Uncertainty Detection
in Natural Language Texts

Veronika Vincze

Research Group on Artificial Intelligence
and
University of Szeged

July 2014

A THESIS SUBMITTED FOR THE DEGREE OF DOCTOR OF PHILOSOPHY
OF THE UNIVERSITY OF SZEGED

Supervisor: János Csirik, DSc

University of Szeged
Doctoral School in Computer Science

Contents

List of Tables

List of Figures

Preface

Uncertainty is an important linguistic phenomenon that is relevant in many fields of language processing. In its most general sense, it can be interpreted as lack of information: the hearer or the reader cannot be certain about some pieces of information. Thus, uncertain propositions are those whose truth value or reliability cannot be determined due to lack of information. Distinguishing between factual (i.e. true or false) and uncertain propositions is of primary importance both in linguistics and natural language processing applications. For instance, in information extraction an uncertain piece of information might be of some interest for an end-user as well, but such information must not be confused with factual textual evidence (reliable information) and the two should be kept separated.

The main objective of this thesis is to detect uncertainty in English and Hungarian natural language texts. As opposed to earlier studies that focused on specific domains and were English-oriented, we will offer here a comprehensive approach to uncertainty detection, which can be easily adapted to the specific needs of many domains and languages. In our investigations, we will pay attention to create linguistically plausible models of uncertainty that will be exploited in creating manually annotated corpora that will serve as the base for the implementation of our uncertainty detectors for several domains, with the help of supervised machine learning techniques. Furthermore, we will also demonstrate that uncertainty detection can be fruitfully applied in a real-world application, namely, information extraction from clinical discharge summaries.

Veronika Vincze
Szeged, July 2014

Acknowledgements

First of all, I would like to thank János Csirik for letting me work at the inspiring Human Language Technology Group and providing the opportunity for me to address a variety of new tasks and interesting challenges throughout the years. I am also grateful for his useful remarks on my work, which helped me create the final version of this thesis.

I am indebted to my colleagues and friends for creating a challenging and inspiring atmosphere at our department, which was indispensable in the development of the projects described here. My special thanks go to György Szarvas and Richárd Farkas, with whom I started to work on uncertainty detection: our vivid discussions helped me carry out research on uncertainty and, in the long run, describe the results in the form of this thesis. I would like to thank Gábor Berend for initiating the idea of writing my PhD thesis in computer science, and István Nagy for his constant mental and technical support. A big *thank you* goes to János Zsibrita, who readily answered all of my questions even in the middle of the night. Without them, this thesis would have never been written.

My thanks go to Károly Bibok, who has had a crucial role in turning my interest to computational linguistics, for which I can never be grateful enough. I am also grateful to the following people for their support and inspiration (in alphabetical order): Attila Almási, András Bánhalmi, Eszter Bártházi, Tibor Csendes, István Hegedűs, Rozália Ivády, Melinda Katona, Tamás Kojedzinszky, Márta Maleczki, György Móra, Ágoston Nagy, Katalin Nagy, Enikő Németh T., Magdolna Ohnmacht, Róbert Ormándi, Petra Anna Papp, Tímea Penk, Katalin Simkó, Balázs Szilárd, Tibor Szécsényi, Tamás Táncos, Viktor Varga and the whole Hungarian NLP community...

Last but not least, I would like to thank my parents, my grandparents and my sister for their constant love and support and for believing in me from the beginnings. Thanks are also due to our cat Puszi, who is always ready to argue with me, no matter what the topic is. As a way of expressing my gratitude, I would like to dedicate this thesis to them.

Nothing is impossible. The word itself says: "I'm possible".

Audrey Hepburn

Part I

Background

Chapter 1

Introduction

1.1 Motivation

Natural language processing (NLP) is an interdisciplinary field between artificial intelligence and linguistics: it aims at understanding human language with automatic methods. One of the most widely studied areas of NLP is text mining (TM), which seeks to collect relevant information from free (unstructured) texts. In this way, new knowledge can be quickly gathered from a large amount of texts. However, due to the human linguistic ability of speaking about nonrealistic (non-existing or possible) events or things, the acquisition of reliable information from texts is not straightforward. There are some propositions whose truth value cannot be unequivocally determined, due to the presence of linguistic devices (like modal verbs or adverbs with speculative meaning): these propositions are uncertain and they may be true in some possible worlds but they may be false in other ones. Let us illustrate this with the following examples:

(1.1) It is raining.

(1.2) It is not raining.

