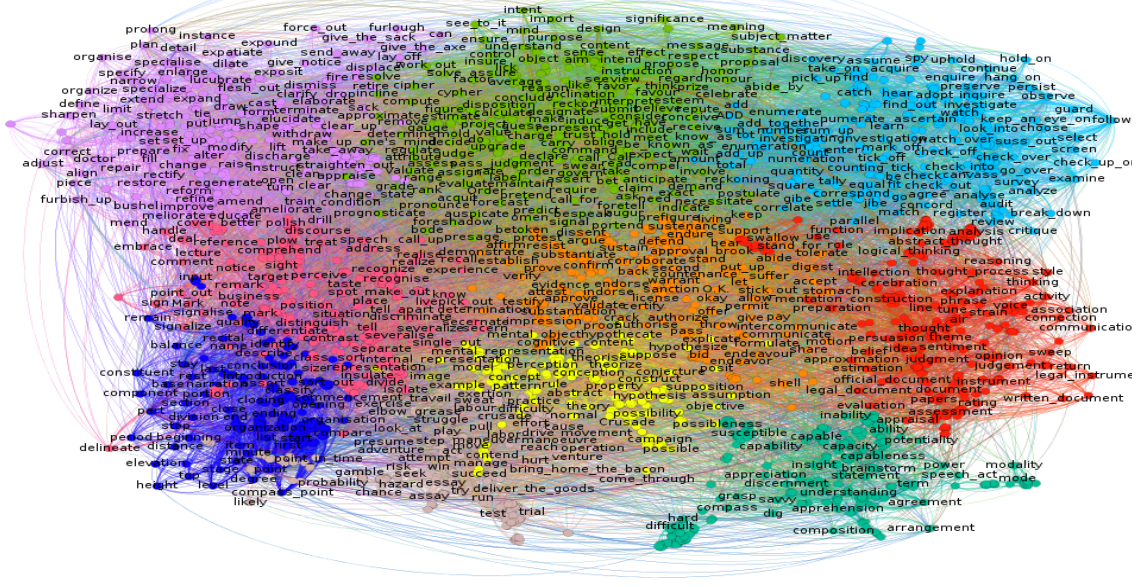

Figure 8: Generated word graph based on WordNet relations for GENIA-MK cues.