Weasels, Hedges and Peacocks: Discourse-level Uncertainty in Wikipedia Articles

Veronika Vincze[1,2]
[1]Hungarian Academy of Sciences, Research Group on Artificial Intelligence
[2]Department of Informatics, University of Szeged
vinczev@inf.u-szeged.hu

Abstract

Uncertainty is an important linguistic phenomenon that is relevant in many areas of language processing. While earlier research mostly concentrated on the semantic aspects of uncertainty, here we focus on discourse- and pragmatics-related aspects of uncertainty. We present a classification of such linguistic phenomena and introduce a corpus of Wikipedia articles in which the presented types of discourse-level uncertainty – weasel, hedge and peacock – have been manually annotated. We also discuss some experimental results on discourse-level uncertainty detection.

1 Introduction

In many areas of natural language processing, it is essential to distinguish between factual and non-factual information. Thus, depending on the precise task, negated or uncertain propositions should be treated separately by e.g. information extraction systems or they should be neglected. For instance, in medicine, if it is uncertain whether the patient suffers from an illness, the doctor should undertake further examinations to determine the final diagnosis. In another case, only those pieces of news are relevant in news media that are true and come from a reliable source. Uncertain information or unreliable sources should not be part of the news. In order to be able to find uncertain propositions in a huge amount of texts, a reliable uncertainty detector is needed, which can be only developed if annotated resources are at hand.

Previous studies on uncertainty detection concentrated mostly on the semantic dimensions. Indeed, in many cases it is the lexical content (meaning) of the uncertainty marker (cue) that is responsible for uncertainty, i.e. it can be identified in texts with the help of semantic tools. However, there are other types of uncertainty which cannot be described by just concentrating on semantics. For instance, *many* may denote quite different approximations: in the sentence *Many of the students did not read the book*, *many* may signal about 60-70% of the students (or at least more than 50%), while in *This airline loses many suitcases*, *many* may be only 20% but this number is still high enough for passengers to call it *many*. Here, the context and world knowledge determine how the quantifier *many* should be interpreted.

Here, we will focus on pragmatics- and discourse-related aspects of uncertainty. We will examine the concepts of source, fuzziness and subjectivity and their connection with uncertainty. As a first contribution, we will present a language-independent classification of such linguistic phenomena. As another contribution, we will introduce a corpus of Wikipedia articles in which linguistic cues of the presented types of discourse-level uncertainty have been manually annotated, hence empirical data on the frequency of such phenomena can also be provided. We will report the results of our experiments and we will also compare them with those of previous studies.

2 Discourse-level Uncertainty

Different concepts and terms that are related to uncertainty phenomena are employed. Modality is usually associated with uncertainty (Palmer, 1986), but the terms factuality (Saurí and Pustejovsky, 2012), veridicality (de Marneffe et al., 2012), evidentiality (Aikhenvald, 2004) and commitment (Diab et al., 2009) are also used. They all represent related but slightly different linguistic phenomena, which lie mostly in the category of semantic uncertainty. Propositions can be uncertain at the semantic level, that is, their truth value cannot be determined just given the speaker's mental state. Szarvas et al. (2012) offer a classi-

International Joint Conference on Natural Language Processing, pages 383–391,
Nagoya, Japan, 14-18 October 2013.

fication of semantic uncertainty phenomena.

Here, we use the term uncertainty similar to Szarvas et al. (2012), who aimed at giving a unified framework for the above-mentioned phenomena: "uncertain propositions are those [...] whose truth value or reliability cannot be determined due to lack of information". They contrast semantic uncertainty with discourse-level uncertainty: if the scheme "*cue* ***x*** but it is certain that not ***x***" is invalid (where ***x*** denotes a proposition, and *cue* denotes an uncertainty cue), that is, an uncertain proposition and its negated version cannot be coordinated, it is an instance of semantic uncertainty (e.g. *##It may be raining in New York but it is certain that it is not raining in New York*).

Besides semantic uncertainty, uncertainty can be found at the level of discourse as well. Here, the missing or intentionally omitted information is not related to the propositional content of the utterance but to other factors. In contrast to semantic uncertainty (Szarvas et al., 2012), the truth value of such propositions can be determined, but uncertainty arises if the proposition is analyzed in detail. For instance, the sentence *Some people are running* evokes questions like *Who exactly are those people that are running?* Here, the answer usually depends on the context, the speaker and the discourse and it cannot be determined out of context, thus henceforth such phenomena will be labeled discourse-level uncertainty.