(1.3) It is probably raining.

Although each of the above sentences contains the word *raining*, their truth value is quite different. Only the first one states explicitly that there was an event of raining (i.e. the proposition is true), whereas the second one negates it (i.e. the proposition *It is raining* is false) and in the third case, we cannot decide whether the proposition is true or not. The third case is an instance of linguistic uncertainty.

Uncertainty is an important linguistic phenomenon that is relevant in many fields of language processing. In its most general sense, it 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. Thus, uncertain propositions are those whose truth value or reliability cannot be determined due to lack of information. Distinguishing between factual (i.e. true or false) and uncertain propositions is of primary importance both in linguistics and natural language processing applications. For instance, in information extraction (IE) many applications seek to extract factual information from text. Applications should handle detected modified parts in a different manner. A typical example is protein-protein interaction extraction from biological texts, where the aim is to mine text evidence for biological entities that are in a particular relation with each other. Here, while an uncertain relation might be of some interest for an end-user as well, such information must not be confused with factual textual evidence (reliable information). In machine translation, it is also necessary to identify linguistic cues of uncertainty since the source and the target language may differ in their toolkit to express uncertainty (one language employs an auxiliary, the other employs 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.

Although uncertainty has been widely investigated in the literature, there are different terms in usage that denote slightly different linguistic phenomena. Modality is usually associated with uncertainty (Palmer, 1986), but the terms factuality (Saurí and Pustejovsky, 2012), veridicality (de Marneffe et al., 2012), evidentiality (Aikhenvald, 2004) and commitment (Diab et al., 2009) are also used. On the other hand, there are several NLP applications that seek to detect uncertainty in natural language texts in a couple of domains (e.g. biomedical texts (Velldal et al., 2012) or news (Saurí and Pustejovsky, 2012)), however, the variability in the terminology makes it difficult to compare these approaches and evaluate them on a uniform basis. Most of these approaches use annotated databases for evaluation: several uncertainty corpora like BioScope (Vincze et al., 2008b), Genia (Kim et al., 2008), FactBank (Saurí and Pustejovsky, 2009), the CoNLL-2010 Shared Task corpora (Farkas et al., 2010) etc. have been constructed in the last few years. Nevertheless, the lack of unified annotation principles leads to the impossibility of direct comparison of these corpora, which means that each of the above uncertainty detectors is optimized for the corpus or domain on which it was trained. In other words, existing uncertainty detectors can hardly be used across domains, and creating new resources and tools for each domain is time consuming and costly. Instead, a unified comprehensive approach would be optimal, which can be adapted to the specific needs of each domain without extra efforts, and the language independence of the model would also be desirable.

In this thesis, we focus on uncertainty detection in natural language texts. This research question can be investigated from a dual perspective since it is situated in the field of natural language processing, i.e. in the intersection of linguistics and computer science. Thus, in our investigations, we will also make use of linguistic background but the emphasis will be put on computer science: we will solve the task of uncertainty detection with the help of artificial intelligence, and we will make use of machine learning algorithms that we will adapt to the special needs of uncertainty detection.

The main aims of this thesis can be set up as follows. We first list our aims that are related to the field of computer science and then we list those related to linguistics. First, we will detect semantic uncertainty cues in English and Hungarian texts, we will examine the domain specificity of uncertainty cue distribution in the domains of biological texts, Wikipedia articles and news media and we will show how domain adaptation techniques may be exploited to across these domains. Second, we will detect discourse-level uncertainty in English and Hungarian, thus testing the language independence of uncertainty categories. Third, we demonstrate how uncertainty detection may prove useful in a real-world application, namely, information extraction from clinical discharge summaries, where the main task is to classify documents according to whether the patient suffers, probably suffers or does not suffer from a specific illness. Fourth, for these aims, it is essential to offer a unified framework in which all kinds of linguistic uncertainty may be easily placed, forming two main groups: semantic uncertainty and discourse-level uncertainty. This classification is based on considerations from both linguistics and computer science. Fifth, we will present our own annotated corpora that conform to the principles of the above-mentioned unified framework of uncertainty. Sixth, we will compare existing corpora and annotation schemes, with special emphasis on scope-based and event-based annotation schemes and we will argue that these differences may have different implications in practical NLP applications.

1.2 Thesis Roadmap

This thesis can be divided into three main parts. In the first part of the thesis (Background), a general background to uncertainty detection is offered (Chapter 1) and basic concepts of machine learning (ML) are introduced (Chapter 2).

