Cross-Genre and Cross-Domain Detection of Semantic Uncertainty

György Szarvas*
Technische Universität Darmstadt

Veronika Vincze**
Hungarian Academy of Sciences

Richárd Farkas†
Universität Stuttgart

György Móra‡
University of Szeged

Iryna Gurevych*
Technische Universität Darmstadt

Uncertainty is an important linguistic phenomenon that is relevant in various Natural Language Processing applications, in diverse genres from medical to community generated, newswire or scientific discourse, and domains from science to humanities. The semantic uncertainty of a proposition can be identified in most cases by using a finite dictionary (i.e., lexical cues) and the key steps of uncertainty detection in an application include the steps of locating the (genre- and domain-specific) lexical cues, disambiguating them, and linking them with the units of interest for the particular application (e.g., identified events in information extraction). In this study, we focus on the genre and domain differences of the context-dependent semantic uncertainty cue recognition task.

We introduce a unified subcategorization of semantic uncertainty as different domain applications can apply different uncertainty categories. Based on this categorization, we normalized the annotation of three corpora and present results with a state-of-the-art uncertainty cue recognition model for four fine-grained categories of semantic uncertainty.

* Technische Universität Darmstadt, Ubiquitous Knowledge Processing (UKP) Lab, TU Darmstadt - FB 20 Hochschulstrasse 10, D-64289 Darmstadt, Germany. E-mail: {szarvas,gurevych}@tk.informatik.tu-darmstadt.de.

** Hungarian Academy of Sciences, Research Group on Artificial Intelligence, Tisza Lajos krt. 103, 6720 Szeged, Hungary. E-mail: vinczev@inf.u-szeged.hu.

† Universität Stuttgart, Institut für Maschinelle Sprachverarbeitung. Azenbergstrasse 12, D-70174 Stuttgart, Germany. E-mail: farkas@ims.uni-stuttgart.de.

‡ University of Szeged, Department of Informatics, Árpád tér 2, 6720 Szeged, Hungary. E-mail: gymora@inf.u-szeged.hu.

Submission received: 6 April 2011; revised submission received: 1 October 2011; accepted for publication: 30 November 2011.

Our results reveal the domain and genre dependence of the problem; nevertheless, we also show that even a distant source domain data set can contribute to the recognition and disambiguation of uncertainty cues, efficiently reducing the annotation costs needed to cover a new domain. Thus, the unified subcategorization and domain adaptation for training the models offer an efficient solution for cross-domain and cross-genre semantic uncertainty recognition.

1. Introduction

In computational linguistics, especially in information extraction and retrieval, it is of the utmost importance to distinguish between uncertain statements and factual information. In most cases, what the user needs is factual information, hence uncertain propositions should be treated in a special way: Depending on the exact task, the system should either ignore such texts or separate them from factual information. In machine translation, it is also necessary to identify linguistic cues of uncertainty because the source and the target language may differ in their toolkit to express uncertainty (one language uses an auxiliary, the other uses just a morpheme). To cite another example, in clinical document classification, medical reports can be grouped according to whether the patient definitely suffers, probably suffers, or does not suffer from an illness.

There are several linguistic phenomena that are referred to as uncertainty in the literature. We consider propositions to which no truth value can be attributed, given the speaker's mental state, as instances of **semantic** uncertainty. In contrast, uncertainty may also arise at the **discourse** level, when the speaker intentionally omits some information from the statement, making it vague, ambiguous, or misleading. Determining whether a given proposition is uncertain or not may involve using a finite dictionary of linguistic devices (i.e., cues). Lexical cues (such as modal verbs or adverbs) are responsible for semantic uncertainty whereas discourse-level uncertainty may be expressed by lexical cues and syntactic cues (such as passive constructions) as well. We focus on four types of semantic uncertainty in this study and henceforth the term **cue** will be taken to mean **lexical cue**.

The key steps of recognizing semantically uncertain propositions in a natural language processing (NLP) application include the steps of locating lexical cues for uncertainty, disambiguating them (as not all occurrences of the cues indicate uncertainty), and finally linking them with the textual representation of the propositions in question. The linking of a cue to the textual representation of the proposition can be performed on the basis of syntactic rules that depend on the word class of the lexical cue, but they are independent of the actual application domain or text type where the cue is observed. The set of cues used and the frequency of their certain and uncertain usages are domain and genre dependent, however, and this has to be addressed if we seek to craft automatic uncertainty detectors. Here we interpret **genre** as the basic style and formal characteristics of the writing that is independent of its topic (e.g., scientific papers, newswire texts, or business letters), and **domain** as a particular field of knowledge and is related to the topic of the text (e.g., medicine, archeology, or politics).

Uncertainty cue candidates do not display uncertainty in all of their occurrences. For instance, the mathematical sense of *probable* is dominant in mathematical texts whereas its ratio can be relatively low in papers in the humanities. The frequency of the two distinct meanings of the verb *evaluate* (which can be a synonym of *judge* [an uncertain meaning] and *calculate*) is also different in the bioinformatics and cell biology domains. Compare:

(1) To **evaluate**$_{CUE}$ the PML/RARalpha role in myelopoiesis, transgenic mice expressing PML/RARalpha were engineered.

(2) Our method was **evaluated** on the Lindahl benchmark for fold recognition.

In this article we focus on the domain-dependent aspects of uncertainty detection and we examine the recognition of uncertainty cues in context. We do not address the problem of linking cues to propositions in detail (see, e.g., Chapman, Chu, and Dowling [2007] and Kilicoglu and Bergler [2009] for the information extraction case).

For the empirical investigation of the domain dependent aspects, data sets are required from various domains. To date, several corpora annotated for uncertainty have been constructed for different genres and domains (BioScope, FactBank, WikiWeasel, and MPQA, to name but a few). These corpora cover different aspects of uncertainty, however, being grounded on different linguistic models, which makes it hard to exploit cross-domain knowledge in applications. These differences in part stem from the varied application needs across application domains. Different types of uncertainty and classes of linguistic expressions are relevant for different domains. Although hypotheses and investigations form a crucial part of the relevant cases in scientific applications, they are less prominent in newswire texts, where beliefs and rumors play a major role. This finding motivates a more fine-grained treatment of uncertainty. In order to bridge the existing gaps between application goals, these typical cases need to be differentiated. A fine-grained categorization enables the individual treatment of each subclass, which is less dependent on domain differences than using one coarse-grained uncertainty class. Moreover, this approach enables each particular application to identify and select from a pool of models only those aspects of uncertainty that are relevant in the specific domain.

As one of the main contributions of this study, we propose a uniform subcategorization of semantic uncertainty in which all the previous corpus annotation works can be placed, and which reveals the fundamental differences between the currently existing resources. In addition, we manually harmonized the annotations of three corpora and performed the fine-grained labeling according to the suggested subcategorization so as to be able to perform cross-domain experiments.

An important factor in training robust cross-domain models is to focus on shallow features that can be reliably obtained for many different domains and text types, and to craft models that exploit the shared knowledge from different sources as much as possible, making the adaptation to new domains efficient. The study of learning efficient models across different domains is the subject of transfer learning and domain adaptation research (cf. Daumé III and Marcu 2006; Pan and Yang 2010). The domain adaptation setting assumes a target domain (for which an accurate model should be learned with a limited amount of labeled training data), a source domain (with characteristics different from the target and for which a substantial amount of labeled data is available), and an arbitrary supervised learning model that exploits both the target and source domain data in order to learn an improved target domain model.

The success of domain adaptation mainly depends on two factors: (i) the similarity of the target and source domains (the two domains should be sufficiently similar to allow knowledge transfer); and (ii) the application of an efficient domain adaptation method (which permits the learning algorithm to exploit the commonalities of the domains while preserving the special characteristics of the target domain).

As our second main contribution, we study the impact of domain differences on uncertainty detection, how this impact depends on the distance between target and source domains concerning their domains and genres, and how these differences can be reduced to produce accurate target domain models with limited annotation effort.

Because previously existing resources exhibited fundamental differences that made domain adaptation difficult,[1] to our knowledge this is the first study to analyze domain differences and adaptability in the context of uncertainty detection in depth, and also the first study to report consistently positive results in cross-training.

The main contributions of the current paper can be summarized as follows:

- We provide a uniform subcategorization of semantic uncertainty (with definitions, examples, and test batteries for annotation) and classify all major previous studies on uncertainty corpus annotation into the proposed categorization system, in order to reveal and analyze the differences.
- We provide a harmonized, fine-grained reannotation of three corpora, according to the suggested subcategorization, to allow an in-depth analysis of the domain dependent aspects of uncertainty detection.
- We compare the two state-of-the-art approaches to uncertainty cue detection (i.e., the one based on token classification and the one on sequence labeling models), using a shared feature set, in the context of the CoNLL-2010 shared task, to understand their strengths and weaknesses.[2]
- We train an accurate semantic uncertainty detector that distinguishes four fine-grained categories of semantic uncertainty (epistemic, doxastic, investigation, and condition types) and thus is better for future applications in various domains than previous models. Our experiments reveal that, similar to the best model of the CoNLL-2010 shared task for biological texts but in a fine-grained context, shallow features provide good results in recognizing semantic uncertainty. We also show that this representation is less suited to detecting discourse-level uncertainty (which was part of the CoNLL task for Wikipedia texts).
- We examine in detail the differences between domains and genres as regards the language used to express semantic uncertainty, and learn how the domain or genre distance affects uncertainty recognition in texts with unseen characteristics.
- We apply domain adaptation techniques to fully exploit out-of-domain data and minimize annotation costs to adapt to a new domain, and we report successful results for various text domains and genres.

The rest of the paper is structured as follows. In Section 2, our classification of uncertainty phenomena is presented in detail and it is compared with the concept of

[1] Only 3 out of the more than 20 participants of the related CoNLL-2010 shared task (Farkas et al. 2010) managed to exploit out-of-domain data to improve their results, and only by a negligible margin.

[2] The most successful CoNLL systems were based on these approaches, but different feature representations make direct comparisons difficult.

uncertainty used in existing corpora. A framework for detecting semantic uncertainty is then presented in Section 3. Related work on cue detection is summarized in Section 4, which is followed by a description of our cue recognition system and a presentation of our experimental set-up using various source and target genre and domain pairs for cross-domain learning and domain adaptation in Section 5. Our results are elaborated on in Section 6 with a focus on the effect of domain similarities and on the annotation effort needed to cover a new domain. We then conclude with a summary of our results and make some suggestions for future research.

2. The Phenomenon *Uncertainty*

In order to be able to introduce and discuss our data sets, experiments, and findings, we have to clarify our understanding of the term *uncertainty*. Uncertainty—in its most general sense—can be interpreted as lack of information: The receiver of the information (i.e., the hearer or the reader) cannot be certain about some pieces of information. In this respect, uncertainty differs from both factuality and negation; as regards the former, the hearer/reader is sure that the information is true and as for the latter, he is sure that the information is not true. From the viewpoint of computer science, uncertainty emerges due to partial observability, nondeterminism, or both (Russell and Norvig 2010). Linguistic theories usually associate the notion of modality with uncertainty: Epistemic modality encodes how much certainty or evidence a speaker has for the proposition expressed by his utterance (Palmer 1986) or it refers to a possible state of the world in which the given proposition holds (Kiefer 2005). The common point in these approaches is that in the case of uncertainty, the truth value/reliability of the proposition cannot be decided because some other piece of information is missing. Thus, uncertain propositions are those in our understanding whose truth value or reliability cannot be determined due to lack of information.

In the following, we focus on semantic uncertainty and we suggest a tentative classification of several types of semantic uncertainty. Our classification is grounded on the knowledge of existing corpora and uncertainty recognition tools and our chief goal here is to provide a computational linguistics-oriented classification. With this in mind, our subclasses are intended to be well-defined and easily identifiable by automatic tools. Moreover, this classification allows different applications to choose the subset of phenomena to be recognized in accordance with their main task (i.e., we tried to avoid an overly coarse or fine-grained categorization).

2.1 Classification of Uncertainty Types

Several corpora annotated for uncertainty have been published in different domains such as biology (Medlock and Briscoe 2007; Kim, Ohta, and Tsujii 2008; Settles, Craven, and Friedland 2008; Shatkay et al. 2008; Vincze et al. 2008; Nawaz, Thompson, and Ananiadou 2010), medicine (Uzuner, Zhang, and Sibanda 2009), news media (Rubin, Liddy, and Kando 2005; Wilson 2008; Saurí and Pustejovsky 2009; Rubin 2010), and encyclopedia (Farkas et al. 2010). As can be seen from publicly available annotation guidelines, there are many overlaps but differences as well in the understanding of uncertainty, which is sometimes connected to domain- and genre-specific features of the texts. Here we introduce a domain- and genre-independent classification of several types of semantic uncertainty, which was inspired by both theoretical and computational linguistic considerations.