We will carefully analyze discourse-level uncertainty phenomena below which are named after their most typical linguistic markers, i.e. cues. Although for the sake of simplicity we only provide English examples here, our categorization is based on pragmatic and cognitive considerations, and we will implicitly assume that our categories are language-independent. We will focus on Wikipedia articles, which – as indicated by previous studies (Ganter and Strube, 2009; Farkas et al., 2010) – seem to contain a certain amount of uncertainty phenomena like this. We will concentrate on three key aspects of discourse-level uncertainty, namely, sources, fuzziness and subjectivity.

2.1 Weasels

The notion of source is important for deciding the reliability of information conveyed (Saurí and Pustejovsky, 2012; Wiebe et al., 2005; Nawaz et al., 2010). It is not a matter of indifference to whom the information / opinion belongs to, especially in news media: people are more likely to believe a statement if it is communicated by a reliable source as opposed to a piece of sourceless information. In the public mind, experts, scientists, ministers, etc. are viewed as credible sources (cf. Bell (1991)) while unnamed or unidentifiable sources are considered less reliable. If some pieces of information are backed by a credible source, they are more likely to be treated as trustworthy, however, sourceless information is given less credence.

Events with no obvious sources are called *weasels* in Wikipedia[1] (Ganter and Strube, 2009): their source is missing or is specified only vaguely or too generally, hence, it cannot be exactly determined who the holder of the opinion is (undetermined source) as it is either not expressed or expressed by an indefinite noun phrase. Weasel sentences usually invoke questions like *Who said that?* and *Who thinks that?* The following sentence illustrates this:

> **Some** have claimed that Bush would have actually increased his lead if state wide recounts had taken place.

The ultimate source of the proposition expressed in the embedded sentence is not known since it is denoted by the pronoun **some**. Thus, it is not known who provided the opinion and therefore it is uncertain whether this is an important (reliable) piece of information (e.g. the opinion of experts) or whether it should be ignored.

Passive constructions which do not express the agent comprise a special type of weasels:

> It has been suggested **[by whom?]** that he should have involved Clinton much more heavily in his campaign.

The sentence does not reveal who has suggested the involvement of Clinton in the campaign. Hence, the source of the information is unclear and the source is missing from the sentence.

The basic idea behind weasel phenomena is the lack of a reference: it is not known who the source of the opinion is. This view is supported by the fact that a weasel candidate ceases to be uncertain if it is enhanced by citations:

> Most authors now prefer to place it within the genus *Pezoporus*, e.g. Leeton et al. (1998).

[1] `http://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style/Words_to_watch`

The phrase *most authors* would indicate a weasel (it is not clear whose opinion is this) but the citation at the end of the sentence clearly identifies the source.

In this paper, we extend the original notion of *weasel* and we argue that propositions that have an underspecified argument that would be relevant or is not common knowledge in the situation can be also viewed as weasels. Thus, a proposition is considered to be an instance of weasel if any of its relevant arguments is underspecified, i.e. it evokes questions like *Who/what exactly? Which?* Here, we give an example:

> While the Skyraider is not as iconic as **some other** aircraft, it has been featured in some Vietnam-era films such as *The Green Berets* (1968) and *Flight of the Intruder* (1991).

The sentence does not determine what kind of aircraft is considered iconic, so it is a vague or underspecified statement: we only know that there are "iconic aircraft", but no more details are specified. Again, the weasel type of uncertainty is expressed here by the adjectives *some* and *other*. Note that there is another occurrence of the word *some* in the sentence, but it does not denote any uncertainty in this case since the relevant Vietnam-era films are then listed.

2.2 Hedges

Another type of discourse-level uncertainty that will be discussed later on is called a hedge. Although a lot of studies used the term *hedge*, it may denote different linguistic phenomena for different authors. For instance, *hedge* means mostly *speculation* in the biomedical domain (see e.g. Medlock and Briscoe (2007), Vincze et al. (2008), and Farkas et al. (2010)). When contrasting epistemic modality and hedging, Rizomilioti (2006) categorizes approximators, passive voice and attribution to unnamed sources, among others, as instances of hedging and Hyland (1996) also cites them among common hedging devices.