In the second part of the thesis (Uncertainty phenomena in language), uncertainty phenomena are presented as they occur in natural language. Semantics in itself is not sufficient to give account of every uncertainty related linguistic phenomenon. There are syntactic tools to express uncertainty, for instance, passive constructions without explicitly marking the agent usually evoke the concept of weasel, i.e. sourceless propositions (Ganter and Strube, 2009). Furthermore, there are other uncertainty phenomena which cannot be described with the help of semantic or syntactic tools. They are usually related to the pragmatic factors of the discourse or world knowledge. For instance, the reliability of the information may be undermined by extralinguistic facts (e.g. the speaker is known to be a liar or no evidence is given for a statement in a scientific paper) hence it cannot be determined whether the proposition is true or false. In this way, pragmatic aspects of uncertainty and the progressive nature of discourse should be also taken into account (see e.g. de Marneffe et al. (2012), Saurí and Pustejovsky (2012)). Thus, uncertainty is a complex phenomenon that can be understood in depth only if syntactic, semantic and pragmatic aspects are considered at the same time. In Chapter 3, we will propose an interdisciplinary unified framework for all phenomena related to uncertainty, taking into account semantic, syntactic, pragmatic and computational linguistic considerations and we will describe semantic and discourse-level uncertainty phenomena in detail.

Determining whether a given proposition is uncertain or not may involve using a finite dictionary of linguistic devices, i.e. cues. On the other hand, in order to develop a supervised machine learning framework for uncertainty detection, manually annotated corpora are necessary. In Chapter 4, we will present English and Hungarian corpora annotated on the basis of the above-mentioned framework and we will also show the specialties of cue distribution across languages, corpora, genres and domains.

It is not only the concept of uncertainty that might differ from corpus to corpus but the linguistic unit that is marked as uncertain or not can also be different. Moreover, some corpora distinguish levels of uncertainty, i.e. more or less probable statements are separately annotated. In Chapter 5, we will compare the event-based and scope-based methods of uncertainty annotation by contrasting the Genia Event and BioScope 1.0 corpora and we also touch upon the question how levels of uncertainty are distinguished in several corpora.

In the third part of the thesis (Uncertainty detection), uncertainty phenomena are detected in natural language texts. Uncertainty cue candidates do not display uncertainty in all of their occurrences. For instance, the mathematical sense of *probable* is dominant in mathematical texts while 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:

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

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

In order to differentiate between cue and non-cue uses of the same lexical items, we developed a machine learning algorithm, which will be described in Chapter 6. There, we will also focus on the domain-dependent aspects of uncertainty detection in English and we will examine the recognition of uncertainty cues in context.

In Chapter 7, we will address the problem of identifying uncertainty cues in Hungarian texts and we will present our methods developed for solving the task.

Uncertainty detection proves to be useful in real-world applications as well. In Chapter 8, we will illustrate this with the example of identifying obesity and related morbidities in the flow-text parts of clinical discharge summaries. Finally, thesis results will be summarized in Chapter 9.

1.3 Contributions

The main contributions of this thesis can be summarized in six theses, which will be listed in the order of relevance for computer science. As portions of this PhD thesis have previously appeared in several papers by the author (and her co-authors), it seems reasonable to summarize which results were achieved by the author in which publications, which is provided for each thesis. Co-authors of the papers have seen and approved all of the theses and the author's contributions to the papers.

- **Thesis 1**: Detecting different types of semantic uncertainty in several types of English texts by domain adaptation techniques. Proving by statistical means and machine learning experiments that the distribution of uncertainty types and cues is highly dependent on domains and genres. Detecting different types of semantic uncertainty in Hungarian texts, thus proving the language independence of semantic uncertainty categories.

In Szarvas et al. (2012), semantic uncertainty phenomena are identified by a cross-domain uncertainty detector. The author participated in the data preparation and corpus annotation, she designed the uncertainty categories to be identified, she defined some of the features implemented in the machine learning algorithm, she compared the domain- and genre-specific characteristics of the texts concerning uncertainty detection and she carried out the error analysis of the experiments. The co-authors implemented the machine-learning based uncertainty detector and carried out the experiments for English, however, experimental results are considered as a shared contribution of all authors. Experiments on Hungarian (Vincze, 2014) are exclusively the author's own work (see Chapters 6 and 7).

- **Thesis 2**: Detecting different types of discourse-level uncertainty in English and Hungarian texts. Proving that discourse-level uncertainty categories are language-independent.

