A hedging annotation scheme focused on epistemic phrases for informal language

Liliana Mamani Sanchez
CNGL/Computational Linguistics Group
Center for Computing and Language Studies
School of Computer Science and Statistics
Trinity College Dublin,
The University of Dublin
Dublin 2, Ireland
mamanisl@tcd.ie

Carl Vogel
Computational Linguistics Group
Center for Computing and Language Studies
School of Computer Science and Statistics
Trinity College Dublin,
The University of Dublin
Dublin 2, Ireland
vogel@tcd.ie

Abstract

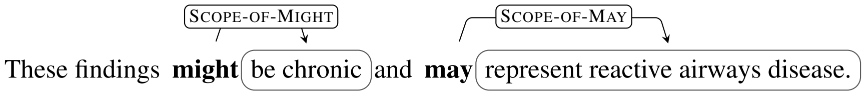
Most existing annotation schemes for hedging were created to aid in the automatic identification of hedges in formal language styles, such as used in scholarly prose. Language with informal tone, typical in much web content, poses a challenge and provides illuminating case studies for the analysis of the use of hedges. We have analysed conversations from a web forum and identified the manners individuals express hedging through expressions which differ slightly regarding to their lexical form from hedges used in formal writing. Based on these observations, we propose an annotation scheme composed of three main categories of hedges where the main class comprises first person epistemic expressions that explicitly note an individual's involvement in what they express. We provide here an overview of our insights obtained by annotating a dataset of web forum posts according to this scheme. These observations will be useful in the design of automatic methods for the detection of hedges in texts in informal language.

1 Introduction

This paper presents an annotation scheme for hedging focused on epistemic modality expressions in informal language. Hedges are used by a speaker to modulate the degree of commitment expressed by his or her statements. Hedging is deployed within various speaker states such as uncertainty, possibility, or politeness. A number of schemes have been proposed directly or indirectly to annotate hedges.

The elements involved in a hedging event occurring in a sentence are: the *hedging expression*, the *source* and *scope* of the hedge. The *source* refers to the entity experiencing or/and expressing the mental state represented by the hedge. The *scope* refers to the sentential constituents that are affected by a hedge expression. In (1), the hedges *might* and *may* are linked to their respective scopes. The source corresponding to both hedges, although only implicit in the sentence, is the sentence's author.

(1)



These findings **might** be chronic and **may** represent reactive airways disease.

Most existing annotation schemes label hedges as single units, while a minority have studied more complex classifications of hedges.

Bioscope is among the corpora which scheme is in the first group (Szarvas et al., 2008). Bioscope is a biomedical dataset of sentences from medical and biomedical articles tagged with speculation and negation information. Minimal units in a sentence are annotated as expressing hedging or negation, if such phenomena occur, and discuss cases where these lexical units are not actually used to imply speculation/negation. They noted that keywords differ in their propensity to be speculative depending on the domain. Ganter and Strube (2009) exploited the concept of "weasel word" in Wikipedia to semi-automatically build a dataset of hedges. Weasel words comprise expressions that Wikipedia editing policies discourage, such as *some people say*, *it is believed*, *many are of the opinion*, *most feel*, *research has shown*, etc. Vincze (2013) divided speculative cues in this Wikipedia dataset into three types: weasels, hedges and peacocks. Weasels signal uncertainty regarding an argument identity. For instance, the uncertainty in *some other* is caused by the lack of specification in 'While the Skyraider is not as iconic as **some other** aircraft…'. Hedges follow the regular conception limited to expressing uncertainty. Peacocks signal various sorts of subjective judgements such as *ardent* and *most distinguished.*

In the second group there are the works of Rubin et. al. (2010), Wiebe et al. (2005) and Hyland (1998), to mention few important ones. Rubin (2010) proposed a multi-dimensional certainty annotation model, where the main dimension qualifies expressions according to five categories that range from total uncertainty to total certainty. Wiebe et al. (2005); Wilson and Wiebe (2005) place speculation within a larger and more complex framework for annotation of opinions. This annotation was centred on the concept of Private States; i.e. they are not open to observation or verification, and they comprise opinions, beliefs, thoughts, feelings, emotions, goals, evaluations, and judgements and speculations. Hyland (1998) proposed a categorization taxonomy of hedges in scientific articles according to their function as: content-oriented, reader-oriented or writer-oriented.