Here, we understand *hedge* in the sense introduced by Lakoff (1973). For him, hedges are "words whose job is to make things fuzzier or less fuzzy", that is, the exact meaning of some qualities or quantities is blurred by them. Intensifiers (*very*, *much*), deintensifiers (*a bit*, *less*) and circumscribers (*approximately*) also belong to this group. Their effect is to add uncertainty to some elements in the proposition: they shift the value of some quality / quantity and the truth value of the proposition can only be decided if it is known what the reference point in the discourse is as the following example shows:

> Specialized services will **very often** provide a **much** more reliable service based on trusted publications.

In this sentence, there are several hedge cues. First, there is *often*, which informs us that it is not always the case that specialized services provide much more reliable service. It is modified by the intensifier *very*, which indicates that it is almost always the case (but still not always). Next, their service is *much* more reliable than any other service (at least those relevant in the context), that is, it is very reliable.

However, it should be noted that there is no absolute way to determine the truth value of this proposition without agreeing on what is meant by e.g. *often*: for now, let us say that *often* means at least seven out of ten times (but not ten times out of ten) and then *very often* may denote eight or nine times out of ten. It depends on the context, the speakers and the event described in the sentence to determine the reference point according to which the quantity or quality of events or entities can be evaluated. In the above example, the reference point may be 70%, and intensifiers denote that the quality or frequency of the event / entity is above the reference point, in this case, above 70%. Deintensifiers, however, assert that the quality or frequency is below the reference point.

Circumscribers – as their name states – circumscribe the exact amount or quality of the event or entity, which can be above or below the reference point. To represent this visually, they denote a set around the reference point in which the exact amount or quality is situated (see Figure 1 below). Here are some linguistic examples:

> This may explain why it has a lower than average estimated albedo of **˜0.03**.
>
> The duration of attacks averages **3-7** days.

It is interesting to note that in such cases not only cue words but also cue characters are responsible for uncertainty: the tilde and hyphen in these specific cases. Moreover, there are cue words that

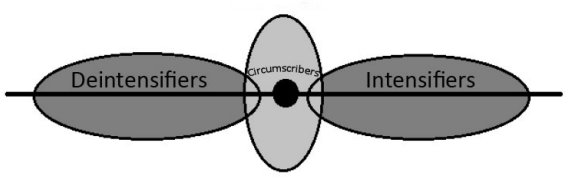

Figure 1: Types of hedges.

function as circumscribers as well like *approximately* and another use of *some*:

> Amsterdam Zuidoost has **approximately** 86,000 inhabitants and consists of **some** 38,000 houses.

Figure 1 shows the relationship of hedge types relative to the reference point. Thus, each type of hedge denotes a set in which the exact amount, quality or frequency of the relevant event or entity is situated but its exact place remains unclear.

Hedging is also one of the politeness strategies mentioned by Brown and Levinson (1987): they may function as mitigators in order to minimize disagreement, and to acknowledge that the speaker is imposing a task on the hearer. In the request *Could you please sort of correct this very short text for me?* the phrase *sort of* is a hedge, and the "very short" text may in fact be rather long. Here, hedges have pragmatic functions and they do not refer to uncertainty.

2.3 Peacocks

Subjectivity by its very nature contains aspects of uncertainty. People's opinions may differ from each other concerning specific things or events: they do not necessarily agree on what is good, neutral or bad. Thus, we cannot unequivocally determine what is good or what is bad.

Words that express unprovable qualifications or exaggerations are called *peacock* by Wikipedia editors.[2] Their meaning often inherently contain positive or negative subjective judgments, that is, they are polar expressions. Peacock terms include *brilliant*, *excellent* and *best-known*. Although their usage may be acceptable in other contexts, the objective style of Wikipedia editing requires that peacocks should be avoided.

[2] http://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style/Words_to_watch

Although they are not called peacocks by Wikipedia editors, we classify other subjective elements as peacocks as well. For instance, editorial remarks that refer to the subjective opinion of the author of the article (like *ironically* and *unfortunately*) or contentious labels (*controversial* and *legendary*) may all express subjectivity in certain contexts, hence we treat them here as peacock terms. The uncertainty in their meaning again lies in the fact that it cannot be objectively judged what can be called *excellent* for instance – it can be only deduced from discourse or contextual information and it may differ from speaker to speaker.