  In Vincze (2013), the author presents some baseline experiments on identifying discourse-level uncertainty phenomena in English and she also compares her results with those of previous studies. Vincze (2014) introduces machine learning methods for identifying uncertainty in Hungarian texts, based on a rich feature set that includes semantic and pragmatic features as well. All of the results described in these papers are the author's own work (see Chapters 6 and 7).

- **Thesis 3**: Application of uncertainty and negation detection in the medical field: identifying the status of obesity and related diseases in patients.

  In Farkas et al. (2009), it is empirically shown how uncertainty detection can be fruitfully applied in a real-world task, namely, predicting morbidities from clinical texts. The author's main contributions to the paper were offering linguistics-based rules for uncertainty and negation detection, collecting uncertainty cues typical of the medical domain, determining the linguistic scope of such cues and collating dictionaries of relevant medical terms and morbidity names. The latter is a shared contribution with another co-author and statistical methods for term identification and context detection and the application of biomarkers in the system were the contributions of other co-authors. Again, the final results of the system are considered as a shared contribution of all authors (see Chapter 8).

- **Thesis 4**: A language-independent classification of uncertainty phenomena based on theoretical background and empirical evidence, both from the fields of computer science and linguistics.

  In Szarvas et al. (2012) and Vincze (2013), the classification of semantic and discourse-level uncertainty phenomena is presented in detail, which is solely the author's contribution (see Chapter 3).

- **Thesis 5**: Creating and manually annotating benchmark databases for several types of uncertainty: BioScope, CoNLL-2010 Shared Task Corpora, Szeged Uncertainty Corpus, WikiWeasel 2.0, hUnCertainty. Writing the annotation guidelines, annotating, supervising the annotation work, disambiguating annotation differences.

  Vincze et al. (2008b), Vincze (2010b), Farkas et al. (2010), Szarvas et al. (2012), Vincze (2013) and Vincze (2014) introduce the corpora that are annotated for uncertainty phenomena and are exploited in this thesis. The author was responsible for designing the methodology of corpus building, preparing the annotation guidelines, supervising the annotation process, moreover, she also participated in annotating and checking the data. She also carried out a statistical analysis of cue distribution in each corpus, thus proving the domain specificity of uncertainty phenomena. The co-authors of the above papers made only marginal contributions to this thesis like statistical analysis of data and defining some of the general annotation principles in BioScope 1.0 (see Chapter 4).

- **Thesis 6**: Contrasting the differences between the linguistic-based and event-oriented annotation of negation and speculation in biological documents. Proving that the scope-oriented annotation system is more adaptable to non-biomedical applications because of the high level of domain specificity in the event-oriented annotation system.

  In Vincze et al. (2011c), the principles behind scope-based and event-based uncertainty detection are compared on the basis of two corpora. The author's main contributions to this paper were categorizing and analyzing the mismatches between the corpora, providing the principles behind scope-based annotation, offering resolution strategies for mismatches and discussing some of the practical implications of the annotation methodology on uncertainty detection. The co-authors of the paper were responsible for principles behind event-based annotation and statistical analysis of the mismatches (see Chapter 5).

The interrelationship of thesis topics, chapters and theses is visualized in Figure 1.1.

1.4 The Relation of the Author's Publications and Thesis Topics

The relationship of the publications and the above listed theses is visually represented in Table 1.1.

As the author of this PhD thesis has already written a PhD thesis in linguistics, it seems plausible to state that there are no overlaps in the publications used in her PhD thesis in computer science and that in linguistics. Table 1.2 summarizes the relationship of the author's papers and her two PhD theses.