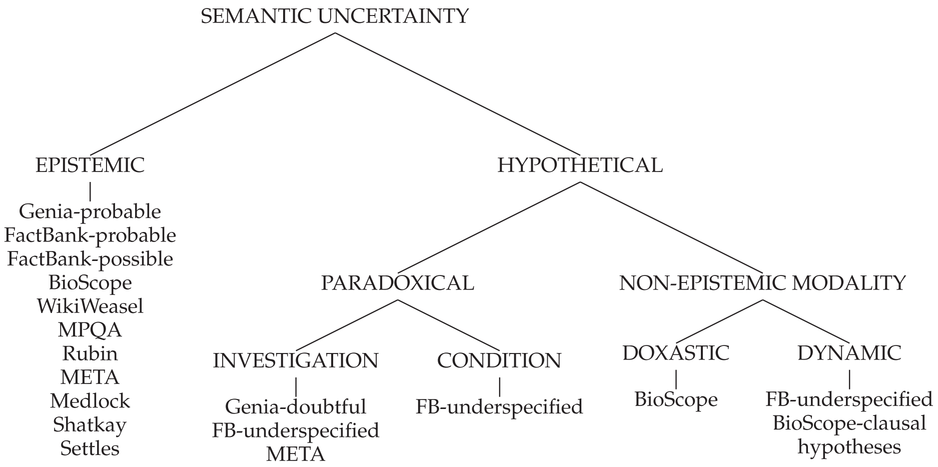

Figure 1
Types of uncertainty. FB = FactBank; Genia = Genia Event; Rubin = the data set described in Rubin, Liddy and Noriko (2005); META = the data set described in Nawaz, Thompson and Ananiadou (2010); Medlock = the data set described in Medlock and Briscoe (2007); Shatkay = the data set described in Shatkay et al. (2008); Settles = the data set described in Settles et al. (2008).

2.1.1 A Tentative Classification. Based on corpus data and annotation principles, the expression *uncertainty* can be used as an umbrella term for covering phenomena at the semantic and discourse levels.[3] Our classification of semantic uncertainty is assumed to be language-independent, but our examples presented here come from the English language, to keep matters simple.

Semantically uncertain propositions can be defined in terms of truth conditional semantics. They cannot be assigned a truth value (i.e., it cannot be stated for sure whether they are true or false) given the speaker's current mental state.

Semantic level uncertainty can be subcategorized into *epistemic* and *hypothetical* (see Figure 1). The main difference between epistemic and hypothetical uncertainty is that whereas instances of hypothetical uncertainty can be true, false or uncertain, epistemically uncertain propositions are definitely uncertain—in terms of possible worlds, hypothetical propositions allow that the proposition can be false in the actual world, but in the case of epistemic uncertainty the factuality of the proposition is not known.

In the case of **epistemic** uncertainty, it is known that the proposition is neither true nor false: It describes a possible world where the proposition holds but this possible world does not coincide with the speaker's actual world. In other words, it is certain that the proposition is uncertain. Epistemic uncertainty is related to epistemic modality: a sentence is epistemically uncertain if on the basis of our world knowledge we cannot decide at the moment whether it is true or false (hence the name) (Kiefer 2005). The source of an epistemically uncertain proposition cannot claim the uncertain proposition and be sure about its opposite at the same time.

(3) EPISTEMIC: It **may** be raining.

[3] The entire typology of semantic uncertainty phenomena and a test battery for their classification are described in a supplementary file. Together with the corpora and the experimental software, they are available at `http://www.inf.u-szeged.hu/rgai/uncertainty`.

As for **hypothetical** uncertainty, the truth value of the propositions cannot be determined either and nothing can be said about the probability of their happening. Propositions under *investigation* are an example of such statements: Until further analysis, the truth value of the proposition under question cannot be stated. **Conditionals** can also be classified as instances of hypotheses. It is also common in these two types of uncertain propositions that the speaker can utter them while it is certain (for others or even for him) that its opposite holds hence they can be called instances of paradoxical uncertainty.

Hypothetical uncertainty is connected with non-epistemic types of modality as well. **Doxastic modality** expresses the speaker's beliefs—which may be known to be true or false by others in the current state of the world. Necessity (duties, obligation, orders) is the main objective of deontic modality; dispositional modality is determined by the dispositions (i.e., physical abilities) of the person involved; and circumstantial modality is defined by external circumstances. Buletic modality is related to wishes, intentions, plans, and desires. An umbrella term for deontic, dispositional, circumstantial, and buletic modality is **dynamic modality** (Kiefer 2005).

HYPOTHETICAL:

(4) DYNAMIC: I **have to** go.

(5) DOXASTIC: He **believes** that the Earth is flat.

(6) INVESTIGATION: We **examined** the role of NF-kappa B in protein activation.

(7) CONDITION: **If** it rains, we**'ll** stay in.

Conditions and instances of dynamic modality are related to the future: In the future, they may happen but at the moment it is not clear whether they will take place or not / whether they are true, false, or uncertain.

2.1.2 Comparison with other Classifications. The feasibility of the classification proposed in this study can be justified by mapping the annotation schemes used in other existing corpora to our subcategorizations of uncertainty. This systematic comparison also highlights the major differences between existing works and partly explains why examples for successful cross-domain application of existing resources and models are hard to find in the literature.

Most of the annotations found in biomedical corpora (Medlock and Briscoe 2007; Settles, Craven, and Friedland 2008; Shatkay et al. 2008; Thompson et al. 2008; Nawaz, Thompson, and Ananiadou 2010) fall into the epistemic uncertainty class. BioScope (Vincze et al. 2008) annotations mostly belong to the epistemic uncertainty category, with the exception of clausal hypotheses (i.e., hypotheses that are expressed by a clause headed by *if* or *whether*), which are instances of the investigation class. The *probable* class of Genia Event (Kim, Ohta, and Tsujii 2008) is of the epistemically uncertain type and the *doubtful* class belongs to the investigation class. Rubin, Liddy, and Kando (2005) consider uncertainty as a phenomenon belonging to epistemic modality: The high, moderate, and low levels of certainty coincide with our epistemic uncertainty category. The speculation annotations of the MPQA corpus also belong to the epistemic uncertainty class, with four levels (Wilson 2008). The *probable* and *possible* classes found in FactBank (Saurí and Pustejovsky 2009) are of the epistemically uncertain type, events with a generic source belong to discourse-level uncertainty, whereas underspecified events are classified as hypothetical uncertainty in our system as, by definition, their truth value cannot be determined. WikiWeasel (Farkas et al. 2010) contains annotation for epistemic uncertainty, but discourse-level uncertainty is also annotated in the corpus (see Figure 1 for an overview). The categories used for the machine reading task described in Morante and Daelemans (2011) also overlap with our fine-grained classes: Uncertain events in their system fall into our epistemic uncertainty class. Their modal events expressing purpose, need, obligation, or desire are instances of dynamic modality, whereas their conditions are understood in a similar way to our condition class. The modality types listed in Baker et al. (2010) can be classified as types of dynamic modality, except for their belief category. Instances of the latter category are either certain (*It is certain that he met the president*) or epistemic or doxastic modality in our system.

2.2 Types of Semantic Uncertainty Cues

We assume that the nature of the lexical unit determines the type of uncertainty it represents, that is, semantic uncertainty is highly lexical in nature. The part of speech of the uncertainty cue candidates serves as the basis for categorization, similar to the ones found in Hyland (1994, 1996, 1998) and Rizomilioti (2006). In English, modality is often associated with modal auxiliaries (Palmer 1979), but, as Table 1 shows, there are many other parts of speech that can express uncertainty. It should be added that there are cues where it depends on the context, rather than the given lexical item, what subclass of uncertainty the cue refers to, for example, *may* can denote epistemic modality (*It may rain...*) or dynamic modality (*Now you may open the door*). These categories are listed in Table 1.

3. A Framework for Detecting Semantic Uncertainty

In our model, uncertainty detection is a standalone task that is largely independent of the underlying application. In this section, we briefly discuss how uncertainty detection

Table 1
Uncertainty cues.

Adjectives / adverbs		
	probable, likely, possible, unsure, possibly, perhaps, etc.	epistemic
Auxiliaries		
	may, might, can, would, should, could, etc.	semantic
Verbs		
speculative:	suggest, question, seem, appear, favor, etc.	epistemic
psych:	think, believe, etc.	doxastic
analytic:	investigate, analyze, examine, etc.	investigation
prospective:	plan, want, order, allow, etc.	dynamic
Conjunctions		
	if, whether, etc.	investigation
Nouns		
nouns derived from uncertain verb:	speculation, proposal, consideration, etc.	same as the verb
other uncertain nouns:	rumor, idea, etc.	doxastic

can be incorporated into an information extraction task, which is probably the most relevant application area (see Kim et al. [2009] for more details). In the information extraction context, the key steps of recognizing uncertain propositions are locating the cues, disambiguating them (as not all occurrences of the cues indicate uncertainty; recall the example of *evaluate*), and finally linking them with the textual representation of the propositions in question. We note here that marking the textual representations of important propositions (often referred to as *events* in information extraction) is actually the main goal of an information extraction system, hence we will not focus on their identification and just assume that they are already marked in texts.

The following is an example that demonstrates the process of uncertainty detection:

(8) In this study we **hypothesized**$_{CUE}$ that the phosphorylation of TRAF2 *inhibits*$_{EVENT}$ binding to the CD40 cytoplasmic domain.

Here the EVENT mark-up is produced by the information extraction system, and uncertainty detection consists of i) the recognition of the cue word *hypothesized*, and determining whether it denotes uncertainty in this specific case (producing the CUE mark-up) and ii) determining whether the cue *hypothesized* modifies the event triggered by *inhibits* or not (positive example in this case).

3.1 Uncertainty Cue Detection and Disambiguation

The cue detection and disambiguation problem can be essentially regarded as a token labeling problem. Here the task is to assign a label to each of the tokens of a sentence in question according to whether it is the starting token of an uncertainty cue (`B-CUE_TYPE`), an inside token of a cue (`I-CUE_TYPE`), or it is not part of any cue (`O`). Most previous studies assume a binary classification task, namely, each token is either part of an uncertainty cue, or it is not a cue. For fine-grained uncertainty detection, a different label has to be used for each uncertainty type to be distinguished. This way, the label sequence of a sentence naturally identifies all uncertainty cues (with their types) in the sentence, and disambiguation is solved jointly with recognition.

Because the uncertainty cue vocabulary and the distribution of certain and uncertain senses of cues vary in different domains and genres, uncertainty cue detection and disambiguation are the main focus of the current study.

3.2 Linking Uncertainty Cues to Propositions

The task of linking the detected uncertainty cues to propositions can be formulated as a binary classification task over uncertainty cue and event marker pairs. The relation holds and is considered true if the cue modifies the truth value (confidence) of the event; it does not hold and is considered false if the cue does not have any impact on the interpretation of the event. That is, the pair *(hypothesized, inhibits)* in Example (8) is an instance of positive relation.

The linking of uncertainty cues and event markers can be established by using dependency grammar rules (i.e., the problem is mainly syntax driven). As the grammatical properties of the language are similar in various domains and genres, this task is largely domain-independent, as opposed to the recognition and disambiguation task. Because of this, we sketch the most important matching patterns, but do not address the linking task in great detail here.

The following are the characteristic rules that can be used to link uncertainty cues to event markers. For practical implementations of heuristic cue/event matching, see Chapman, Chu, and Dowling (2007) and Kilicoglu and Bergler (2009).

- If the event clue has an uncertain verb, noun, preposition, or auxiliary as a (not necessarily direct) parent in the dependency graph of the sentence, the event is regarded as uncertain.
- If the event clue has an uncertain adverb or adjective as its child, it is treated as uncertain.

4. Related Work on Uncertainty Cue Detection

Here we review the published works related to uncertainty cue detection. Earlier studies focused either on in-domain cue recognition for a single domain or on cue lexicon extraction from large corpora. The latter approach is applicable to multiple domains, but does not address the disambiguation of uncertain and other meanings of the extracted cue words. We are also aware of several studies that discussed the differences of cue distributions in various domains, without developing a cue detector. To the best of our knowledge, our study is the first to address the genre- and domain-adaptability of uncertainty cue recognition systems and thus uncertainty detection in a general context.