Here is a sentence with some peacock terms:

> Through the **ardent** efforts of Rozsnyai, the Philharmonia Hungarica quickly matured into one of Europe's **most distinguished** orchestras.

The words *ardent* and *most distinguished* are clearly positive in polarity, and again it cannot be objectively decided what level of enthusiasm is called ardent or which orchestras belong to the most distinguished ones.

All peacock terms are similar to hedges to some extent. They can be called scalar uncertainties since in both cases, a scale is involved in the interpretation of the uncertain term. In the case of peacock, there is a scale of polarity on which phrases can be judged as positive or negative whereas in the case of hedges, there is a scale on which there is a reference point, on the basis of which the uncertain part of the utterance is placed. Although they are similar, we suggest that peacocks and hedges be differentiated in our classification because peacocks are related to subjectivity while hedges are more neutral, hence they can be relevant for different NLP applications (e.g. in opinion mining, which seeks to collect subjective opinions on different topics, peacocks may prove more useful than hedges). Still, hedges shift the value of the quantity / quality mentioned in the text while peacocks denote a specific point on the scale, without modifying it, which again suggests that they should not be lumped in the same class.

3 Related Work

These days, uncertainty and modality detection is a widely studied area in natural language processing, which manifests itself in a number of corpora annotated for uncertainty in domains like biology

(Medlock and Briscoe, 2007; Kim et al., 2008; Vincze et al., 2008; Nawaz et al., 2010), medicine (Uzuner et al., 2009), news media (Wilson, 2008; Saurí and Pustejovsky, 2009; Rubin, 2010), and encyclopedia texts (Farkas et al., 2010). Although some authors have called attention to the fact that the progressive nature of discourse and dimensions of time should be also taken into account (de Marneffe et al., 2012; Saurí and Pustejovsky, 2012), as can be judged on the basis of available guidelines, most of these corpora make use of semantic uncertainty, with some exceptions that take into account pragmatic or discourse-level information as well (see below).

The concept of source has played a significant role in the literature. FactBank (Saurí and Pustejovsky, 2009) explicitly annotates the factuality of events according to their sources' perspective and Wiebe et al. (2005) also emphasize the role of sources annotated in the MPQA corpus for opinion mining. The notion of perspective – both in Nawaz et al. (2010) and in Morante and Daelemans (2011) – is similar to the one of sources applied in FactBank and MPQA. In Wikipedia, the lack of identifiable sources is explicitly discouraged by editors. They call such phenomena *weasels* (see also Ganter and Strube (2009)) and weasel detection was one of the subtasks of the CoNLL-2010 shared task (Farkas et al., 2010).

The lack of source characteristics to weasels can be paired with a certain strategy that Hyland (1996) calls impersonal constructions. It is a type of writer-oriented hedges[3] in his system. It is interesting to note that in his system, the opposite of this strategy can also be found, which could be called *anti-weasel*: the writer emphasizes his responsibility by using first person pronouns. However, this latter strategy does not represent any form of uncertainty in our view.

Fuzziness is another dimension of uncertainty. Lakoff (1973) gave an account of some lexical items – which he calls hedges – that "make things fuzzier", that is, words such as *approximately*, *kind of*, *at least* etc. Due to the presence of such words, the quality or quantity under investigation is shifted on a scale. If modified by the adverb *very* for instance, it moves towards one end of the scale on which this quality/quantity is determined. The phenomenon of hedging in scientific articles is analyzed and categorized according to the functions it can fulfill in Hyland (1996).

Subjectivity is also related to uncertainty. There is a great diversity among individual views and opinions: a feature of a product may be appreciated by some customers but it might be considered intolerable for others. Thus, what should be considered positive or negative seems subjective. Many approaches to subjectivity or sentiment analysis rely on lexicons and databases of subjective terms. For instance, the database SentiWordNet (Baccianella et al., 2010) contains a subset of the synsets of the Princeton Wordnet with positivity, negativity and neutrality scores assigned to each concept, depending on the use of its sentiment orientation, thus it is a lexicon where subjective terms are listed and ranked. Wilson (2008) defines subjectivity clues as words and phrases that express private states, that is, individual opinions. She distinguishes lexical cues and syntactic cues that are responsible for subjectivity. She lists several modifiers among her syntactic clues of subjectivity like *quite* and *really*. However, in contrast with other subjective elements, we do not regard them as peacock cues since – as Wilson (2008) herself states – they "work to intensify", so in our system they are classified as hedge cues. On the other hand, some instances of biased language can also be classified as peacocks in our system (Recasens et al., 2013).