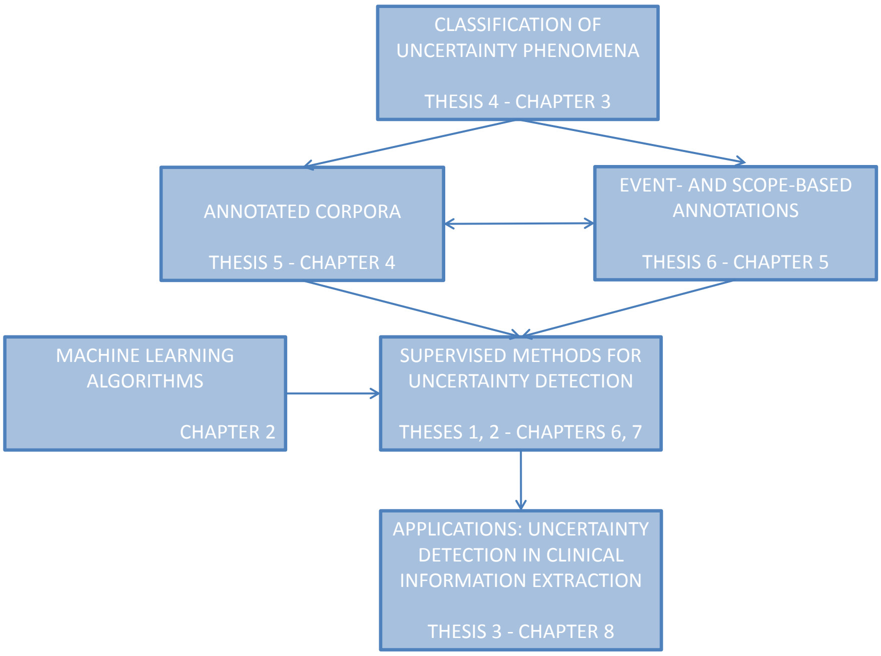
CLASSIFICATION OF UNCERTAINTY PHENOMENA
THESIS 4 - CHAPTER 3
ANNOTATED CORPORA
THESIS 5 - CHAPTER 4
EVENT- AND SCOPE-BASED ANNOTATIONS
THESIS 6 - CHAPTER 5
MACHINE LEARNING ALGORITHMS
CHAPTER 2
SUPERVISED METHODS FOR UNCERTAINTY DETECTION
THESES 1, 2 - CHAPTERS 6, 7
APPLICATIONS: UNCERTAINTY DETECTION IN CLINICAL INFORMATION EXTRACTION
THESIS 3 - CHAPTER 8

Figure 1.1: Thesis topics, chapters and theses.

	Thesis					
	1	2	3	4	5	6
BMC 2008 (Vincze et al., 2008b)					●	
JAMIA 2009 (Farkas et al., 2009)			●			
NESP 2010 (Vincze, 2010b)					●	
CoNLL 2010 (Farkas et al., 2010)					●	
JBMS 2011 (Vincze et al., 2011c)						●
CL 2012 (Szarvas et al., 2012)	●			●	●	
IJCNLP 2013 (Vincze, 2013)		●		●	●	
COLING 2014 (Vincze, 2014)	●	●			●	

Table 1.1: The author's most important publications and the theses.

	PhD thesis	
	computer science	linguistics
ALKNYELVDOK1 2007 (Vincze, 2007)		●
LINGDOK7 2008 (Vincze, 2008)		●
LREC 2008 (Vincze et al., 2008a)		●
BMC 2008 (Vincze et al., 2008b)	●	
ALKNYELVDOK2 2009 (Vincze, 2009a)		●
LINGDOK8 2009 (Vincze, 2009b)		●
MSZNY 2009 (Vincze, 2009c)		●
JAMIA 2009 (Farkas et al., 2009)	●	
ALKNYELVDOK3 2009 (Vincze, 2009d)		●
APPLINGPHD 2009 (Vincze, 2009e)		●
NESP 2010 (Vincze, 2010b)	●	
CoNLL 2010 (Farkas et al., 2010)	●	
ALKNYELV 2010 (Vincze, 2010a)		●
COLING 2010 (Vincze and Csirik, 2010)		●
MSZNY 2010 (Vincze et al., 2010)		●
JBMS 2011 (Vincze et al., 2011c)	●	
LINGDOK10 2011 (Vincze, 2011)		●
MWE 2011 (Vincze et al., 2011a)		●
RANLP 2011 (Vincze et al., 2011b)		●
RANLP 2011 (Nagy T. et al., 2011)		●
CL 2012 (Szarvas et al., 2012)	●	
IJCNLP 2013 (Vincze, 2013)	●	
COLING 2014 (Vincze, 2014)	●	

Table 1.2: The author's most important publications and her Phd theses.

Chapter 2

Basic Concepts in Machine Learning

2.1 Introduction

In this chapter, we offer a brief overview of basic concepts in machine learning that are essential for our approaches to uncertainty detection to be described in Chapters 6, 7 and 8. We first present some basic methods in machine learning and then we describe the most important issues in evaluation methodology.