We should add that there are plenty of studies on end-application oriented uncertainty detection, that is, how to utilize the recognized cues (see, for instance, Kilicoglu and Bergler [2008], Uzuner, Zhang, and Sibanda [2009] and Saurí [2008] for information extraction or Farkas and Szarvas [2008] for document labeling applications), and a recent pilot task sought to exploit negation and hedge cue detectors in machine reading (Morante and Daelemans 2011). As the focus of our paper is cue recognition, however, we omit their detailed description here.

4.1 In-Domain Cue Detection

In-domain uncertainty detectors have been developed since the mid 1990s. Most of these systems use hand-crafted lexicons for cue recognition and they treat each occurrence of the lexicon items as a cue—that is, they do not address the problem of disambiguating cues (Friedman et al. 1994; Light, Qiu, and Srinivasan 2004; Farkas and Szarvas 2008; Saurí 2008; Conway, Doan, and Collier 2009; Van Landeghem et al. 2009). ConText (Chapman, Chu, and Dowling 2007) uses regular expressions to define cues and "pseudo-triggers". A pseudo-trigger is a superstring of a cue and it is basically used for recognizing contexts where a cue does not imply uncertainty (i.e., it can be regarded as a hand-crafted cue disambiguation module). MacKinlay, Martinez, and Baldwin (2009) introduced a system which also used non-consecutive tokens as cues (like *not+as+yet*).

Utilizing manually labeled corpora, machine learning–based uncertainty cue detectors have also been developed (to the best of our knowledge each of them uses an in-domain training data set). They use token classification (Morante and Daelemans 2009; Clausen 2010; Fernandes, Crestana, and Milidiú 2010; Sánchez, Li, and Vogel 2010) or sequence labeling approaches (Li et al. 2010; Rei and Briscoe 2010; Tang et al. 2010; Zhang et al. 2010). In both cases the tokens are labeled according to whether they are part of a cue. The latter assigns a label sequence to a sentence (a sequence of tokens) thus it naturally deals with the context of a particular word. On the other hand, context information for a token is built into the feature space of the token classification approaches. Özgür and Radev (2009) and Velldal (2010) match cues from a lexicon then apply a binary classifier based on features describing the context of the cue candidate.

Each of these approaches uses a rich feature representation for tokens, which usually includes surface-level, part-of-speech, and chunk-level features. A few systems have also used dependency relation types originating at the cue (Rei and Briscoe 2010; Sánchez, Li, and Vogel 2010; Velldal, Øvrelid, and Oepen 2010; Zhang et al. 2010); the CoNLL-2010 Shared Task final ranking suggests that it has only a limited impact on the performance of an entire system (Farkas et al. 2010), however. Özgür and Radev (2009) further extended the feature set with the other cues that occur in the same sentence as the cue, and positional features such as the section header of the article in which the cue occurs (the latter is only defined for scientific publications). Velldal (2010) argues that the dimensionality of the uncertainty cue detection feature space is too high and reports improvements by using the sparse random indexing technique.

Ganter and Strube (2009) proposed a rather different approach for (weasel) cue detection—exploiting weasel tags[4] in Wikipedia articles given by editors. They used syntax-based patterns to recognize the internal structure of the cues, which has proved useful as discourse-level uncertainty cues are usually long and have a complex internal structure (as opposed to semantic uncertainty cues).

As can be seen, uncertainty cue detectors have mostly been developed in the biological and medical domains. All of these studies, however, focus on only one domain, namely, in-domain cue detection is carried out, which assumes the availability of a training data set of sufficient size. The only exception we are aware of is the CoNLL-2010 Shared Task (Farkas et al. 2010), where participants had the chance to use Wikipedia data on biomedical domain and vice versa. Probably due to the differences in the annotated uncertainty types and the stylistic and topical characteristics of the texts, very few participants performed cross-domain experiments and reported only limited success (see Section 5.3.2 for more on this).

Overall, the findings of these studies indicate that disambiguating cue candidates is an important aspect of uncertainty detection and that the domain specificity of disambiguation models and domain adaptation in general are largely unexplored problems in uncertainty detection.

4.2 Weakly Supervised Extraction of Cue Lexicon

Similar to our approach, several studies have addressed the problem of developing an uncertainty detector for a new domain using as little annotation effort as possible. The aim of these studies is to identify uncertain sentences; this is carried out by semi-automatic construction of cue lexicons. The weakly supervised approaches start with very small seed sets of annotated certain and uncertain sentences, and use bootstrapping to induce a suitable training corpus in an automatic way. Such approaches collect potentially certain and uncertain sentences from a large unlabeled pool based on their similarity to the instances in the seed sets (Medlock and Briscoe 2007), or based on the known errors of an information extraction system that is itself sensitive to uncertain texts (Szarvas 2008). Further instances are then collected (in an iterative fashion) on the basis of their similarity to the current training instances. Based on the observation

[4] See http://en.wikipedia.org/wiki/Wikipedia:Embrace_weasel_words.

that uncertain sentences tend to contain more than one uncertainty cue, these models successfully extend the seed sets with automatically labeled sentences, and can produce an uncertainty classifier with a sentence-level F-score of 60–80% for the uncertain class, given that the texts of the seed examples, the unlabeled pool, and the actual evaluation data share very similar properties.

Szarvas (2008) showed that these models essentially learn the uncertainty lexicon (set of cues) of the given domain, but are otherwise unable to disambiguate the potential cue words—that is, to distinguish between the uncertain and certain uses of the previously seen cues. This deficiency of the derived models is inherent to the bootstrapping process, which considers all occurrences of the cue candidates as good candidates for positive examples (as opposed to unlabeled sentences without any previously seen cue words).

Kilicoglu and Bergler (2008) proposed a semi-automatic method to expand a seed cue lexicon. Their linguistically motivated approach is also based on the weakly supervised induction of a corpus of uncertain sentences. It exploits the syntactic patterns of uncertain sentences to identify new cue candidates.

The previous studies on weakly supervised approaches to uncertainty detection did not tackle the problem of disambiguating the certain and uncertain uses of cue candidates, which is a major drawback from a practical point of view.

4.3 Cue Distribution Analyses

Besides automatic uncertainty recognition, several studies investigated the distribution of hedge cues in scientific papers from different domains (Hyland 1998; Falahati 2006; Rizomilioti 2006). The effect of different domains on the frequency of uncertain expressions was examined in Rizomilioti (2006). Based on a previously defined dictionary of hedge cues, she analyzed the linguistic tools expressing epistemic modality in research papers from three domains, namely, archeology, literary criticism, and biology. Her results indicated that archaeological papers tend to contain the most uncertainty cues (which she calls downtoners) and the fewest uncertainty cues can be found in literary criticism papers. Different academic disciplines were contrasted in Hyland (1998) from the viewpoint of hedging: Papers belonging to the humanities contain significantly more hedging devices than papers in sciences. It is interesting to note, however, that in both studies, biological papers are situated in the middle as far as the percentage rate of uncertainty cues is concerned. Falahati (2006) examined hedges in research articles in medicine, chemistry, and psychology and concluded that it is psychology articles that contain the most hedges.

Overall, these studies demonstrate that there are substantial differences in the way different technical/scientific domains and different genres express uncertainty in general, and in the use of semantic uncertainty in particular. Differences are found not just in the use of different vocabulary for expressing uncertainty, but also in the frequency of certain and uncertain usage of particular uncertainty cues. These findings underpin the practical importance of domain portability and domain adaptation of uncertainty detectors.

5. Uncertainty Cue Recognition

In this section, we present our uncertainty cue detector and the results of the cross-genre and -domain experiments carried out by us. Before describing our model and discussing the results of the experiments, a short overview of the texts used as training and test data sets will be given along with an empirical analysis of the sense distributions of the most frequent cues.

5.1 Data Sets

In our investigations, we selected three corpora (i.e., BioScope, WikiWeasel, and FactBank) from different domains (biomedical, encyclopedia, and newswire, respectively). Genres also vary in the corpora (in the scientific genre, there are papers and abstracts whereas the other corpora contain pieces of news and encyclopedia articles). We preferred corpora on which earlier experiments had been carried out because this allowed us to compare our results with those of previous studies. This selection makes it possible to investigate domain and genre differences because each domain has its characteristic language use (which might result in differences in cue distribution) and different genres also require different writing strategies (e.g., in abstracts, implications of experimental results are often emphasized, which usually involves the use of uncertain language).

The BioScope corpus (Vincze et al. 2008) contains clinical texts as well as biological texts from full papers and scientific abstracts; the texts were manually annotated for hedge cues and their scopes. In our experiments, 15 other papers annotated for the CoNLL-2010 Shared Task (Farkas et al. 2010) were also added to the set of BioScope papers. The WikiWeasel corpus (Farkas et al. 2010) was also used in the CoNLL-2010 Shared Task and it was manually annotated for weasel cues and semantic uncertainty in randomly selected paragraphs taken from Wikipedia articles. The FactBank corpus contains texts from the newswire domain (Saurí and Pustejovsky 2009). Events are annotated in the data set and they are evaluated on the basis of their factuality from the viewpoint of their sources.

Table 2 provides statistical data on the three corpora. Because in our experimental set-up, texts belonging to different genres also play an important role, data on abstracts and papers are included separately.

5.1.1 Genres and Domains. Texts found in the three corpora to be investigated can be categorized into three genres, which can be further divided to subgenres at a finer level of distinction. Figure 2 depicts this classification.

The majority of BioScope texts (papers and abstracts) belong to the scientific discourse genre. FactBank texts can be divided into broadcast and written news, and Wikipedia texts belong to the encyclopedia genre.

As for the domain of the texts, there are three broad domains, namely, biology, news, and encyclopedia. Once again, these domains can be further divided into narrower

Table 2
Data on the corpora. sent. = sentence; epist. = epistemic cue; dox. = doxastic cue; inv. = investigation cue; cond. = condition cue.

Data Set	#sent.	#epist.	#dox.	#inv.	#cond.	Total
BioScope papers	7676	1373	220	295	187	2075
BioScope abstracts	11797	2478	200	784	24	3486
BioScope total	19473	3851	420	1079	211	5561
WikiWeasel	20756	1171	909	94	491	3265
FactBank	3123	305	201	36	178	720
Total	43352	5927	1530	1209	880	9546

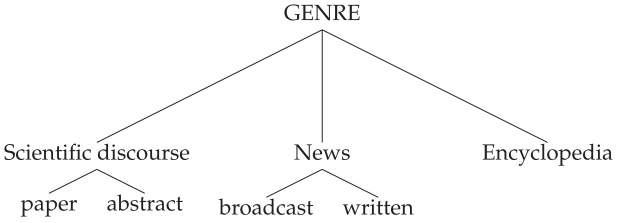

Figure 2
Genres of texts.

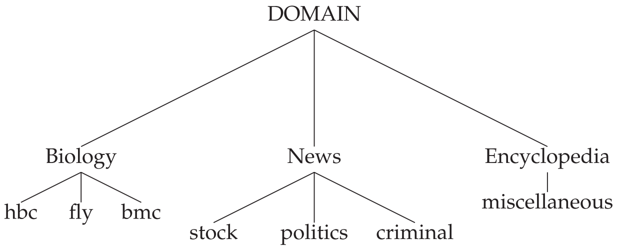

Figure 3
Domains of texts.

topics at a fine-grained level, which is shown in Figure 3. All abstracts and five papers in BioScope are related to the MeSH terms *human*, *blood cell*, and *transcription factor* (`hbc` in Figure 3). Nine BMC Bioinformatics papers come from the bioinformatics domain (`bmc` in Figure 3), and ten papers describe some experimental results on the Drosophila species (`fly`). FactBank news can be classified as stock news, political news, and criminal news. Encyclopedia articles cover a broad range of topics, hence no detailed classification is given here.

5.1.2 The Normalization of the Corpora. In order to uniformly evaluate our methods in each domain and genre (and each corpus), the evaluation data sets were normalized. This meant that cues had to be annotated in each data set and differentiated for types of semantic uncertainty. This resulted in the reannotation of BioScope, WikiWeasel, and FactBank.[5] In BioScope, the originally annotated cues were separated into epistemic cues and subtypes of hypothetical cues and instances of hypothetical uncertainty not yet marked were also annotated. In FactBank, epistemic and hypothetical cues were annotated: Uncertain events were matched with their uncertainty cues and instances of hypothetical uncertainty that were originally not annotated were also marked in the corpus. In the case of WikiWeasel, these two types of cues were separated from discourse-level cues.

One class of hypothetical uncertainty (i.e., dynamic modality) was not annotated in any of the corpora. Although dynamic modality seems to play a role in the news domain, it is less important and less represented in the other two domains we investigated here. The other subclasses are more of general interest for the applications. For example, one of our training corpora comes from the scientific domain, where it is more important to distinguish facts from hypotheses and propositions under investigation

[5] The corpora are available at `http://www.inf.u-szeged.hu/rgai/uncertainty`.