Human communication and discourse is incremental in nature (Cristea and Webber, 1997). Information may be added at a later point of the discourse that clarifies a previously missing piece of information. Applying this to discourse-level uncertainty, it may be the case that an apparent weasel phrase is elaborated on later in the discourse, or the exact value of an apparent hedge expression is later provided. In such cases, the phrases should not be marked as uncertain, which indicates the essential role of co-text – i.e. surrounding words in the text (Brown and Yule, 1983) – in detecting discourse-level uncertainty.

4 The Annotated Corpus

In order to test the practical applicability of the new classification of discourse-level uncertainty phenomena, and to investigate the frequency of each uncertainty type, we also created an annotated corpus. We selected WikiWeasel, the Wikipedia subset of the CoNLL-2010 Shared

[3]However, in our classification, it should be called a weasel.

Task corpora (Farkas et al., 2010) for annotation. By doing this, our results could be contrasted with those of the original annotation carried out specifically for the shared task. Moreover, as the corpus has recently been annotated for semantic uncertainty (Szarvas et al., 2012), interesting comparisons can also be made between semantic and discourse-level uncertainty. The annotated corpus is available free of charge for research purposes at `www.inf.u-szeged.hu/rgai/uncertainty`.

4.1 Statistical Data on the Corpus

The dataset consists of 4,530 Wikipedia articles and 20,756 sentences. Texts were manually annotated by two linguists for linguistic cues denoting all types of discourse-level uncertainty, i.e. weasel, peacock and hedge. 200 articles were annotated by both linguists and the inter-annotator agreement rate for the categories weasel, peacock and hedge were 0.4837, 0.4512 and 0.4606, respectively (in terms of κ-measure), which reflects that identifying discourse-level phenomena is not straightforward, however, it can be reasonably well solved considering the subjective nature of the task. During the annotation, special emphasis was laid on the discourse structure of the text. For instance, weasel cue candidates do not denote uncertainty when the sentence is enhanced with citations. Also, a weasel-like element may be elaborated on in the next sentence, thus it is not to be marked as weasel as in:

> Some ship names are references to other games created by Jordan Weisman. The "Black Swan" is a reference to a character from Crimson Skies, and also possibly to the ship Black Pearl from Pirates of the Caribbean.

In order to attain the gold standard for the commonly annotated parts, the two annotators discussed problematic cases and reached a consensus for each case. The final version of the corpus contains these disambiguated cases.

The dataset contains 10,794 discourse-level uncertainty cues[4], which occur in 7,336 uncertain sentences. A sentence was considered to be uncertain if it contained at least one uncertainty cue. But, as the results show, many sentences include more than one uncertainty cue. Statistical data on the uncertainty cues found in the WikiWeasel corpus are listed in Table 1, together with available data on semantic uncertainty types, taken from Szarvas et al. (2012).

Uncertainty cue	#	%	Diff. cues
Hedge	4,743	35.24	260
Weasel	4,138	30.75	99
Peacock	1,913	14.21	540
Discourse-level total	10,794	80.2	899
Epistemic	1,171	8.7	114
Doxastic	909	6.75	36
Conditional	491	3.65	15
Investigation	94	0.7	12
Semantic level total	2,665	19.8	166
Total	13,459	100	1065

Table 1: Uncertainty cues in WikiWeasel.

As can be seen, most of the uncertainty cues found in the corpus belong to the discourse-level uncertainty class, the ratio of semantic to discourse-level uncertainty cues being 1:4. Among the types of discourse-level uncertainty, hedges are the most frequent, followed by weasels and peacocks. All this suggests that discourse-level uncertainty is very typical of Wikipedia articles, about 35% of the sentences being uncertain at the discourse level. As regards the specific classes, 3,807 (18.3%), 3,497 (16.8%) and 1,359 (6.5%) sentences contain at least one hedge, weasel or peacock cue, respectively.