2.2 Methods in Machine Learning

There are two basic approaches to solve problems in natural language processing. The first one concentrates on exploiting expert knowledge and usually heavily relies on rules defined by linguists. These systems are called rule-based systems and they make use of linguistic information directly integrated into the workflow. However, the second approach relies on (textual) data from which the algorithm learns generalizations on its own (i.e. it does not make use of linguistic information directly), most typically on the basis of statistical inferences, and later on, with the help of these generalizations, it can solve the task on previously unseen data as well. This process is called statistical machine learning.

Machine learning approaches can be classified according to several aspects (Alpaydin, 2010). When the task is to give an instance a label from a set of pre-defined classes (e.g. to classify e-mails as spam or not), it is called classification. On the other hand, when an instance is paired with a real value within an interval (like temperature values), it is called regression. In this thesis, we will concentrate on classification tasks: we will predict what uncertainty class the given unit belongs to (see Chapters 6 and 7 for more details).

The level of supervision is also an important aspect in machine learning. In the case of supervised learning, the algorithm is given a so-called training dataset in which instances are manually labeled by human experts and it extracts some patterns based on them. However, there is no such a prelabeled dataset available for unsupervised learning and in this case, the algorithm finds some regularities and patterns in the data without relying on labeled instances. Finally, semi-supervised learning methods make use of both labeled and unlabeled data. For uncertainty detection, we will apply supervised methods, and the description of the labeled datasets to be used can be found in Chapter 4.

Machine learning models can also be distinguished concerning the unit they make predictions for. Token-based models treat tokens individually, that is, they assign one label to one instance (token) at a time. For instance, token-based models in natural language processing treat each word in a sentence separately and labels are assigned to each word independently. On the other hand, sequence labeling aims at assigning a sequence of labels to a sequence of tokens at a time, e.g. words in a sentence are assigned their particular label at the very same time.

Another possible dimension for distinguishing supervised methods is how they model probabilities. Joint (or generative) methods model the joint probability of the observations and the labels: $P(X, Y)$, where X is the observation set and Y is the set of labels. Conditional (or discriminative) methods, however, model the conditional probability $P(Y|X)$, that is, given the observed data X, what the conditional probability of the class Y is.

In the literature, we can find several supervised machine learning algorithms. However, in the following, we will restrict ourselves to present only those that are relevant for our work here and that will be used in our experiments on uncertainty detection described in Chapters 6 and 7, namely, the Maximum Entropy model (Maxent), which is a token-based discriminative model and Conditional Random Fields (CRF), which is a sequential discriminative model.

2.2.1 Maximum Entropy Model

Maximum Entropy models (Berger et al., 1996) are discriminative models, that is, they work with conditional probabilities, given some observations (features). Features f are elements that link the observation x with the label y. Features are weighted in such a way that the likelihood of the observed labeling (found in the training data) is maximized, that is, the Maxent model provides the probability distribution which has maximum entropy subject to the constraints:

$$p(y|x) = \frac{exp(\sum_{i=1}^{m} \lambda_i f_i(x,y))}{\sum_{y \in Y} exp(\sum_{i=1}^{m} \lambda_i f_i(x,y))}$$

Maxent models will be used in Chapter 6 for detecting semantic uncertainty in English texts.

2.2.2 Conditional Random Fields

Conditional random fields (CRFs) are a type of discriminative undirected probabilistic graphical model used for structured prediction (Lafferty et al., 2001). The most important feature of a CRF model is that it can take context into account: the linear chain CRF predicts sequences of labels for sequences of input samples. Thus, the model does not work with local probabilities like $p(y_t|x_t)$ where t is the position of x within the sequence, instead, it estimates the conditional probability of the whole sequence:

$$p(y|x) = \frac{1}{Z(x)} exp\{\sum_t \sum_{j=1}^{K} \lambda_j f_j(x, y_t, y_{t-1})\}$$

The estimation of weights (λ_j) for each feature f_j is carried out by maximizing the conditional log likelihood:

$$\max_{\lambda} \ell(\lambda) = \max \sum_{i=1}^{N} p(y^{(i)}|x^{(i)})$$

where N is the number of observation sequences $x^{(i)}$ and label sequences $y^{(i)}$.

Training CRFs might be time-consuming for some tasks since the time needed for training depends quadratically on the number of class labels and linearly on the number of training instances and the average sequence length. However, state-of-the-art solutions use CRF models for many NLP tasks where time consumption is still tolerable. In this thesis, we will also employ CRF models for detecting uncertainty both in English and Hungarian (see Chapters 6 and 7).