We studied hedges in informal language in a dataset of web forum posts, but some issues emerged when attempting to annotate hedges according to conceptions in the aforementioned approaches. In Section 2, we provide an overall description of these issues. Our attempts to address these issues led us to propose a categorization scheme for hedges in language with informal tone which is described in Section 3. In Section 4, we present a summary of our findings from analysis and manual annotation in the dataset of forum posts, and in Section 5, we present our conclusions and views for future work.

2 Problem description

In this section we describe some issues related to the annotation of hedges in texts from an informal language type. The web forum from which posts were extracted belongs to a commercial software vendor; it is a forum in which users look to other users for advice in solving their software-related issues.

Most of annotation schemes mentioned in the previous section focus uniquely on the identification of a hedge and its scope in texts of formal tone and are only concerned with the interpretation of hedges as speculative signals. Automatic approaches for the identification of speculation (that implies automatic annotation of hedges) such as used by Light et al. (2004), Medlock and Briscoe (2007) and Farkas et al. (2010) follow the same line by targeting academic articles. Few approaches address annotation in informal register texts (Vincze et al. (2014) provide an exception). However, the informal texts of online web fora typically do not benefit from rounds of editing: the text is typically "noisy". Noisy text often contains ungrammatical language, misspellings, typos and non-linguistic strings (like emoticons), which limit the efficacy of natural language processing tools and methods (that perform moderately in well-formed text) when applied to texts with such informal language.

Despite acceptance that uncertainty expressed in a proposition is partly a product of reporting other points of view, as in (2), only few approaches such Rubin (2010), Wiebe et al. (2005) and Hendrickx et al. (2012) address the annotation of the hedging source, and none of the projects for automatic detection of hedges also address the detection of the hedging source. Disregard for the source follows from the fact that regardless of who the speculation experiencer is, a hedge is worth identifying since it points out to non-factual information.

(2) The existence of such an independent mechanism has also been **suggested** in mammals.

This approach can be thought of as 'content-centered'. Potentially, also identifying the experiencer of the hedging event could aid the building of the statements such as 'Individual A knows X and has certainty about it' or 'Individual B does not know whether X', and the like. This would be a 'user-centered' approach since explores the qualities and properties of a writer's utterances to find out whether a hedging expression reflects the writer's perspective or not.

In a domain of web forum posts generated by forum participants, the kind of expressions used to convey hedging are slightly different from hedges such as *suggest*, *potential*, *likely*, and *may* used frequently in more formal contexts. Informal expressions of hedging include phrasal expressions, acronyms and spoken-register transcriptions (e.g. *not sure*, *IMO* (*In My Opinion*), *AFAIK (As Far As I Know)*, *dunno*). When the hedging experiencer is explicit in these sentences, this fact is lexically realized with the use of first person expressions such as *I am not sure*, *My opinion*, *IMO* or *to me, it looks like* in sentences (3), (4), (5) and (6) respectively.[1] These sentences convey the forum post writer's direct involvement in the hedging phenomenon, which is revealed by the use of 'I' as subject, the first-person possessor in *IMO* for 'In My Opinion', or by the pronoun *me*. This sort of phrasal hedges can also appear in a discontinuous manner in a sentence such as *I think* in (7).

(3) **I am not sure** which SP is on here, or how to check. Post: 18706

(4) **I'd suggest** the following additional steps: Post: 3655

(5) **IMO** it is best to always leave tamper protection on to prevent threats · · · Post: 16687

(6) **To me, it looks like** the O/P wants to try out the 2011 beta for testing · · · Post: 15134

I THINK

(7) **I** don't know how it's in other countries, but **think** it 's almost the same Post: 35934

Our research seeks insights that contribute in both content and user-centered studies of hedging. Therefore, our development of an annotation scheme and subsequent analysis take into account both perspectives. To this end, features the annotation scheme should consider are: a) identification of the hedging source, b) identification of the scope, c) domain-generality of annotated expressions, d) inclusion of different interpretations of hedging, e) functionality with noisy text, and d) capacity to annotate non-contiguous hedging elements in a sentence.

3 A scheme for hedging in informal language

The annotation scheme was built around three elements: Entities, Relations and Attributes. Entities are used to represent: a) a hedge, b) its source c) its scope, d) non-hedge and e) other discourse markers. Relations are used to link the hedge entities with their source or scope. Attributes are additional information about hedge entities that can be filled in during the annotation process.