4.2 Cue Distribution in the Corpus

On the number of different cues, Table 1 tells us that the set of linguistic cues expressing weasels are the most limited, with almost 100 cues. In contrast, peacock cues vary the most with 540 cues. This suggests that weasels have the most restricted vocabulary in contrast to peacocks, and hedges being in the middle. This also means that the average frequency of a weasel cue is much higher than that of a peacock cue: the average frequency of occurrence of weasel, hedge and peacock cues is 41.8, 18.24 and 3.54, respectively.

We did a more detailed analysis on the lexical distribution of the cues as well. The ten most frequent cues for each type are listed in Table 2. These are responsible for about 86%, 45% and 42% of the occurrences of weasel, hedge and peacock cues, respectively. Thus, a limited vocabu-

[4]We should mention that our corpus contained 680 passive constructions, which were annotated as weasels. As we focus now on lexical cues of discourse-level uncertainty, and they belong to syntactic cues, the investigation of such cases will be subject to further studies.

Weasel	#	%	**Hedge**	#	%	**Peacock**	#	%
some	887	25.64	often	539	11.36	most	318	16.62
many	631	18.24	usually	263	5.55	popular	112	5.85
other	539	15.58	many	217	4.58	famous	81	4.23
several	204	5.90	generally	210	4.43	well-known	50	2.61
most	202	5.84	very	206	4.34	notable	50	2.61
various	177	5.12	most	179	3.77	notably	45	2.35
others	175	5.06	almost	152	3.20	important	40	2.09
certain	82	2.37	several	140	2.95	best	38	1.99
number	43	1.24	common	127	2.68	traditionally	38	1.99
critics	37	1.07	much	119	2.51	controversial	37	1.93

Table 2: The most frequent discourse-level uncertainty cues in the WikiWeasel corpus.

lary can account for over 85% of weasels.

However, some terms can belong to more than one uncertainty type. For example, *most* occurs in all the three types (weasel: ***Most*** *agree that this puts her at about 12 years of age*, hedge: *He spent* ***most*** *of his time working on questions of theology* and peacock: *Kathu is the district which covers the* ***most*** *touristical beach of Phuket*), but *some*, *many* and *several* can all be instances of weasels and hedges. This is due to the linguistic variability of these items: e.g. *some* may refer to "an indefinite quantity" or "something unspecified".

As can be seen, there are some overlapping cues among the types. This is especially so in the case of hedges and weasels: 25 cues can denote hedges or weasels as well, thus 25% of the weasel cues are ambiguous. These cues were also responsible for most of the differences between the two annotations, which indicates that their identification requires special attention both for human annotators and NLP tools: it is mostly the neighbouring words that can determine whether it is a weasel or hedge. For instance, if *some* occurs before a verb and constitutes a noun phrase on its own, then it is almost certainly a weasel cue (*Some think that*...) but if it occurs before a noun denoting time, it is probably a hedge (*some minutes ago*).

5 Experiments

We carried out some baseline experiments on the corpus. We divided the corpus into training (80%) and test (20%) sets and applied a simple dictionary-based approach which classified each cue candidate as uncertain if it was tagged as uncertain in at least 50% of its occurrences in the training dataset. For ambiguous cues, the most frequent label was chosen (e.g. *most* was used as a peacock cue). Similar to the CoNLL-2010 shared task, we evaluated our results at the cue level as well as at the sentence level.

	Cue level			**Sentence level**		
	P	R	F	P	R	F
Weasel	0.7088	0.6724	0.6901	0.7443	0.7183	0.7311
Hedge	0.8780	0.6616	0.7546	0.9185	0.7193	0.8068
Peacock	0.4222	0.4730	0.4462	0.4034	0.5341	0.4597
Micro F	0.7196	0.6348	0.6745	0.7458	0.6924	0.7181

Table 3: Baseline results in terms of precision / recall / F-score.

Table 3 shows that the peacock class is the most difficult to detect, which may be due to the fact that this class has the most diverse cues and thus applying a dictionary-based method leads to a lower recall. Still, the lower precision was due to the higher level of ambiguity concerning the most typical peacock cues (like *most*). As for hedges, a simple lexical approach can result in a good precision score, which suggests that hedge cues are less ambiguous than weasel or peacock cues. It is also seen that sentence-level results are significantly higher than cue-level results (ANOVA, $p = 0.0026$). Uncertain sentences typically contain more than one cue and in the former scenario, it is sufficient to recognize only one cue in the sentence to regard the sentence as uncertain and false negatives do not affect the performance significantly.