(which can be later confirmed or rejected, compare the meta-knowledge annotation scheme developed for biological events [Nawaz, Thompson, and Ananiadou 2010]), and from propositions that depend on each other (conditions).

5.1.3 Uncertainty Cues in the Corpora. An analysis of the cue distributions reveals some interesting trends that can be exploited in uncertainty detection across domains and genres. The most frequent cue stems in the (sub)corpora used in our study can be seen in Table 3 and they are responsible for about 74% of epistemic cue occurrences, 55% of doxastic cue occurrences, 70% of investigation cue occurrences, and 91% of condition cue occurrences.

As can be seen, one of the most frequent epistemic cues in each corpus is *may*. *If*, *possible*, *might*, and *suggest* also occur frequently in our data set.

The distribution of the uncertainty cues was also analyzed from the perspective of uncertainty classes in each corpus, which is presented in Figure 4. In most of the corpora, epistemic cues are the most frequent (except for FactBank) and they vary the most: Out of the 300 cue stems occurring in the corpora, 206 are epistemic cues. Comparing the domains, it can readily be seen that in biological texts, doxastic uncertainty is not frequent, which is especially true for abstracts, whereas in FactBank and WikiWeasel they cover about 27% of the data. The most frequent doxastic keywords exhibit some domain-specific differences, however: In BioScope, the most frequent ones include *putative* and *hypothesis*, which rarely occur in FactBank and WikiWeasel. Nevertheless, cues belonging to the investigation class can be found almost exclusively in scientific texts (89% of them are in BioScope), which can be expected because the aim of scientific publications is to examine whether a hypothesized phenomenon occurs. Among the most

Table 3
The most frequent cues in the corpora. epist. = epistemic cue; dox. = doxastic cue; inv. = investigation cue; cond. = condition cue.

	Global		Abstracts		Full papers		BioScope		FactBank		WikiWeasel	
Epist.	may	1508	suggest	616	may	228	suggest	810	may	43	may	721
	suggest	928	may	516	suggest	194	may	744	could	29	probable	112
	indicate	421	indicate	301	indicate	103	indicate	404	possible	26	suggest	108
	possible	304	appear	143	possible	84	appear	213	likely	24	possible	93
	appear	260	or	119	might	83	or	197	might	23	likely	80
	might	256	possible	101	or	78	possible	185	appear	15	might	78
	likely	221	might	72	can	73	might	155	seem	11	seem	67
	or	198	potential	72	appear	70	can	117	potential	10	could	55
	could	196	likely	60	likely	57	likely	117	probable	10	perhaps	51
	probable	157	could	56	could	56	could	112	suggest	10	appear	32
Dox.	consider	276	putative	43	putative	37	putative	80	expect	75	consider	250
	believe	222	think	43	hypothesis	33	hypothesis	77	believe	25	believe	173
	expect	136	hypothesis	43	assume	24	think	66	think	24	allege	81
	think	131	believe	14	think	24	assume	32	allege	8	think	61
	putative	83	consider	10	expect	22	predict	26	accuse	7	regard	58
Invest.	whether	247	investigate	177	whether	73	investigate	221	whether	26	whether	52
	investigate	222	examine	160	investigate	44	examine	183	if	3	if	20
	examine	183	whether	96	test	25	whether	169	remain to be seen	2	whether or not	7
	study	102	study	88	examine	23	study	101	question	1	assess	3
	determine	90	determine	67	determine	20	determine	87	determine	1	evaluate	3
Cond.	if	418	if	14	if	85	if	99	if	65	if	254
	would	238	would	6	would	46	would	52	would	50	would	136
	will	80	until	2	will	20	will	20	will	21	will	39
	until	40	could	1	should	11	should	11	until	16	until	15
	could	30	unless	1	could	9	could	10	could	9	unless	14

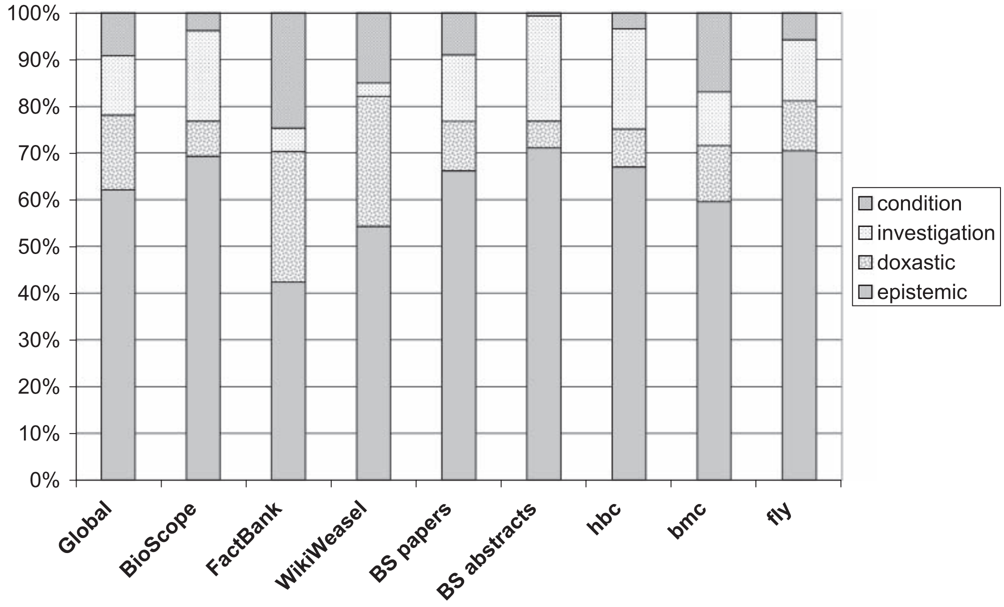

Figure 4
Cue type distributions in the corpora.

frequent cues, *investigate*, *examine*, and *study* belong to this group. These data reveal that the frequency of doxastic and investigation cues is strongly domain-dependent, and this explains the fact that the investigation vocabulary is very limited in Factbank and WikiWeasel. Only about 10 cue stems belong to this uncertainty class in these corpora. The set of condition cue stems, however, is very small in each corpus; altogether 18 condition cue stems can be found in the data, although *if* and *would* are responsible for almost 75% of condition cue occurrences. It should also be mentioned that the percentage of condition cues is higher in FactBank than in the other corpora.

Another interesting trend was observed when word forms were considered instead of stemmed forms: Certain verbs in third person singular (e.g., *expects* or *believes*) occur mostly in FactBank and WikiWeasel. The reason for this may be that when speaking about someone else's opinion in scientific discourse, the source of the opinion is usually provided in the form of references or citations—usually at the end of the sentence—and due to this, the verb is often used in the passive form, as in Example (9).

(9) It is currently **believed** that both RAG1 and RAG2 proteins were originally encoded by the same transposon recruited in a common ancestor of jawed vertebrates **[3,12,13,16]**.

In contrast, impersonal constructions are hardly used in news media, where the objective is to inform listeners about the source of the news presented as well in order to enable them to judge the reliability of a piece of news. Here, a clause including the source and a communication verb is usually attached to the proposition.

A genre-related difference between scientific abstracts and full papers is that condition cues can rarely be found in abstracts, although they occur more frequently in papers (with the non-cue usage still being much more frequent). Another difference is the percentage of cues of the investigation type, which may be related to the structure of abstracts. Biological abstracts usually present the problem they examine and describe methods they use. This entails the application of predicates belonging to the investigation class of uncertainty. It can be argued, however, that scientific papers also have these characteristics but abstracts are much shorter than papers (generally, they contain about 10–12 sentences). Hence, investigation cues are responsible for a greater percentage of cues.

There are some lexical differences among the corpora that are related to domain or genre specificity. For instance, due to their semantics, the words *charge*, *accuse*, *allege*, *fear*, *worry*, and *rumor* are highly unlikely to occur in scientific publications, but they occur relatively often in news texts and in Wikipedia articles. As for lexical divergences between abstracts and papers, many of them are related to verbs of investigation and their different usage. In the corpora, verbs of investigations were marked only if it was not clear whether the event/phenomenon would take place or not. If it has already happened (*The police are investigating the crime*) or the existence of the thing under investigation can be stated with certainty, independently of the investigation (*The top ten organisms were examined*), then they are not instances of hypotheses, so they were not annotated. As the data sets make clear, there were some candidates of investigation verbs that occurred in the investigation sense mostly in abstracts but in another sense in papers, especially in the `bmc` data set (e.g. *assess* or *examine*). *Evaluate* also had a special mathematical sense in `bmc` papers, which did not occur in abstracts.

It can also be seen that some of the very frequent cues in papers do not occur (or only relatively rarely) in abstracts. This is especially true for the `bmc` data set, where *can*, *if*, *would*, *could*, and *will* are among the 15 most frequent cues and represent 23.21% of cue occurrences, but only 3.85% in abstracts. It is also apparent that the rate of epistemic cues is lower in `bmc` papers than in abstracts or other types of papers.

Genre-dependent characteristics can be analyzed if BioScope abstracts and `hbc` papers are compared because their fine-grained domain is the same. Thus, it may be assumed that differences between their cues are related to the genre. The sets of cues used are similar, but the sense distributions may differ for certain ambiguous cues. For instance, *indicate* mostly appears in the 'suggest' sense in abstracts, whereas in papers it is used in the 'signal' sense. Another difference is that the percentage rate of doxastic cues is almost twice as high in papers as in abstracts (10.6% and 5.7%, respectively). Besides these differences, the two data sets are quite similar.

Domain-related differences can be analyzed when the three subdomains of biological papers are contrasted. As stressed earlier, `bmc` papers contain fewer instances of epistemic uncertainty, but condition cues occur more frequently in them. Nevertheless, `fly` and `hbc` papers are rather similar in these respects but `hbc` papers contain more investigation cues than the other two subcorpora. As regards lexical issues, the non-cue usage of *possible* in comparative constructions is more frequent in the `bmc` data set than in the other papers and many occurrences of *if* in `bmc` are related to definitions, which were not annotated as uncertain. On the basis of this information, the `fly` and the `hbc` domains seem to be more similar to each other than to the `BMC` data set from a linguistic point of view.

From the perspective of genre and domain adaptation, the following points should be highlighted concerning the distribution of uncertainty cues across corpora. Doxastic uncertainty is of primary importance in the news and encyclopedia domains whereas the investigation class is characteristic of the biological domain. Within the latter, there is a genre-related difference as well: It is the epistemic and investigation classes that are mainly present in abstracts whereas in papers cues belonging to other uncertainty classes can also be found. Thus, when applying techniques developed for biological texts or abstracts to news texts, for example, doxastic uncertainty cues deserve special attention as it might well be the case that there are insufficient training examples for this class of uncertainty cues. The adaptation of an uncertainty cue detector constructed for encyclopedia texts requires the special treatment of investigation cues, however, if, for instance, scientific discourse is the target genre since they are underrepresented in the source genre.

5.2 Evaluation Metrics

As evaluation metrics, we used cue-level and sentence-level $F_{\beta=1}$ scores for the uncertain class (the standard evaluation metrics of Task 1 of the CoNLL-2010 shared task) and denote them by F_{cue} and F_{sent}, respectively. We report cue-level $F_{\beta=1}$ scores on the individual subcategories of uncertainty and the unlabeled (binary) $F_{\beta=1}$ scores as well. A sentence is treated as uncertain (in the gold standard and prediction) iff it contains at least one cue. Note that the cue-level metric is quite strict as it is based on recognized phrases—that is, only cues with perfect boundary matches are true positives. For the sentence-level evaluation we simply labeled those sentences as uncertain that contained at least one recognized cue.

5.3 Cross-Domain Cue Recognition Model

In order to minimize the development cost of a labeled corpus and an uncertainty detector for a new genre/domain, we need to induce an accurate model from a minimal amount of labeled data, or take advantage of existing corpora for different genres and/or domains and use a domain adaptation approach. Experiments investigating the value and sufficiency of existing corpora (which are usually out-of-domain) and simple domain adaptation methods were carried out. For this purpose, we implemented a cue recognition model, which is described in this section.

To train our models, we applied surface level (e.g., capitalization) and shallow syntactic features (part-of-speech tags and chunks) and avoided the use of lexicon-based features listing potential cue words, in order to reduce the domain dependence of the learned models. Now we will introduce our model, which is competitive with the state-of-the-art systems and focus on its domain adaptability. We will also describe the implementation details of the learning model and the features employed. We should add that the optimization of a cue detector was not the main focus of our study, however.