2.3 Evaluation Methodology

In order to test the efficiency of machine learning approaches, a manually annotated test database is needed, which is not used during the training phase and thus contains unseen examples for the system. Based on generalizations got from the training dataset, the system emits label predictions for each instance in the test, which are later compared to the gold standard labels in the original manual annotation.

A label prediction is considered to be *true positive* (TP) in the case when the instance belongs to the target class in the gold standard dataset and the system correctly predicts its target label, e.g. to a word that is marked as *uncertain* in the test data it correctly assigns the label *uncertain*.

A label prediction is considered to be *false positive* (FP) in the case when the instance does not belong to the target class in the gold standard dataset but the system falsely predicts the target label to be in the target class, e.g. to a word that is not marked as *uncertain* in the test data it falsely assigns the label *uncertain*.

A label prediction is considered to be *false negative* (FN) in the case when the instance belongs to the target class and the system falsely predicts the target label to be in a different class, e.g. to a word that is marked as *uncertain* in the test data it does not assign the label *uncertain*.

A label prediction is considered to be *true negative* (TN) in the case when the instance does not belong to the target class and the system correctly predicts the target label to be different from that of the target class, e.g. to a word that is not marked as *uncertain* in the test data it does not assign the label *uncertain*.

On the basis of the above terms, precision and recall can be calculated. Precision measures how precisely a system predicts a target class (or set of classes), in other words, how many of the instances predicted to belong to a given target class are genuine members of that class.

$$P = \frac{TP}{TP+FP}$$

Recall measures the ratio of instances of a class (or set of classes) that the system actually recognizes as members of the target class in question:

$$R = \frac{TP}{TP+FN}$$

F-measure is defined as the harmonic mean of precision and recall and measures the performance of a system with respect to a target class (or set of classes):

$$F_\beta = (1+\beta^2) \times \frac{P*R}{\beta^2*P+R}$$

Here, β can be used to change the weights of precision and recall, depending on the given application. However, in our evaluation, we will use $\beta = 1$, that is, we will apply F_1-measure, giving equal weights to precision and recall.

When there are several classes of instances which are not distributed equally in the data, it makes sense to distinguish between macro and micro F-measures. Macro F-measure is calculated by simply averaging the F-measures for individual target classes, however, micro F-measure takes into account the frequency of instances for each class, that is, F-measures calculated for each class are weighted according to the frequency of the instances belonging to that class. In the latter way, performance on classes with only few examples does not greatly influence the overall results achieved by the system.

2.4 Summary

In this chapter, we introduced the basic machine learning concepts that will be used in our experiments on uncertainty detection. We described Maximum Entropy and Conditional Random Fields based models and we also presented the evaluation methodology to measure the performance of our systems developed to detect uncertainty in natural language texts, which will be presented in Chapters 6, 7 and 8.

Part II

Uncertainty Phenomena in Language

Chapter 3

A Classification of Uncertainty Phenomena

3.1 Introduction

In order to be able to detect uncertainty in natural language texts, 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 the above 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 this chapter, we suggest a tentative classification of uncertainty phenomena, paying attention to both semantic and discourse-level 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, besides, we also aim to offer a unified framework for all types of uncertainty phenomena. 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).

3.2 Corpora Annotated for Uncertainty

Uncertainty has been paid considerable attention in the last few years in NLP applications. Several corpora annotated for uncertainty have been published in different domains such as biology (Medlock and Briscoe, 2007; Kim et al., 2008; Settles et al., 2008; Shatkay et al., 2008; Vincze et al., 2008b; Nawaz et al., 2010a), medicine (Uzuner et al., 2009), news media (Saurí and Pustejovsky, 2009; Wilson, 2008; Rubin et al., 2005; Rubin, 2010), encyclopedia (Farkas et al., 2010), reviews (Konstantinova et al., 2012; Cruz Díaz, 2013) and microblogs (Wei et al., 2013). Here we briefly summarize their characteristics:

- The Genia Event corpus (Kim et al., 2008), which annotates biological events with negation and two types of uncertainty (9,372 sentences).

- The BioScope corpus (Vincze et al., 2008b), which includes three types of texts from the biomedical domain – namely, radiological reports, biological full papers and abstracts from the Genia corpus – annotated for both negation and hedge keywords and their linguistic scopes (20,924 sentences).

- The system developed by Medlock and Briscoe (2007) made use of a corpus consisting of six papers from genomics literature in which 1,537 sentences were annotated for speculation. These texts – with re-annotation – are also included in BioScope.