If we compare the data with the CoNLL-2010 version of the corpus, it is seen that the new annotation scheme leads to many more cues (6,725 cue phrases in 4,718 uncertain sentences in the original version vs. 10,794 cues in 7,336 sentences in the version described here) and – although the datasets are not directly comparable – it gives a much better performance: the best system achieved an F-score of 60.2 on weasel detection at the sentence level and 36.5 at the cue level and no classes of cues were distinguished there (Farkas et al., 2010). This difference may be attributed to several factors. First, not all hedge phenomena (used in the sense introduced here) were systematically annotated in the CoNLL-2010 corpus.

Second, complex syntactic structures that contained several types of uncertainty were annotated as one complex cue (e.g. the phrase *it has been widely suggested*, which contains epistemic uncertainty (*suggested*), weasel (passive sentence with no agent) and hedge (*widely*) as well). Third, the CoNLL-2010 version did not distinguish subtypes of cues, i.e. semantic uncertainty and weasels were annotated in the same way. It was probably because of this lack of distinction that participants of the shared task got considerably lower results for Wikipedia articles than for biological papers, which contained fewer weasel cues (Farkas et al., 2010). However, the new annotation makes it possible to select those types of uncertainty that are relevant for a given application, see Section 6.

6 Discourse-level Uncertainty and NLP

Detecting weasels is of utmost importance in every information extraction application where it should be known who the author/source is. Thus, information extraction applied for the news media may certainly profit from finding weasels, i.e. missing or undeterminable sources. Pieces of information without an identifiable (and reliable) source require special treatment: they will be excluded from the news or they will be communicated to the public in a special form, using phrases such as *according to unnamed sources* etc.

In sentiment analysis and opinion mining, the identification of subjective terms is essential. These terms are often ambiguous hence a subjectivity word sense disambiguation is needed (Wiebe, 2012). In our corpus, peacock terms and intensifiers – a subtype of hedges – are manually annotated, thus it can be used in the development and evaluation of tools that seek to disambiguate elements of a subjectivity lexicon in running texts.

Information retrieval may also be enhanced by detecting discourse-level uncertainty. In order to find relevant documents for queries that contain numbers, more specifically, to improve recall in such cases, it is important to handle numeric hedges. For instance, if someone looks for websites describing games appropriate for ten year old children, he also may be interested in games that are for children over eight. Thus, the search engine should be prepared for recognizing that the number specified in the query ("ten") is part of other numeric sets (e.g. "over eight") and in this way, more relevant hits can be retrieved.

The linguistic processing of patents especially requires that hedges should be recognized. There is a tendency to generalize over the scope of the patent (i.e. hedges are used) in order to prevent further abuse (Osenga, 2006). Thus, the scope of the patents can be expanded or other use cases can later be included in the patent. Hence, any NLP system that aims at patent processing must target hedge detection as well.

Document classification may also profit from detecting discourse-level uncertainty since different genres of texts involve different types of uncertainty.For instance, papers in the humanities contain significantly more hedges than papers in sciences (Rizomilioti, 2006). Thus, the frequency of hedges may be indicative of the domain of the text as well, which again may be exploited in document classification.

7 Conclusions

In this paper, we presented a classification of discourse-level uncertainty phenomena, and focused on the concepts of source, fuzziness and subjectivity. We also introduced a corpus of Wikipedia articles in which linguistic cues for each type of discourse-level uncertainty – weasel, peacock and hedge – were manually annotated. We carried out some baseline experiments on discourse-level uncertainty detection, which may prove useful in information extraction and retrieval, sentiment analysis and opinion mining.

In the future, we intend to develop a machine-learning based uncertainty detector. We would also like to investigate the distribution of weasels, hedges and peacocks in other types of texts (e.g. news media or scientific papers) and in other languages as our three categories are language-independent. Moreover, to learn how domain-dependent the model is, we plan to do some domain adaptation experiments as well.

Acknowledgments

This work was supported in part by the European Union and the European Social Fund through the project FuturICT.hu (grant no.: TÁMOP-4.2.2.C-11/1/KONV-2012-0013).

References

Alexandra Y. Aikhenvald. 2004. *Evidentiality*. Oxford University Press, Oxford.