5.3.1 Feature Set. We extracted two types of features for each token to describe the token itself, together with its local context in a window of limited size (1, 2, or no window, depending on the feature).

The first group consists of features describing the surface form of the tokens. Here we provide the list of the surface features with the corresponding window sizes:

- **Stems** of the tokens by the Porter stemmer in a window of size 2 (current token and two tokens to the left and right).
- **Surface pattern** of the tokens in a window of size one (current token and 1 token to the left and right). These patterns are similar to the *word shape* feature described in Sun et al. (2007). This feature can describe the capitalization and other orthographic features as well. Patterns represent

character sequences of the same type with one single character for a given word. There are six different pattern types denoting capitalized and lowercased character sequences with the characters "A" and "a", number sequences with "0", Greek letter sequences with "G" and "g", Roman numerals with "R" and "r", and non-alphanumerical characters with "!".

- **Prefixes and suffixes** of word forms from three to five characters long.

The second group of features describes the syntactic properties of the token and its local context. The list of the syntactic features with the corresponding window sizes is the following:

- **Part-of-speech** (POS) tags of the tokens by the C&C POS-tagger in a window of size 2.
- **Syntactic chunk** of the tokens, as given by the C&C chunker,[6] and the chunk code of the tokens in a window of size 2.
- **Concatenated stem, POS, and chunk labels** similar to the features used by Tang et al. (2010). These feature strings were a combination of the stem and the chunk code of the current token, the stem of the current token combined with the POS-codes of the token left and right, and the chunk code of the current token with the stems of the neighboring tokens.

5.3.2 CoNLL-2010 Experiments. The CoNLL-2010 shared task *Learning to detect hedges and their scope in natural language text* focused on uncertainty detection. Two subtasks were defined at the shared task: The first task sought to recognize sentences that contain some uncertain language in two different domains and the second task sought to recognize lexical cues together with their linguistic scope in biological texts (i.e., the text span in terms of constituency grammar that covers the part of the sentence that is modified by the cue). The lexical cue recognition subproblem of the second task[7] is identical to the problem setting used in this study, with the only major difference being the types of uncertainty addressed: In the CoNLL-2010 task biological texts contained only epistemic, doxastic, and investigation types of uncertainty. Apart from these differences, the CoNLL-2010 shared task offers an excellent testbed for comparing our uncertainty detection model with other state-of-the-art approaches for uncertainty detection and to compare different classification approaches. Here we present our detailed experiments using the CoNLL data sets, analyze the performance of our models, and select the most suitable models for further experiments.

CoNLL systems. The uncertainty detection systems that were submitted to the CoNLL shared task can be classified into three major types. The first set of systems treats the problem as a sentence classification task, that is, one to decide whether a sentence contains any uncertain element or not. These models operate at the sentence level and are unsuitable for cue detection. The second group handles the problem as a token

[6] POS-tagging and chunking were performed on all corpora using the C&C Tools (Curran, Clark, and Bos 2007).

[7] As an intermediate level, participants of the first task could submit the lexical cues found in sentences for evaluation, without their scope, which gave some insight into the nature of cue detection on the Wikipedia corpus (where scope annotation does not exist) as well.

Table 4
Results on the original CoNLL-2010 data sets. The first three rows correspond to our baseline, token-based, and sequence labeling models. The *BEST/SEQ* row shows the results of the best sequence labeling approach of the CoNLL shared task (for both domains), the *BEST/TOK* rows show the best token-based models, and the *BEST/SENT* rows show the best sentence-level classifiers (these models did not produce cue-level results).

	BIOLOGICAL		WIKIPEDIA	
	F_{cue}	F_{sent}	F_{cue}	F_{sent}
BASELINE	74.5	81.4	19.5	58.6
TOKEN/MAXENT	79.7	85.8	22.3	58.1
SEQUENCE/CRF	81.4	87.0	32.7	47.0
BEST/SEQ (Tang et al. 2010)	81.3	86.4	36.5	55.0
BEST/TOK BIO (Velldal, Øvrelid, and Oepen 2010)	78.7	85.2	–	–
BEST/TOK WIKI (Morante, Van Asch, and Daelemans 2010)	76.7	81.7	11.3	57.3
BEST/SENT BIO (Täckström et al. 2010)	–	85.2	–	55.4
BEST/SENT WIKI (Georgescul 2010)	–	78.5	–	60.2

classification task, and classifies each token independently as uncertain (or not). Contextual information is only included in the form of feature functions. The third group of systems handled the task as a sequential token labeling problem, that is, determined the most likely label sequence of a sentence in one step, taking the information about neighboring labels into account. Sequence labeling and token classification approaches performed best for biological texts and sentence-level models and token classification approaches gave the best results for Wikipedia texts (see Table 6 in Farkas et al. [2010]). Here we compare a state-of-the-art token classification and sequence labeling approach using a shared feature representation to decide which model to use in further experiments.

Classifier models. We used a first-order linear chain conditional random fields (CRF) model as a sequence labeler and a Maximum Entropy (Maxent) classifier model as a token classifier, implemented in the Mallet (McCallum 2002) package for training the uncertainty cue detectors. This choice was motivated by the fact that these were the most popular classification approaches among the CoNLL-2010 participants, and that CRF models are known to provide high accuracy for the detection of phrases with accurate boundaries (e.g., in named entity recognition). We trained the CRF and Maxent models with their default settings in Mallet for 200 iterations or until convergence (CRF), and also until convergence (Maxent) in each experimental set-up.

As a baseline model, we applied a simple dictionary-based approach which classifies every uni- and bigram as uncertain that is tagged as uncertain in over 50% of the cases in the training data. Hence, it is a similar system to that presented by Tjong Kim Sang (2010), without tuning the decision threshold for predicting uncertainty.

CoNLL results. An overview of the results achieved on the CoNLL-2010 data sets can be found in Table 4. A comparison of our models with the CoNLL systems reveals that our uncertainty detection model is very competitive when applied on the biological data set. Our CRF model trained on the official training data set of the shared task achieved a cue-level F-score of 81.4 and sentence-level F-score of 87.0 on the biological evaluation data set. These results would have come first in the shared task, with a marginal difference compared to the top performing participant. In contrast, our model is less competitive on the Wikipedia data set: The Maxent model achieved a cue-level F-score of 22.3 and sentence-level F-score of 58.1 on the Wikipedia evaluation data set, whereas our CRF model was not competitive with the best participating systems. The observation that sequence-labeling models perform worse than token-based approaches on Wikipedia, especially for sentence-level evaluation measures, coincides with the findings of the shared task: The discourse-level uncertainty cues in the Wikipedia data set are rather long and heterogeneous and sequence labeling models often revert to not annotating any token in a sentence when the phrase boundaries are hard to detect. Still, sequence labeling models have an advantage in terms of cue-level accuracy. This is not surprising because CRF is a state-of-the-art model for chunking / sequence labeling tasks.

We conclude from Table 4 that our model is competitive with the state-of-the-art systems for detecting semantic uncertainty (which is closer to the biological subtask), but it is less suited to recognizing discourse-level uncertainty. In the subsequent experiments we used our CRF model, which performed best in detecting uncertainty *cues* in natural language sentences.

5.3.3 Domain Adaptation Model. In supervised machine learning, the task is to learn how to make predictions on previously unseen, new examples based on a statistical model learned from a collection of labeled training examples (i.e., a set of examples coupled with the desired output for them). The classification setting assumes a set of labels L, a set of features X, and a probability distribution $p(X)$ describing the examples in terms of their features. Then the training examples are assumed to be given in the form of $\{x_i, l_i\}$ pairs and the goal of classification is to estimate the label distribution $p(L|X)$, which can be used later on to predict the labels for unseen examples.

Domain adaptation focuses on the problem where the same (or a closely related) learning task has to be solved in multiple domains which have different characteristics in terms of their features: The set of features X may be different or the probability distributions $p(X)$ describing the inputs may be different. When the target tasks are treated as different (but related), the label distribution $p(L|X)$ is dependent on the domain. That is, given a domain d, the problem can be formalized as modeling $p(L|X)_d$ based on X_d, $p(X)_d$ and a set of examples: $\{x_{i,d}, l_i\}$.[8] In the context of domain adaptation, there is a target domain t and a source domain s, with labeled data available for both, and the goal is to induce a more accurate target domain model $p(L|X)_t$ from $\{x_{i,t}, l_i\} \cup \{x_{i,s}, l_i\}$ than the one learned from $\{x_{i,t}, l_i\}$ only. In practical scenarios, the goal is to exploit the source data to acquire an accurate model from just limited target data which are alone insufficient to train an accurate in-domain model, and thus to port the model to a new domain with moderate annotation costs. The problem is difficult because it is nontrivial for a learning method to account for the different data (and label) distributions between target and source, which causes a remarkable drop in model accuracy when it is applied to classifying examples taken from the target domain.

In our experimental context, both topic- and genre-related differences of texts pose an adaptation problem as these factors have an impact on both the vocabulary ($p(X)$) and the sense distributions of the cues ($p(L|X)$) found in different texts. There is some

[8] The literature also describes the case when the set of labels depends on the domain, but we omit this case to simplify our notation and discussion. For details, see Pan and Yang (2010).

confusion in the literature regarding the terminology describing the various domain mismatches in the learning problem. For example, Daumé III (2007) describes a domain adaptation method where he assumes that the label distribution is unchanged (we note here that this assumption is not exploited in the method, and that the label distribution changes in our problem), whereas Pan and Yang (2010) uses the term *inductive transfer learning* to refer to our scenario (in their paper, *domain adaptation* refers to a different setting).[9] In this study we always use the term *domain adaptation* to refer to our problem setting, that is, where both $p(X)$ and $p(L|X)$ are assumed to change.

In our experiments, we used various data sets taken from multiple genres and domains (see Section 5.1.1 for an overview) and applied a simple but effective domain adaptation model (Daumé III 2007) for training our classifiers. In this model, domain adaptation is carried out by defining each feature over the target and source data sets twice—just once for target domain instances, and once for both the target and source domain instances. Formally, having a target domain t and a source domain s and n features $\{f_1, f_2, \ldots f_n\}$, for each f_i we have a target-only version $f_{i,t}$ and a shared version $f_{i,t+s}$. Each target domain example is described by $2n$ features: $\{f_{1,t}, f_{2,t}, \ldots f_{n,t}, f_{1,t+s}, f_{2,t+s}, \ldots f_{n,t+s}\}$ and source domain examples are described by only the n shared features: $\{f_{1,t+s}, f_{2,t+s}, \ldots f_{n,t+s}\}$. Using the union of the source and target training data sets $\{x_{i,t}, l_i\} \cup \{x_{i,s}, l_i\}$ and this feature representation, any standard supervised machine learning technique can be used and it becomes possible for the algorithm to learn target-dependent and shared patterns at the same time and handle the changes in the underlying distributions. This easy domain adaptation technique has been found to work well in many NLP-oriented tasks. We used the CRF models introduced herein and in this way, we were able to exploit feature–label correspondences across domains (for features that behave consistently across domains) and also to learn patterns specific to the target domain.

5.4 Cross-Domain and Genre Experiments

We defined several settings (target and source pairs) with varied domain and genre distances and target data set sizes. These experiments allowed us to study the potential of transferring knowledge across existing corpora for the accurate detection of uncertain language in a wide variety of text types. In our experiments, we used all the combinations of genres and domains that we found plausible. News texts (and their subdomains) were not used as source data because FactBank is significantly smaller than the other corpora (WikiWeasel or scientific texts). As the source data set is typically larger than the target data set in practical scenarios, news texts can only be used as target data. Abstracts were only used as source data because information extraction typically addresses full texts whereas abstracts just provide annotated data for development purposes. Besides these restrictions, we experimented with all possible target and source pairs.

We used four different machine-learning settings for each target–source pair in our investigations. In the purely cross-domain (CROSS) setting, the model was trained on the source domain and evaluated on the target (i.e., no labeled target domain data sets were used for training). In the purely in-domain setting (TARGET), we performed

[9] More on this can be found in Pan and Yang (2010) and at `http://nlpers.blogspot.com/2007/11/domain-adaptation-vs-transfer-learning.html`.

Table 5
Experimental results on different target and source domain pairs. The third column contains the ratio of the target train and source data sets' sizes in terms of sentences. DIST shows the distance of the source and target domain/genre ('-' same; '+' fine-grade difference; '++' coarse-grade difference; bio = biological; enc = encyclopedia; sci_paper = scientific paper; sci_abs = scientific abstract; sci_paper_hbc = scientific papers on human blood cell experiments; sci_paper_fly = scientific papers on Drosophila; sci_paper_bmc = scientific papers on bioinformatics).