We address three main types of hedges: a) Single-hedges, b) Not-Claiming-Knowledge epistemic phrases and c) Syntactic hedges. These are described in next sections. An additional label, "Non-hedge", marks entities that were deemed as potential hedges but not actually used in any hedging sense.

3.1 Single hedges

This hedge category corresponds to the traditional conception of hedges as single words conveying uncertainty. They are usually modal and lexical verbs expressing epistemic modality such as *may*, *appear* and *suggest*. The initial lexicon of single hedge instances considered for manual annotation were extracted from Rubin's (2010) work. Some lexical items such as "*can not know*" and "*don't understand*"

[1] The number preceded by 'Post' in examples throughout this paper corresponds to the post identification number for the forum post in the data set from which the example was extracted.

could overlap with Not-Claiming-Knowledge (NCK) epistemic phrases, but they are only deemed as NCK expressions if they are associated to a first person pronoun in the sentence.

3.2 Source

We define two categories of Source for a hedging expression: a) **Inner Epistemic Source** and b) **Outer Epistemic Source**. The Outer Epistemic Source is always the post's writer, as the writer selects a statement's content. The Inner Epistemic Source corresponds to the entity whose hedged point of view is expressed in the sentence; thus, the Inner Epistemic Source can be the writer or not. The use of *suggest* and *suggested* in examples (8) and (9), respectively, illustrates these two categories.

(8) User1: [...] **I'd suggest** the following additional steps: ...

(9) User2: User1 **suggested** following some steps, and you should consider [...]

In (8), User1 is the one asserting his/her own hedged point of view. In this case, the Inner Epistemic Source is attributed to User1, while in (9) it is attributed to User1 despite the fact that User2 is the proposition's author. In cases where the writer express his/her own point of view in the hedging event, the Outer Epistemic Source coincides with the Inner Epistemic Source as in (8).

When the Source is not explicit in the sentence, it can occur either implicitly as in (10) or as subject ellipsis as in (11).[2] Particularly, web forum text is prone to subject ellipsis due to its informal style.

(10) [...] *some sort of malware* **might** be preventing you from seeing the stock quotes.

(11) and $\epsilon_{subject}$ **don't know** if this is the correct place to ask it so pls lemme know if i shud ask elsewhere on the forum [...]

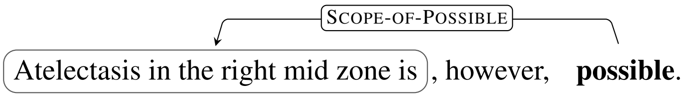
The implementation of the annotation scheme we describe here provides the means for marking the different cases when the Inner Epistemic Source is explicit or not.

3.3 Scope

In this research, scope annotation is mainly focused on the syntactic dependency head of the phrase affected by the hedge. This means that our approach to annotating the hedging scope is not as strict as it might be: at least the syntactic head of the scope has to be annotated and linked to the entity representing the corresponding hedging expression. This is mainly due to the inherent complexity of identifying a particular clause boundaries inside a sentence.

In our annotation scheme, the scope is separated from the hedging entity and linked to it by a Scope-of relationship as in example (12). This has the advantage of avoiding the marking of extra words that do not belong to the scope. For instance for (12), Szarvas et al. (2008) by including the hedge 'possible' within the scope boundaries, tag *however* as part of the scope.[3]

(12)



3.4 Epistemic phrases

This category includes first person epistemic phrases: they typically contain a first person subject, an epistemic verb such as such as *I think*, *I suppose* and *I wonder*, and a complement clause is embedded under the epistemic modality in this kind of phrase. In this section, we show how the concept of first person epistemic phrases moves away from the conceptualization of hedges coming from the epistemic modality tradition.

[2] Here, $\epsilon_{subject}$ is used to signal an unexpressed but interpreted subject.

[3] This is done to keep the hedge linked to its scope. This way of representing the relation was chosen because of limitations in the annotation tools they were using (Szarvas et al., 2008).

Early discussion about interpreting epistemic phrases as hedges originated in the analysis *I think*. Particularly, Thompson and Mulac (1991) consider this epistemic phrase has achieved a hedging state through a process of grammaticalization. Their view is that *I think* is roughly similar to *maybe* when used to express the degree of speaker commitment, thus comprising a grammatical sub-category of adverbs.