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. In *Proceedings of LREC'10*, Valletta, Malta, May. ELRA.

Allan Bell. 1991. *The language of the News Media*. Blackwell, Oxford.

Penelope Brown and Stephen C. Levinson. 1987. *Politeness: Some universals in language usage*. CUP, Cambridge, UK.

Gillian Brown and George Yule. 1983. *Discourse Analysis*. CUP, Cambridge, UK.

Dan Cristea and Bonnie Webber. 1997. Expectations in incremental discourse processing. In *Proceedings of ACL97/EACL97*, pages 88–95. Morgan Kaufmann.

Marie-Catherine de Marneffe, Christopher D. Manning, and Christopher Potts. 2012. Did it happen? The pragmatic complexity of veridicality assessment. *Computational Linguistics*, 38:301–333.

Mona T. Diab, Lori S. Levin, Teruko Mitamura, Owen Rambow, Vinodkumar Prabhakaran, and Weiwei Guo. 2009. Committed belief annotation and tagging. In *Proceedings of LAW 2009*, pages 68–73. The Association for Computer 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 CoNLL-2010 Shared Task*, pages 1–12, Uppsala, Sweden, July. ACL.

Viola Ganter and Michael Strube. 2009. Finding Hedges by Chasing Weasels: Hedge Detection Using Wikipedia Tags and Shallow Linguistic Features. In *Proceedings of ACL-IJCNLP 2009*, pages 173–176, Suntec, Singapore, August. ACL.

Ken Hyland. 1996. Writing without conviction? Hedging in scientific research articles. *Applied Linguistics*, 17(4):433–454.

Jin-Dong Kim, Tomoko Ohta, and Jun'ichi Tsujii. 2008. Corpus annotation for mining biomedical events from literature. *BMC Bioinformatics*, 9(Suppl 10).

George Lakoff. 1973. Hedges: A study in meaning criteria and the logic of fuzzy concepts. *Journal of Philosophical Logic*, 2(4):458–508.

Ben Medlock and Ted Briscoe. 2007. Weakly Supervised Learning for Hedge Classification in Scientific Literature. In *Proceedings of the ACL*, pages 992–999, Prague, Czech Republic, June.

Roser Morante and Walter Daelemans. 2011. Annotating Modality and Negation for a Machine Reading Evaluation. In *Proceedings of CLEF 2011*.

Raheel Nawaz, Paul Thompson, and Sophia Ananiadou. 2010. Evaluating a meta-knowledge annotation scheme for bio-events. In *Proceedings of NESP 2010*, pages 69–77, Uppsala, Sweden, July. University of Antwerp.

Kristen Osenga. 2006. Linguistics and Patent Claim Construction. *Rutgers Law Journal*, 38(61):61–108.

Frank Robert Palmer. 1986. *Mood and Modality*. CUP, Cambridge.

Marta Recasens, Cristian Danescu-Niculescu-Mizil, and Dan Jurafsky. 2013. Linguistic models for analyzing and detecting biased language. In *Proceedings of ACL-2013*, pages 1650–1659, Sofia, Bulgaria, August. ACL.

Vassiliki Rizomilioti. 2006. Exploring epistemic modality in academic discourse using corpora. In *Information Technology in Languages for Specific Purposes*, pages 53–71. Springer US.

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 and James Pustejovsky. 2009. FactBank: a corpus annotated with event factuality. *Language Resources and Evaluation*, 43:227–268.

Roser Saurí and James Pustejovsky. 2012. Are you sure that this happened? Assessing the factuality degree of events in text. *Computational Linguistics*, 38:261–299.

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

Özlem Uzuner, Xiaoran Zhang, and Tawanda Sibanda. 2009. Machine Learning and Rule-based Approaches to Assertion Classification. *Journal of the American Medical Informatics Association*, 16(1):109–115, January.

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(Suppl 11):S9.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39(2-3):164–210.

Janyce Wiebe. 2012. Subjectivity word sense disambiguation. In *Proceedings of WASSA 2012*, page 2, Jeju, Korea, July. ACL.

Theresa Ann Wilson. 2008. *Fine-grained Subjectivity and Sentiment Analysis: Recognizing the Intensity, Polarity, and Attitudes of Private States*. Ph.D. thesis, University of Pittsburgh, Pittsburgh.