				CROSS		TARGET		DA/ALL		DA/CUE	
TARGET	SOURCE	$\frac{SOURCE}{TARGET}$	DIST	F_{cue}	F_{sent}	F_{cue}	F_{sent}	F_{cue}	F_{sent}	F_{cue}	F_{sent}
enc	sci_paper+_abs	0.9	++/++	68.0	74.2	82.4	87.4	82.6	87.6	82.6	87.6
news	sci_paper+_abs	6.2	++/++	64.4	70.5	68.7	77.1	72.7	79.5	73.8	81.0
news	enc	6.6	++/++	68.2	74.8	68.7	77.1	73.7	81.2	73.1	80.0
sci_paper	enc	2.7	++/++	67.8	75.1	78.8	84.4	80.0	85.9	79.8	85.4
sci_paper_bmc	sci_abs_hbc	4.3	+/+	58.2	70.5	64.0	74.5	68.1	76.7	69.3	77.8
sci_paper_fly	sci_abs_hbc	3.4	+/+	70.5	79.1	80.0	85.1	83.3	88.2	82.9	87.8
sci_paper_hbc	sci_abs_hbc	8.2	−/+	76.5	82.9	74.2	80.2	84.2	88.6	83.0	88.9
sci_paper_bmc	sci_paper_fly+_hbc	1.8	+/−	69.8	77.6	64.0	74.5	70.0	78.2	69.4	78.1
sci_paper_fly	sci_paper_bmc+_hbc	1.2	+/−	78.4	83.5	80.0	85.1	82.6	87.0	82.9	87.0
sci_paper_hbc	sci_paper_bmc+_fly	4.4	+/−	81.7	85.9	74.2	80.2	80.7	86.9	80.7	85.9
		AVERAGE:		70.4	77.4	73.5	80.6	77.8	84.0	77.8	84.0

10-fold cross-validation on the target data (i.e., no source domain data were used). In the two domain adaptation settings, we again performed 10-fold cross-validation on the target data but exploited the source data set (as described in Section 5.3). Here, we either used each sentence of the source data set (DA/ALL) or only those sentences that contained a cue observed in the target train data set (DA/CUE).

Table 5 lists the results obtained on various target and source domains in various machine learning settings and Table 6 contains the absolute differences between a particular result and the in-domain (TARGET) results.

Fine-grained semantic uncertainty classification results are summarized in Tables 7 and 8. Table 7 contrasts the coarse-grained F_{cue} with the unlabeled/binary F_{cue} of fine-grained experiments, therefore it quantifies the difference in accuracy due to the more difficult classification setting and the increased sparseness of the task. Table 8 shows the per class F_{cue} scores, namely, how accurately our model recognizes the individual uncertainty types.

Table 6
The absolute difference between the F-scores of Table 5 relative to the baseline TARGET setting.

				CROSS		TARGET		DA/ALL		DA/CUE	
TARGET	SOURCE	$\frac{SOURCE}{TARGET}$	DIST	F_{cue}	F_{sent}	F_{cue}	F_{sent}	F_{cue}	F_{sent}	F_{cue}	F_{sent}
enc	sci_paper+_abs	0.9	++/++	–14.4	–13.2	82.4	87.4	0.2	0.2	0.2	0.2
news	sci_paper+_abs	6.2	++/++	–4.3	–6.6	68.7	77.1	4.0	2.4	5.1	3.9
news	enc	6.6	++/++	–0.5	–2.3	68.7	77.1	5.0	4.1	4.4	2.9
sci_paper	enc	2.7	++/++	–11.0	–9.3	78.8	84.4	1.2	1.5	1.0	1.0
sci_paper_bmc	sci_abs_hbc	4.3	+/+	–5.8	–4.0	64.0	74.5	4.1	2.2	5.3	3.3
sci_paper_fly	sci_abs_hbc	3.4	+/+	–9.5	–6.0	80.0	85.1	3.3	3.1	2.9	2.7
sci_paper_hbc	sci_abs_hbc	8.2	−/+	2.3	2.7	74.2	80.2	10.0	8.4	8.8	8.7
sci_paper_bmc	sci_paper_fly+_hbc	1.8	+/−	5.8	3.1	64.0	74.5	6.0	3.7	5.4	3.6
sci_paper_fly	sci_paper_bmc+_hbc	1.2	+/−	–1.6	–1.6	80.0	85.1	2.6	1.9	2.9	1.9
sci_paper_hbc	sci_paper_bmc+_fly	4.4	+/−	7.5	5.7	74.2	80.2	6.5	6.7	6.5	5.7
		AVERAGE:		–3.1	–3.2	73.5	80.6	4.3	3.4	4.3	3.4

Table 7
Comparison of cue-level binary (F_{bin}) and unlabeled F-scores (F_{unl}). Binary F-score corresponds to coarse-grained classification (uncertain vs. certain), and unlabeled F-score is the fine-grained classification converted to binary (disregarding the fine-grained category labels).

				CROSS		TARGET		DA/ALL		DA/CUE	
TARGET	SOURCE	$\frac{SOURCE}{TARGET}$	DIST	F_{bin}	F_{unl}	F_{bin}	F_{unl}	F_{bin}	F_{unl}	F_{bin}	F_{unl}
enc	sci_paper+_abs	0.9	++/++	68.0	67.4	82.4	82.4	82.6	81.9	82.6	81.7
news	sci_paper+_abs	6.2	++/++	64.4	59.9	68.7	66.4	72.7	71.5	73.8	71.8
news	enc	6.6	++/++	68.2	67.0	68.7	66.4	73.7	73.6	73.1	73.4
sci_paper	enc	2.7	++/++	67.8	67.2	78.8	78.3	80.0	80.2	79.8	79.5
sci_paper_bmc	sci_abs_hbc	4.3	+/+	58.2	66.3	64.0	61.9	68.1	68.5	69.3	67.9
sci_paper_fly	sci_abs_hbc	3.4	+/+	70.5	78.7	80.0	79.2	83.3	83.4	82.9	83.2
sci_paper_hbc	sci_abs_hbc	8.2	−/+	76.5	83.6	74.2	69.3	84.2	83.1	83.0	83.4
sci_paper_bmc	sci_paper_fly+_hbc	1.8	+/−	69.8	69.7	64.0	61.9	70.0	69.5	69.4	65.9
sci_paper_fly	sci_paper_bmc+_hbc	1.2	+/−	78.4	77.7	80.0	79.2	82.6	82.1	82.9	82.5
sci_paper_hbc	sci_paper_bmc+_fly	4.4	+/−	81.7	81.9	74.2	69.3	80.7	81.3	80.7	81.2
		AVERAGE:		70.4	71.9	73.5	71.4	77.8	77.5	77.8	77.0

Table 8
The per class cue-level F-scores in fine-grained classification. F_{crs}, F_{tgt}, and F_{da} correspond to the CROSS, TARGET, and DA/CUE settings, respectively (same as previous). The DA/ALL setting is not shown for space reasons and due to its similarity to the DA/CUE results.

		EPISTEMIC			INVESTIGATION			DOXASTIC			CONDITION		
TARGET	SOURCE	F_{crs}	F_{tgt}	F_{da}	F_{crs}	F_{tgt}	F_{da}	F_{crs}	F_{tgt}	F_{da}	F_{crs}	F_{tgt}	F_{da}
enc	sci_paper+_abs	75.9	83.4	82.8	67.3	67.5	70.4	48.8	89.2	88.1	54.4	62.6	61.2
news	sci_paper+_abs	70.9	65.4	75.2	79.5	75.9	83.1	39.1	68.9	71.3	47.2	57.1	57.5
news	enc	65.4	65.4	74.5	74.6	75.9	87.5	76.3	68.9	78.0	50.6	57.1	56.7
sci_paper	enc	72.9	81.2	81.9	36.5	72.9	72.4	63.6	74.9	79.8	57.0	58.9	59.7
sci_paper_bmc	sci_abs_hbc	71.5	68.3	72.6	56.1	37.7	58.1	68.1	61.9	69.4	45.5	45.0	49.5
sci_paper_fly	sci_abs_hbc	82.9	82.1	85.3	69.0	68.6	76.6	75.1	71.7	75.4	28.6	63.4	64.1
sci_paper_hbc	sci_abs_hbc	87.5	77.7	86.4	76.5	53.5	77.5	80.6	39.0	76.7	26.1	10.0	33.3
sci_paper_bmc	sci_paper_fly+_hbc	74.4	68.3	69.2	55.9	37.7	57.4	63.7	61.9	64.7	57.3	45.0	50.7
sci_paper_fly	sci_paper_bmc+_hbc	80.3	82.1	84.3	66.7	68.6	75.8	77.7	71.7	77.3	53.5	63.4	68.0
sci_paper_hbc	sci_paper_bmc+_fly	85.2	77.7	86.0	74.0	53.5	70.3	75.9	39.0	70.2	58.1	10.0	41.4
	AVERAGE:	76.7	75.2	79.8	65.6	61.2	72.9	66.9	64.7	75.1	47.8	47.3	54.2

The size of the target training data sets proved to be an important factor in these investigations. Hence, we performed experiments with different target data set sizes. We utilized the DA/ALL model (which is more robust for extremely small target data sizes [e.g., 100-400 sentences]) and performed the same 10-fold cross validation on the target data set as in Tables 5-8. For each fold of the cross-validation here, however, we just used n sentences (x axis of the figures) from the target training data set and a fixed set of 4,000 source sentences to alleviate the effect of varying data set sizes. Figure 5 depicts the learning curves for two target/source data set pairs.

6. Discussion

As Table 5 shows, incorporating labeled data from different genres and/or domains consistently improves the performance. The successful applicability of domain adaptation tells us that the problem of detecting uncertainty has similar characteristics across genres and domains. The uncertainty cue lexicons of different domains and

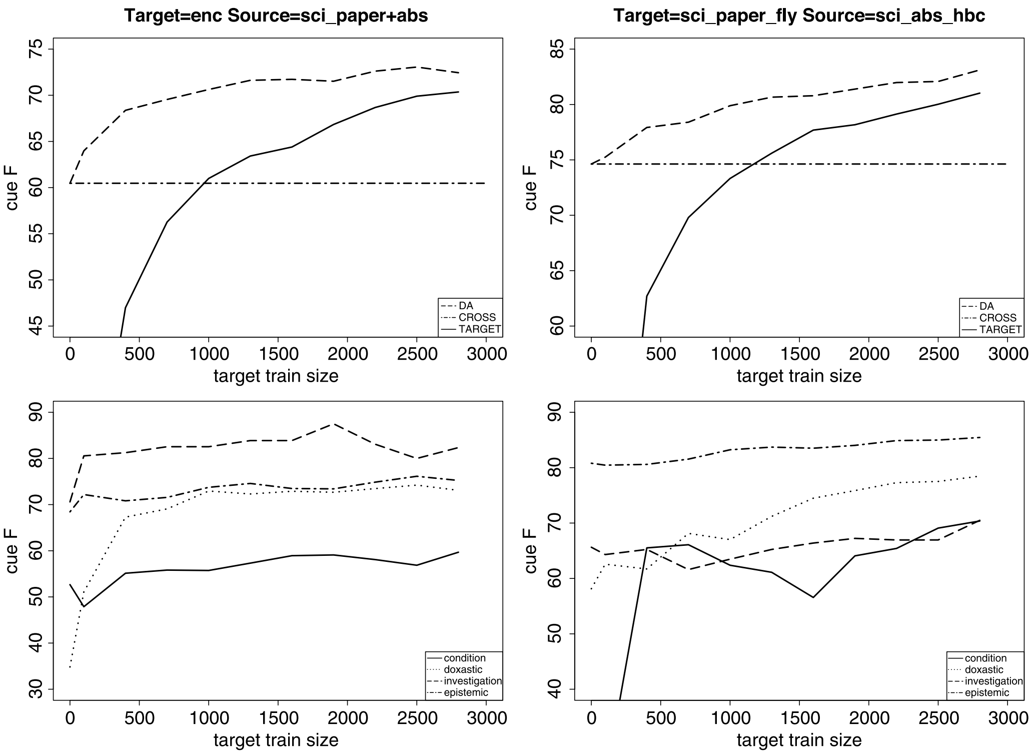

Figure 5
Learning curves: Results achieved with different target train sizes. The left and right figures show two selected source/target pairs. The upper figures depict coarse-grained classification results (F_{cue}); DA, CROSS, and TARGET with the same settings as in Table 5. The lower figures show the per class F_{cue} of the DA/ALL model in the fine-grained classification.

genres indeed share a core vocabulary and despite the differences in sense distributions, labeled data from a different source improves uncertainty classification in a new genre and domain if the different data sets are annotated consistently. This justifies our aim to create a consistent representation of uncertainty that can be applied to multiple domains.