Scheibman (2001), Kärkkäinen (2010) and Wierzbicka (2006) showed that first person epistemic phrases used to express personal stance are highly frequent in various registers in contemporary English. They also describe particular properties of these phrases such as: a) representing the speaker's attitude with respect to the subsequent piece of discourse in contrast to when third person is used, there the piece of discourse is seen as a description (Scheibman, 2001), b) it comprises explicitly subjective claims in contrast to impersonal expressions where the hedging source is obscure (Hyland, 1998) , c) used to express knowledge states, as a boundary marker for turn-taking in conversation, a speaker's perspective marker, and as a way to align the speaker's with the listener's stance (Kärkkäinen, 2010). Besides, Wierzbicka (2006) suggests that this category of phrases merits recognition as a major grammatical and semantic class in modern English. She acknowledges a rigorous semantic and cultural contextual analysis of each type would be needed to provide an accurate interpretation, which she does in part.

A preliminary examination of hedge expression in our dataset showed that writers use first person singular epistemic phrases exhibiting characteristics mentioned above. Expressions of hedging such as *I'd suggest*, *I assumed*, *I am not sure* and *IMO* (*In My Opinion*) (see (14b), (15b), (3) and (16)) emphasize the writer's involvement in a proposition. In this aspect, they are different from epistemic phrases that have second and third person subjects as *They didn't know* in (13b). Annotating these expressions according to the traditional conceptualization of hedges would not capture the subjectivity expressed in epistemic phrases. This would result in annotating only the lexical verbs *suggest*, *assumed* in (14b) and (15b) in a similar manner as they would in (14a) and (15a); or in terms of epistemic phrases only *didn't know* would be annotated in (13a) which makes of this annotation equivalent to the one in (13b). The difference with epistemic modals is also relevant as they act as verbal modifiers, and even in first person subject propositions as in (18) the main constituent is the predicate they hedge. On the other hand, first person epistemic phrases are more subjective in the sense that the subject's involvement (revealed by the use of first person pronouns) is emphasized by the main constituent of these epistemic phrases (predicates related to mental states).[4]

(13) a. **I didn't know** what else to do and I still don't. Post: 288960

b. **They didn't know** the file and recommended to kill the process. Post: 4452

(14) a. . . . and some researchers **suggest** that fibromyalgia and CFS are related.

b. **I'd suggest** the following additional steps: Post: 3655

(15) a. and it is **assumed** that he learned to read and write at the local parish school.

b. **I assumed** I would be able to retrieve all my files, but so far - no such luck. Post: 7492

(16) This is a kind censorship, **IMO**. Post: 171336

(17) I for one **think** the best course of action to take when you believe you are . . . Post: 16687

(18) As I learn more about the problem I **may** ask for more info. Post: 25545

We believe first person epistemic phrases reach the same level of grammaticization as 'I think' showed by Thompson and Mulac (1991), so matching only the main lexical component expressing hedging would only partially represent the writer's commitment. The frequency and variety of epistemic phrases to be detailed in Section 4 support these intuitions. Furthermore, matching hedging expressions that resemble traditional hedges and that are not epistemic phrases is not always possible; that is the cases of *IMO* and *AFAIK*. Both acronyms reinforce our hypothesis that epistemic phrases are a distinctive grammaticalized category of hedges since they stand for the first-person epistemic phrases *In My Opinion* and *As Far As I Know* respectively, and even the case of *IDK* standing for *I don't know*.

[4] In Section 4 the main types of first person epistemic phrases are characterized according to their main constituent.

In some cases additional epistemic phrases modifiers modulate the intensity the subjective force, as in *I for one think* in (17).[5] However, we have not yet addressed degrees of certainty in hedging.

We reckon these expressions and the like constitute units of meaning with distinctive hedging qualities, and we aim to provide theoretical support to account for them as a newly grammaticized category of hedges. Ensuring that the writer is the one experiencing the mental state expressed by a hedge is highly relevant. One of the goals of this research is finding ways the epistemic source of a hedging event can be easily identified. We propose first person singular and plural epistemic phrases as hedge category when the subject of the epistemic experience is relevant to be identified. In Section 4 we provide overall description of our findings around these phrases that we call of Non-Claiming-Knowledge (NCK) epistemic phrases.