6.1 Domain Adaptation Results

The size of the target and source data sets largely influences to what extent external data can improve results. The only case where domain adaptation had only a negligible effect (an F-score gain less than 1%) is where the target data set is itself very large. This is expected as the more target data one has, the less crucial it is to incorporate additional data with some undesirable characteristics (difference in style, domain, certain/uncertain sense distribution, etc.).

The performance scores for the CROSS setting clearly indicate the domain/genre distance of the data sets: The more distant the domain and genre of the source and target data sets are, the more the CROSS performance (where no labeled target data is used) degrades, compared with the TARGET model. In general, when the distance between both the domain and the genre of texts is substantial (++/++ and +/+ rows in Tables 5 and 6), this accounts for a 6–10% decrease in both the sentence and cue-level F-scores. An exception is the case of encyclopedic source and news target domains. Here the performance is very close to the target domain performance. This indicates that these settings are not so different from each other as it might seem at the first glance. The encyclopedic and news genres share quite a lot of commonalities (compare cue distributions in Figure 4, for instance). We verified this observation by using a knowledge-poor quantitative estimator of similarity between domains (Van Asch and Daelemans 2010): Using cosine as the similarity measure, the newswire and encyclopedia texts are found to be the second most similar domain pair in our experiments, with a score comparable to those obtained for the pairs of scientific article types `bmc`, `hbc`, and `fly`.

When there is a domain or genre match between source and target (−/+ and +/− rows in Tables 5 and 6), however, and the distance regarding the other is just moderate, the cross-training performance is close to or even better than the target-only results. That is, the larger amount of source training data balances the differences between the domains. These results indicate that the learned uncertainty classifiers can be directly applied to *slightly* different data sets. This suitability is due to the learned disambiguation models, which generalize well in similar settings. This is contrary to the findings of earlier studies, which built the uncertainty detectors using seed examples and bootstrapping. These models were not designed to learn any disambiguation models for the cue words found, and their performance degraded even for slightly different data (Szarvas 2008).

Comparing the two domain adaptation procedures DA/CUE and DA/ALL, adaptation via transferring only source sentences that contain a target domain cue is, on average, comparable to transferring all the data from the source domain. In other words, when we have a small but sufficient amount of target data available, it is enough to account for source data corresponding to the uncertainty cues we saw in the limited target data set. This observation has several consequences, namely:

- The source-only cues, or to be more precise, their disambiguation models, are not helpful for the target domains as they cannot be adapted. This is due to the differences in the source and target disambiguation models.
- Similarly, domain adaptation improves the disambiguation models for the observed target cues, rather than introducing new vocabulary into the target domain. This mechanism coincides with our initial goal of using domain adaptation to learn better semantic models. This effect is the opposite of how bootstrapping-based weakly supervised approaches improve the performance in an underresourced domain. This observation suggests a promising future direction of combining the two approaches to maximize the gains while minimizing the annotation costs.
- In a general context, we can effectively extend the data for a given domain if we have robust knowledge of the potential uncertainty vocabulary for that domain. Given the wide variety of the domains and genres of our data sets, it is reasonable to suppose that they represent uncertain language in general quite well, and the joint vocabularies provide a good starting point for a targeted data development for further domains.

As regards the fine-grained classification results, Table 7 demonstrates that the fine-grained distinction results in only a small, or no, loss in performance. The coarse-grained model is slightly more accurate than the fine-grained model (counting correctly recognized but misclassified cues as true positives) in most settings. The most significant difference is observed for the target-only settings, where no out-of-domain data are used for the training and thus the data sets are accordingly smaller. A noticeable exception is when scientific abstracts are used for cross training: In those settings the coarse-grained model performs poorly, due to its lower recall, which we attribute to overfitting the special characteristics of abstracts. The fact that in fine-grained classification the CROSS results consistently outperform the TARGET models (see Table 8) even for distant domain pairs, also underlines that the increased sparseness caused by the differentiation of the various subtypes of uncertainty is an important factor only for smaller data sets. The improvement by domain adaptation is clearly more prominent in fine-grained than in coarse-grained classification, however: The individual cue types benefit by 5–10% points in terms of the F-score from out-of-domain data and domain adaptation. Moreover, as Table 8 shows, for the domain pairs and fine-grained classes where a nice amount of positive examples are at hand, the per class F_{cue} scores are also around 80% and above. This means that it is possible to accurately identify the individual subtypes of semantic uncertainty, and thus it also proves the feasibility of the subcategorization and annotation scheme proposed in this study (Section 2). Other important observations here are that domain adaptation is even more significant in the more difficult fine-grained classification setting, and that the condition class represents a challenge for our model. The performance for the condition class is lower than that for the other classes, which can only in part be attributed to the fact that this is the least represented subtype in our data sets: as opposed to other cue types, condition cues are typically used in many different contexts and they may belong to other uncertainty classes as well.

6.2 The Required Amount of Annotation

Based on our experiments, we may conclude that a manually annotated training data set consisting of 3,000–5,000 sentences is sufficient for training an accurate cue detector for a new genre/domain. The results of our learning curve experiments (Figure 5) illustrate the situations where only a limited amount of annotated data (fewer than 3,000 sentences) is available for the target domain. The feasibility of decreasing annotation efforts and the real added value of domain adaptation are more prominent in this range. It is easy to see that the TARGET results approach to DA results with more target data.

Figure 5 shows that the size of the target training data set where the supervised TARGET setting outperforms the CROSS model (trained on 4,000 source sentences) is around 1,000 sentences. As we mentioned earlier, even distant domain data can improve the cue recognition model in the absence of a sufficient target data set. Figure 5 justifies this observation, as the CROSS and DA settings outperform the TARGET setting on each source–target data set pair. It can also be observed that the doxastic type is more domain-dependent than the others and its results consistently improve by increasing the size of the target domain annotation (which coincides with the cue frequency investigations of Section 5.1.3). In the news target domain, however, the investigation and epistemic classes benefit a lot from a small amount of annotated target data but their performance scores increase just slightly after that. This indicates that most of the important domain-dependent (probably lexical) knowledge could be gathered from 100–400 sentences. In the biological experiments, we may conclude that the investigation class is already covered by the source domain (intuitively, the investigation cues are well represented in the abstracts) and its results are not improved significantly by using more target data. The condition class is underrepresented in both the source and target data sets and hence no reliable observations can be made regarding this subclass (see Table 2).

Overall, if we would like to have an uncertainty cue detector for a new genre/domain: (i) We can achieve performance around 60–70% by using cross training depending on the difference between the domains (i.e., without any annotation effort); (ii) By annotating around 3,000 sentences, we can have a performance of 70–80%, depending on the level of difficulty of the texts; (iii) We can get the same 70–80% results with annotating just 1,000 sentences and using domain adaptation.

6.3 Interesting Examples and Error Analysis

As might be expected, most of the erroneous cue predictions were due to vocabulary differences, for example, *fear* or *accuse* occurred only in news texts, which is why they were not recognized by models trained on biological or encyclopedia texts. Another example is the case of *or*, which is a frequent cue in biological texts. Still, it is rarely used as a cue in other domains but without domain adaptation, the model trained on biological texts marks quite a few occurrences of *or* as cues in the news or encyclopedia domains. Many of these anomalies were eliminated by the application of domain adaptation techniques, however.

Many errors were related to multi-class cues. These cues are especially hard to disambiguate because not only can they refer to several classes of uncertainty, but they typically have non-cue usage as well. For instance, the case of *would* is rather complicated because it can fulfill several functions:

(10) EPISTEMIC USAGE ('IT IS HIGHLY PROBABLE'): Further biochemical studies on the mechanism of action of purified kinesin-5 from multiple systems **would** obviously be fruitful. (Corpus: `fly`)

(11) CONDITIONAL: "If religion was a thing that money could buy,/The rich **would** live and the poor **would** die." (Corpus: WikiWeasel)

(12) FUTURE IN THE PAST: This Aarup can trace its history back to 1500, but it **would** be 1860's before it **would** become a town. (Corpus: WikiWeasel)

(13) REPEATED ACTION IN THE PAST ('USED TO'): 'Becker' was the next T.V. Series for Paramount that Farrell **would** co-star in. (Corpus: WikiWeasel)

(14) DYNAMIC MODALITY: Individuals **would** first have a small lesion at the site of the insect bite, which **would** eventually leave a small scar. (Corpus: WikiWeasel)

(15) PRAGMATIC USAGE: Although some **would** dispute the fact, the joke related to a peculiar smell that follows his person. (Corpus: WikiWeasel)

The epistemic uses of *would* are annotated as epistemic cues whereas its occurrences in conditionals are marked as hypothetical cues. The habitual past meaning is not related to uncertainty, hence it is not annotated. The future in the past meaning (i.e., past tense of *will*), however, denotes an event of which it is known that happened later, so it is certain. The dynamically modal *would* is similar to the future *will* (which is an instance of dynamic modality as well), but it is not annotated in the corpora. The pragmatic use of *would* does not refer to semantic uncertainty (the semantic value of the sentence would be exactly the same without it or if it is replaced with *may*, *might*, *will*, etc., that is, *some will/may/might/∅ dispute the fact* mean the same). It is rather a stylistic issue to further express uncertainty at the discourse level (i.e., there are some unidentified people who dispute the fact, hence the opinion cannot be associated with any definite source).

The last two uses of *would* are not typically described in grammars of English and seem to be characteristic primarily of the news and encyclopedia domains. Thus it is advisable to explore such cases and treat them with special consideration when adapting an algorithm trained and tested in a specific domain to another domain.

Another interesting example is *may* in its non-cue usage. Being (one of) the most frequent cues in each subcorpus, its non-cue usage is rather limited but can be found occasionally in FactBank and WikiWeasel. The following instance of *may* in FactBank was correctly marked as non-cue by the cue detector when trained on Wikipedia texts. On the other hand, it was marked as a cue when trained on biological texts since in this case, there were insufficient training examples of *may* not being a cue:

(16) "Well **may** we say 'God save the Queen,' for nothing will save the republic," outraged monarchist delegate David Mitchell said. (Corpus: FactBank)

A final example to be discussed is *concern*. This word also has several uses:

(17) NOUN MEANING 'COMPANY': The insurance **concern** said all conversion rights on the stock will terminate on Nov. 30. (Corpus: FactBank)

(18) NOUN MEANING 'WORRY': **Concern** about declines in other markets, especially New York, caused selling pressure. (Corpus: FactBank)

(19) PREPOSITION: The company also said it continues to explore all options **concerning** the possible sale of National Aluminum's 54.5% stake in an aluminum smelter in Hawesville, Ky. (Corpus: FactBank)

(20) VERB: Many of the predictions in these two data sets **concern** protein pairs and proteins that are not present in other data sets. (Corpus: `bmc`)

Among these examples, only the second one should be annotated as uncertain. POS-tagging seems to provide enough information for excluding the verbal and prepositional uses of the word but in the case of nominal usage, additional information is also required to enable the system to decide whether it is an uncertainty cue or not (in this case, the noun in the 'company' sense cannot have an argument while in the 'worry' sense, it can have [*about declines*]). Again, the frequency of the two senses depends heavily on the domain of the texts, which should also be considered when adapting the cue detector to a different domain. We should mention that the role of POS-tagging is essential in cue detection because many ambiguities can be resolved on the basis of POS-tags. Hence, POS-tagging errors can lead to a serious decline in performance.

We think that an analysis of similar examples can further support domain adaptation and cue detection across genres and domains.

7. Conclusions and Future Work

In this article, we introduced an uncertainty cue detection model that can perform well across different domains and genres. Even though several types of uncertainty exist, available corpora and resources focus only on some of the possible types and thereby only cover particular aspects of the phenomenon. This means that uncertainty models found in the literature are heterogeneous, and the results of experiments on different corpora are hardly comparable. These facts motivated us to offer a unified model of semantic uncertainty enhanced by linguistic and computer science considerations. In accordance with this classification, we reannotated three corpora from several domains and genres using our uniform annotation guidelines.

Our results suggest that simple cross training can be employed and it achieves a reasonable performance (60–70% cue-level F-score) when no annotated data is at hand for a new domain. When some annotated data is available (here *some* means fewer than 3,000 annotated sentences for the target domain), domain adaptation techniques are the best choice: (i) they lead to a significant improvement compared to simple cross training, and (ii) they can provide a reasonable performance with significantly less annotation. In our experiments, the annotation of 3,000 sentences and training only on these is roughly equivalent to the annotation of 1,000 sentences using external data and domain adaptation. If the size of the training data set is sufficiently large (larger than 5,000 sentences) the effect of incorporating additional data—having some undesirable characteristics—is not crucial.

Comparing different domain adaptation techniques, we found that similar results could be attained when the source domain was filtered for sentences that contained cues in the target domain. This tells us that models learn to better disambiguate the cues seen in the target domain instead of finding new, unseen cues. In this sense, this approach can be regarded as a complementary method to weakly supervised techniques for lexicon extraction. A promising way to further minimize annotation costs while maximizing performance would be the integration of the two approaches, which we plan to investigate in the near future.

In our study, we did not pay attention to dynamic modality (due to the lack of annotated resources), but the detection of such phenomena is also desirable. For instance, dynamically modal events cannot be treated as certain—that is, the event of buying cannot be assigned the same truth value in *They agreed to buy the company* and *They bought the company*. Whereas the second sentence expresses a fact, the first one informs us about the intention of buying the company, which will be certainly carried out in a world where moral or business laws are observed but at the moment it cannot be stated whether the transaction takes place (i.e., that it is certain). Hence, in the future, we also intend to integrate the identification of dynamically modal cues into our uncertainty cue detector.

Acknowledgments

This work was supported in part by the NIH grants (project codenames MASZEKER and BELAMI) of the Hungarian government, by the German Ministry of Education and Research under grant SiDiM (grant no. 01IS10054G), and by the Volkswagen Foundation as part of the Lichtenberg-Professorship Program (grant no. I/82806). Richárd Farkas was funded by the Deutsche Forschungsgemeinschaft grant SFB 732.

References

Baker, Kathy, Michael Bloodgood, Mona Diab, Bonnie Dorr, Ed Hovy, Lori Levin, Marjorie McShane, Teruko Mitamura, Sergei Nirenburg, Christine Piatko, Owen Rambow, and Gramm Richardson. 2010. Modality Annotation Guidelines. Technical Report 4, Human Language Technology Center of Excellence, Baltimore, MD.

Chapman, Wendy W., David Chu, and John N. Dowling. 2007. ConText: An algorithm for identifying contextual features from clinical text. In *Proceedings of the ACL Workshop on BioNLP 2007*, pages 81–88, Prague, Czech Republic.

Clausen, David. 2010. HedgeHunter: A system for hedge detection and uncertainty classification. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning (CoNLL-2010): Shared Task*, pages 120–125, Uppsala.

Conway, Mike, Son Doan, and Nigel Collier. 2009. Using hedges to enhance a disease outbreak report text mining system. In *Proceedings of the BioNLP 2009 Workshop*, pages 142–143, Boulder, CO.

Curran, James, Stephen Clark, and Johan Bos. 2007. Linguistically motivated large-scale NLP with C&C and Boxer. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 33–36, Prague.

Daumé III, Hal. 2007. Frustratingly easy domain adaptation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 256–263, Prague.

Daumé III, Hal and Daniel Marcu. 2006. Domain adaptation for statistical classifiers. *Journal of Artificial Intelligence Research*, 26:101–126.

Falahati, Reza. 2006. The use of hedging across different disciplines and rhetorical sections of research articles. In *Proceedings of the 22nd NorthWest Linguistics Conference (NWLC22)*, pages 99–112, Burnaby.

Farkas, Richárd and György Szarvas. 2008. Automatic construction of rule-based ICD-9-CM coding systems. *BMC Bioinformatics*, 9:1–9.

Farkas, Richárd, 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 (CoNLL-2010): Shared Task*, pages 1–12, Uppsala.

Fernandes, Eraldo R., Carlos E. M. Crestana, and Ruy L. Milidiú. 2010. Hedge detection using the RelHunter approach. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning (CoNLL-2010): Shared Task*, pages 64–69, Uppsala.

Friedman, Carol, Philip O. Alderson, John H. M. Austin, James J. Cimino, and Stephen B. Johnson. 1994. A General natural-language text processor for clinical radiology. *Journal of the American Medical Informatics Association*, 1(2):161–174.

Ganter, Viola 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, Suntec.

Georgescul, Maria. 2010. A hedgehop over a max-margin framework using hedge cues. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning (CoNLL-2010): Shared Task*, pages 26–31, Uppsala.

Hyland, Ken. 1994. Hedging in academic writing and EAP textbooks. *English for Specific Purposes*, 13(3):239–256.

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

Hyland, Ken. 1998. Boosters, hedging and the negotiation of academic knowledge. *Text*, 18(3):349–382.

Kiefer, Ferenc. 2005. *Lehetőség és szükségszerűség* [Possibility and necessity]. Tinta Kiadó, Budapest.

Kilicoglu, Halil and Sabine Bergler. 2008. Recognizing speculative language in biomedical research articles: A linguistically motivated perspective. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*, pages 46–53, Columbus, OH.

Kilicoglu, Halil and Sabine Bergler. 2009. Syntactic dependency based heuristics for biological event extraction. In *Proceedings of the BioNLP 2009 Workshop Companion Volume for Shared Task*, pages 119–127, Boulder, CO.

Kim, Jin-Dong, Tomoko Ohta, Sampo Pyysalo, Yoshinobu Kano, and Jun'ichi Tsujii. 2009. Overview of BioNLP'09 Shared Task on Event Extraction. In *Proceedings of the BioNLP 2009 Workshop Companion Volume for Shared Task*, pages 1–9, Boulder, OH.

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

Li, Xinxin, Jianping Shen, Xiang Gao, and Xuan Wang. 2010. Exploiting rich features for detecting hedges and their scope. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning (CoNLL-2010): Shared Task*, pages 78–83, Uppsala.

Light, Marc, Xin Ying Qiu, and Padmini Srinivasan. 2004. The language of bioscience: Facts, speculations, and statements in between. In *Proceedings of the HLT-NAACL 2004 Workshop: Biolink 2004, Linking Biological Literature, Ontologies and Databases*, pages 17–24, Boston, Massachusetts, USA.

MacKinlay, Andrew, David Martinez, and Timothy Baldwin. 2009. Biomedical event annotation with CRFs and precision grammars. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing: Shared Task*, BioNLP '09, pages 77–85, Uppsala.

McCallum, Andrew Kachites. 2002. *MALLET: A Machine Learning for Language Toolkit*. Available at `http://mallet.cs.umass.edu`.

Medlock, Ben and Ted Briscoe. 2007. Weakly supervised learning for hedge classification in scientific literature. In *Proceedings of the ACL*, pages 992–999, Prague.

Morante, Roser and Walter Daelemans. 2009. Learning the scope of hedge cues in biomedical texts. In *Proceedings of the BioNLP 2009 Workshop*, pages 28–36, Boulder, CO.

Morante, Roser and Walter Daelemans. 2011. Annotating modality and negation for a machine reading evaluation. In *Proceedings of CLEF 2011*, Amsterdam, Netherlands.

Morante, Roser, Vincent Van Asch, and Walter Daelemans. 2010. Memory-based resolution of in-sentence scopes of hedge cues. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning (CoNLL-2010): Shared Task*, pages 40–47, Uppsala, Sweden.

Nawaz, Raheel, Paul Thompson, and Sophia Ananiadou. 2010. Evaluating a meta-knowledge annotation scheme for bio-events. In *Proceedings of the Workshop on Negation and Speculation in Natural Language Processing*, pages 69–77, Uppsala.

Özgür, Arzucan and Dragomir R. Radev. 2009. Detecting speculations and their scopes in scientific text. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 1398–1407, Singapore.

Palmer, Frank Robert. 1979. *Modality and the English Modals*. Longman, London.

Palmer, Frank Robert. 1986. *Mood and Modality*. Cambridge University Press, Cambridge.

Pan, Sinno Jialin and Qiang Yang. 2010. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359.

Rei, Marek and Ted Briscoe. 2010. Combining manual rules and supervised learning for hedge cue and scope detection. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning (CoNLL-2010): Shared Task*, pages 56–63, Uppsala.

Rizomilioti, Vassiliki. 2006. Exploring epistemic modality in academic discourse using corpora. In Elisabet Arnó Macia, Antonia Soler Cervera, and Carmen Rueda Ramos, editors, *Information Technology in Languages for Specific Purposes*, volume 7 of *Educational Linguistics*. Springer US, New York, pages 53–71.

Rubin, Victoria 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.

Rubin, Victoria L., Elizabeth D. Liddy, and Noriko Kando. 2005. Certainty identification in texts: Categorization model and manual tagging results. In James G. Shanahan, Yan Qu, and Janyce Wiebe, editors, *Computing Attitude and Affect in Text: Theory and Applications (the Information Retrieval Series)*, Springer Verlag, New York, pages 61–76.

Russell, Stuart J. and Peter Norvig. 2010. *Artificial Intelligence—A Modern Approach (3rd international edition)*. Upper Saddle River, NJ: Pearson Education.

Sánchez, Liliana Mamani, Baoli Li, and Carl Vogel. 2010. Exploiting CCG structures with tree kernels for speculation detection. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning (CoNLL-2010): Shared Task*, pages 126–131, Uppsala.

Saurí, Roser. 2008. *A Factuality Profiler for Eventualities in Text*. Ph.D. thesis, Brandeis University, Waltham, MA.

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

Settles, Burr, Mark Craven, and Lewis Friedland. 2008. Active learning with real annotation costs. In *Proceedings of* *the NIPS Workshop on Cost-Sensitive Learning*, pages 1–10, Vancouver, Canada.

Shatkay, Hagit, Fengxia Pan, Andrey Rzhetsky, and W. John Wilbur. 2008. Multi-dimensional classification of biomedical text: Toward automated, practical provision of high-utility text to diverse users. *Bioinformatics*, 24(18):2086–2093.

Sun, Chengjie, Lei Lin, Xiaolong Wang, and Yi Guan. 2007. Using maximum entropy model to extract protein-protein interaction information from biomedical literature. In De-Shuang Huang, Donald C. Wunsch, Daniel S. Levine, and Kang-Hyun Jo, editors, *Advanced Intelligent Computing Theories and Applications. With Aspects of Theoretical and Methodological Issues*. Springer Verlag, Heidelberg, pages 730–737.

Szarvas, György. 2008. Hedge classification in biomedical texts with a weakly supervised selection of keywords. In *Proceedings of ACL-08: HLT*, pages 281–289, Columbus, OH.

Täckström, Oscar, Sumithra Velupillai, Martin Hassel, Gunnar Eriksson, Hercules Dalianis, and Jussi Karlgren. 2010. Uncertainty detection as approximate max-margin sequence labelling. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning (CoNLL-2010): Shared Task*, pages 84–91, Uppsala.

Tang, Buzhou, 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 (CoNLL-2010): Shared Task*, pages 13–17, Uppsala.

Thompson, Paul, Giulia Venturi, John McNaught, Simonetta Montemagni, and Sophia Ananiadou. 2008. Categorising modality in biomedical texts. In *Proceedings of the LREC 2008 Workshop on Building and Evaluating Resources for Biomedical Text Mining*, pages 27–34, Marrakech, Morocco.

Tjong Kim Sang, Erik. 2010. A baseline approach for detecting sentences containing uncertainty. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning (CoNLL-2010): Shared Task*, pages 148–150, Uppsala.

Uzuner, Özlem, 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.

Van Asch, Vincent and Walter Daelemans. 2010. Using domain similarity for performance estimation. In *Proceedings of the 2010 Workshop on Domain Adaptation for Natural Language Processing*, pages 31–36, Uppsala.

Van Landeghem, Sofie, Yvan Saeys, Bernard De Baets, and Yves Van de Peer. 2009. Analyzing text in search of bio-molecular events: A high-precision machine learning framework. In *Proceedings of the BioNLP 2009 Workshop Companion Volume for Shared Task*, pages 128–136, Boulder, CO.

Velldal, Erik. 2010. Detecting uncertainty in biomedical literature: A simple disambiguation approach using sparse random indexing. In *Proceedings of SMBM 2010*, pages 75–83, Cambridge.

Velldal, Erik, Lilja Øvrelid, and Stephan Oepen. 2010. Resolving speculation: MaxEnt cue classification and dependency-based scope rules. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning (CoNLL-2010): Shared Task*, pages 48–55, Uppsala.

Vincze, Veronika, 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.

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

Zhang, Shaodian, Hai Zhao, Guodong Zhou, and Bao-Liang Lu. 2010. Hedge detection and scope finding by sequence labeling with normalized feature selection. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning (CoNLL-2010): Shared Task*, pages 92–99, Uppsala.