3.5 Syntactic hedges

Syntactic hedges constitute the third category of hedges we propose, given structural differences on the sentence level from Single-hedges and NCK hedges. We have considered in this study the classification of conditionals made by Iatridou (1991) (relevance, factual and hypothetical conditionals), and we have have applied tests proposed by her to manually identify which conditionals convey hedging. For instance, the writer is not expressing uncertainty in (19) about the interlocutor wanting to ask questions, but the speech act taking place is for making him or her aware that he would answer in case when further questions arise. The use of *if* in the previous example differs from its use in (20), where it helps to state a hypothesis.

(19) **If** you have any questions, feel free to ask.

(20) I'm curious what your system profile is, and **if** there is a potential incompatibility here.

In the annotation task at hand, it is not possible to have access to complete dialogue that takes place starting with a question, comment or announcement. Having access to the full text written where the conditional occurs might enable one to determine if this corresponds to the hypothetical type of conditional. Nonetheless, we have to acknowledge that the amount of effort involved on deciding about the speech act qualities of conditionals increases the time required by manual annotation.

4 Findings in annotation of hedges in a web forum dataset

Our annotation dataset was composed of 3,000 web forum posts, from which interrogative sentences, quotations and non-processable sentences[6] were dropped out, leaving a total of 16,720 sentences. To ease the manual annotation task, an initial set of hedge expressions was used to pre-annotate this dataset; however, the final set of hedging expressions surpassed in number and variety this initial set. In this dataset, we found 790 unique types of hedges, 272 of them belong to the Single hedge category, 300 to NCK phrases, 8 to Syntactic and 210 to miscellaneous hedges. In Table 1, we show some frequent types of Single hedges and NCK epistemic phrases.

Recall that we think the annotation scheme should be able to identify the stance of the author of annotated sentences. Thus, the first person pronouns *I*, *we*, *me* and possessive pronoun *my* were targeted to identify (NCK) epistemic phrases. We have classified these phrases in three categories: a) Primary epistemic phrases, b) Semantically extended epistemic phrases, and c) Lexically extended epistemic phrases. The annotation scheme does not exhaust this taxonomy, however these categories are relevant to give a characterization of epistemic phrases found in an informal domain style.

The Primary type of epistemic phrase are expressions composed of a subject and an epistemic lexical verb conveying speculation as a main verb. *I think* and *I hope* in Table 1 are some examples of this kind of phrases. In these examples, the main verb can be categorized as a Single-hedge; it could be identified by

[5] We are aware that *think* may have a non-speculative reading in this sentence.

[6] Non-processable sentences are mostly noisy text composed by non linguistic tokens.

Table 1: Frequency of hedge types for Single and NCK hedges[1] in the annotation dataset of posts.

	Original	Freq.	Original	Freq.
Single hedges	would	441	'd	48
	world	1	wuuld	1
	Normalized: would		Subtotal	491
	try	175	tried	147
	trying	111	tries	17
	Normalized (try)		Subtotal	450
	some	396	396	
	other	305	others	52
	Normalized (other)		Subtotal	357
	may	155	maybe	93
	may be	71		
	Normalized (may)		Subtotal	319
	suggested	49	suggests	5
	suggest	3	suggesting	3
	Normalized: suggest		Subtotal	60
	assuming	3	assume	2
	Normalized (assume)		Subtotal	5

	Original	Freq.	Original	Freq.
NCK phrases	i think	157	i thought	35
	i 'm thinking	6	i * think	5
	i thing	1	i was thinking	1
	still think	1	i am thinking	1
	i now think	1	think	1
	i thinh	1		
	Normalized (i think)		Subtotal	210
	i hope	59	hope	49
	i was hoping	4	i 'm hoping	3
	i sure hope	2	i do hope	2
	i just hope	2	i hoped	1
	i am hoping	1	i had hoped	1
	i am hopeful	1		
	Normalized (i hope)		Subtotal	125
	i do n't know	43	i dont know	8
	i do not know	8	do n't know	6
	i did n't know	5	did n't know	3
	Normalized (i do not know)		Subtotal	89[2]

[1] Lexical types such as *wuuld* and *i dont know* are provided verbatim. They reflect the variety of hedge types in the dataset.

[2] This is a condensed set of hedge types provided for the sake of brevity.

an algorithm aiming to detect hedging based on traditional hedges. The Semantically Extended epistemic phrase category contains phrases equivalent in meaning to the Primary types: the main lexical verb do not necessarily convey uncertainty, but as a whole the phrase conveys uncertainty, such as in *I don't know*. For instance, *know* is an epistemic verb but does not convey a sense of uncertainty and as it is not used for hedging. The negated counterparts of such verbs are equivalent to a primary type of epistemic verb conveying uncertainty as in (Holmes, 1988) – *not know* is categorized as an epistemic verb expressing epistemic modality. Nonetheless, we can easily see that negating *know* is quite versatile, e.g. *never/seldom/hardly/scarcely know*, *improbable (that) (personal pronoun) know(s)*, *hard to know*, etc. The same versatility can be thought of the case of *remember* and *see*. Informal contractions such as *dunno* are included in this category. To this category also belong objective epistemic phrases also known as non-factives (eg. *understand*) which do not presuppose the factivity of the embedded proposition.

The lexically extended epistemic phrase category includes phrases where the main epistemic component is not a verb, but the epistemic force is conveyed by another constituent such as a noun or adjective such as *I am hopeful* – see Table 1. Other NCK phrases are *AFAIK*, *IMO*, and *I am not sure*.

Some normalization techniques reduced the number of lexically extended epistemic phrases to primary type in the case of NCK, and in both Single and NCK categories normalization causes grouping of equivalent types; in Table 1 these groups and their normalized type are shown. Normalization strategies address a) standard and non-standard abbreviations, b) contractions, c) typographical errors and misspellings, d) tense and number variations, e) colloquial forms, and f) inclusion of modifiers. Single hedges were normalized from 272 to 189 types and NCK were normalized from 300 to 137. Some of these groups have particularly a broad range of types such as the group for *I do not know* which has 20 different lexical types of NCK epistemic phrases, including colloquial forms such as *dunno* and *donno*. We believe that these normalisation techniques and normalized groupings are useful to design strategies for the automatic detection of hedges in informal language styles.

Syntactic hedges comprise seven lexical types such as *if*, *or* and *when* composing a total of 1,307 occurrences.

A fourth category comprises miscellaneous hedges that could not be classified as any of the former categories. In this group, 314 occurrences are spread over 210 types (1̃.5 occurrences per type). They are expressions such as *fingers crossed*, numerical ranges and NCK-like phrases that are specifically related to the main discussion topic in the web forum such as *I am not a techie*, and *I am technically challenged*. We kept this kind of phrases aside because we wanted to ensure the set of NCK phrases is topic-neutral.

Regarding other annotated entities, we found that 3.57% (286 occurrences) of hedges have a source that is not the writer of the sentence where the hedge is used. Although minimal, this shows the existence of cases where the mental state expressed by a hedge use does not reflect the writer's one. Occurrences of hedges without scope associated to them such as *somebody* and *confusion* make up to 18.26% (1,496) of overall cases. Further analysis of these cases would be needed to determine if their presence in a sentence changes the value of certainty conveyed by it.

5 Concluding remarks

This paper presented an annotation scheme of hedges for informal language, where the main category is constituted by first person epistemic expressions (see Appendix A for a specification of the annotation scheme elements). We have shown that this kind of expressions has a distinctive character in the expression of hedging, different from hedges in form of epistemic modals and other hedges commonly used in texts that have a formal register. The variety of forms in these phrases reveals this is a relevant category of hedging in domains with informal registers, such as web fora. The expressions characterized here provide structures that can be exploited for the automatic identification of hedges in noisy text, where automatic deep grammatical characterisation is a tough problem. Distinctions between NCK phrases and epistemic phrases with subjects other than first person, and other classes of hedges in first person constructions were drawn. Both kind of distinctions focus on the writer's involvement in a hedging event.

We have proposed the annotation of scope in a way that sentence constituents unrelated to hedging can be excluded in annotation, which we believe contributes to the precision of scope annotation. There are many paths of future research, mainly, finding a way to assess the hedging quality of the expressions found so far by additional independent judges. This was not done this so far, because it requires specialized knowledge about the domain and because the annotation process has so far been exploratory in verifying whether Not-Claiming-Knowledge epistemic phrases are a prevailing category of hedges in the informal register used in web forums. Another path for further development is the automation of some techniques manually done so far in a way detection of hedges can be done in noisy texts.

6 Acknowledgements

This research is supported by the Science Foundation Ireland CNGL (Grant 07/CE/I1142) as part of the Centre for Next Generation Localisation (www.cngl.ie) at Trinity College Dublin and by the Trinity College Research Scholarship Program.

References

Farkas, R., V. Vincze, G. Mra, J. Csirik, and G. Szarvas (2010, July). 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*, Uppsala, Sweden, pp. 1–12. ACL.

Ganter, V. and M. 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*, Suntec, Singapore, pp. 173–176.

Hendrickx, I., A. Mendes, and S. Mencarelli (2012). Modality in text: a proposal for corpus annotation. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Holmes, J. (1988). Doubt and certainty in ESL textbooks. *Applied Linguistics 9*(1), 21–44.

Hyland, K. (1998). *Hedging in Scientific Research Articles.* Pragmatics & beyond. John Benjamins Publishing Company.

Iatridou, S. (1991). *Topics in Conditionals.* Ph. D. thesis, MIT, Cambridge, Massachusetts. Distributed by MIT Working Papers in Linguistics.

Kärkkäinen, E. (2010). Position and scope of epistemic phrases in planned and unplanned american english. In *New approaches to hedging*, pp. 207–241. Amsterdam: Elsevier.

Light, M., X. T. Qui, and P. Srinivasan (2004). The language of bioscience: Facts, speculations, and statements in between. *Proceedings of BioLink 2004 Workshop on Linking Biological Literature, Ontologies and Databases: Tools for Users*, 17 – 24.

Medlock, B. and T. Briscoe (2007). Weakly supervised learning for hedge classification in scientific literature. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, Prague, Czech Republic, pp. 992–999.

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

Scheibman, J. (2001). Local patterns of subjectivity in person and verb type in. In J. L. Bybee and P. Hopper (Eds.), *Frequency and the Emergence of Linguistic Structure*, Volume 45 of *Typological studies in language*, pp. 61–89. Amsterdam; Philadelphia: John Benjamins Publishing Company.

Szarvas, G., V. Vincze, R. Farkas, and J. Csirik (2008). The BioScope corpus: annotation for negation, uncertainty and their scope in biomedical texts. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*, Columbus, Ohio, pp. 38–45.

Thompson, S. A. and A. Mulac (1991). A quantitative perspective on the grammaticization of epistemic parentheticals in english. In *Approaches to Grammaticalization*, pp. 314–329. John Benjamins.

Vincze, V. (2013, October). Weasels, hedges and peacocks: Discourse-level uncertainty in wikipedia articles. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, Nagoya, Japan, pp. 383–391. Asian Federation of Natural Language Processing.

Vincze, V., I. K. Simkó, and V. Varga (2014). *Proceedings of LAW VIII - The 8th Linguistic Annotation Workshop*, Chapter Annotating Uncertainty in Hungarian Webtext, pp. 64–69. Association for Computational Linguistics and Dublin City University.

Vincze, V., G. Szarvas, R. Farkas, G. Mora, and J. Csirik (2008). The BioScope corpus: biomedical texts annotated for uncertainty, negation and their scopes. *BMC Bioinformatics 9*(Suppl 11), S9.

Wiebe, J., T. Wilson, and C. Cardie (2005). Annotating expressions of opinions and emotions in language ANN. *Language Resources and Evaluation 39*(2/3), 164–210.

Wierzbicka, A. (2006). *English: meaning and culture.* Oxford University Press, USA.

Wilson, T. and J. Wiebe (2005). Annotating attributions and private states. In *Proceedings of the Workshop on Frontiers in Corpus Annotations II: Pie in the Sky*, CorpusAnno'05, Stroudsburg, PA, USA, pp. 53–60. ACL.

Appendix A Annotation scheme syntax

```
<entity> => <hedge> | <non-hedge> | <scope>| <source>
<hedge> => <single> | <not-claiming-knowledge> | <syntactic> | <miscellaneous>
<relations> => <source-of> | <scope-is>
<attribute> => <inner-epistemic-source>
<inner-epistemic-source> => writer | other

<has-attribute> => (<hedge>,<attribute>)
<source-of> => (<hedge>,<source>)
<scope-is> =>(<hedge>,<scope>)
```