- Settles et al. (2008) constructed a corpus comprising of 850 sentences from PubMed abstracts. Sentences are classified as either speculative or definite, however, no keywords are marked in the corpus.

- Shatkay et al. (2008) describe a database where 10,000 biomedical sentences are annotated for polarity and three levels of uncertainty.

- 1,469 gene regulation events are also annotated for four levels of certainty in the E. Coli corpus (Thompson et al., 2008).

- The biological events found in 70 abstracts selected from the Genia Pathway corpus were annotated for knowledge type, uncertainty, polarity, manner and source (Nawaz et al., 2010a).

- The corpus WikiWeasel functioned as the training and evaluation database of the CoNLL-2010 Shared Task (Farkas et al., 2010). It is annotated for weasel cues and consists of 20,745 sentences (4,718 of which are uncertain).

- The FactBank database (Saurí and Pustejovsky, 2009) contains annotation for events, sources and factuality among others, distinguishing four types of factuality.

- A substantial part of the MPQA corpus (Wiebe et al., 2005) – 4,499 sentences in size – is annotated for polarity, thus, for uncertainty too (Wilson, 2008).

- 32 newspaper articles were annotated for four dimensions of certainty – level, perspective, focus and time (Rubin et al., 2005), which database was later extended to include 80 articles from the news domain (Rubin, 2010).

- In the medical domain, assertion classification approaches were tested on 20,208 sentences from discharge summaries and 6,406 sentences from radiological reports in which medical problems were annotated for being present, absent or uncertain (Uzuner et al., 2009).

- 400 reviews on books, movies and consumer products were annotated for uncertainty and negation (Konstantinova et al., 2012; Cruz Díaz, 2013), taken from the SFU review corpus (Taboada et al., 2006) and it contains 17,263 sentences. The annotation principles applied were adapted from those used in the construction of BioScope.

- 4,743 tweets are annotated for uncertainty (Wei et al., 2013), which contain 926 uncertain tweets. The annotation scheme relies heavily on the principles described in Szarvas et al. (2012) and also in Section 3.3.1.

Although these corpora are all annotated for uncertainty, a sharper investigation would reveal that the way they interpret the term *uncertainty* is somewhat different in each corpus. First of all, several different terms are used for the phenomenon such as *hedge*, *speculation*, *uncertainty*, *weasel* etc. In the following, the definition used in four corpora, namely, BioScope, Genia Event, FactBank and WikiWeasel is presented briefly and differences and similarities are discussed.

Speculation is understood in BioScope in the following way: sentences that state the possible existence of a thing, i.e. neither its existence nor its non-existence is unequivocally stated, are considered speculative sentences. A sentence is considered a statement if it does not include any speculative element that suggests uncertainty. In connection with BioScope, the terms *hedge* and *speculation* are also used – *hedge* is most typically employed in the biomedical domain (see e.g. Medlock and Briscoe (2007) or Farkas et al. (2010)) [1].

In the Genia Event corpus events can have three labels of uncertainty: *certain*, *probable* and *doubtful*. Events are marked as *doubtful* if they are under investigation or they form part of a hypothesis, etc. Events are considered *probable* if their existence cannot be stated for certain. The attribute *certain* is chosen by default if none of the two others hold: an event the existence of which cannot be questioned in any way.

FactBank makes use of the terms *certain*, *probable*, *possible* and *underspecified*. *Probable* and *possible* are two classes of uncertain phenomena: they differ in their strength of certainty – a probable event is more likely to take place than a possible one.

[1] It must be mentioned that the term *hedge* may denote different linguistic phenomena for different authors: for instance, when contrasting epistemic modality and hedging, Rizomilioti (2006) categorizes approximators, passive voice and attribution to unnamed source among others as instances of hedging and Hyland (1998) also mentions them among hedging devices. In our study, however, all these phenomena are classified as discourse level uncertainty, see Section 3.3.2.

	BioScope	Genia Event	WikiWeasel	FactBank
keyword	●		●	
target word		●		●
event		●		●
scope	●			
negation	●	●		●
speculation	●		●	
probable		●		●
doubtful		●		
possible				●
underspecified				●
concept of source			●	●
weasel			●	

Table 3.1: Features of the corpora.

Underspecified events are those that cannot be attributed a certainty value because the source is unaware of / ignorant to the event.

While the above corpora all understand uncertainty at the semantic level, weasels, which can be found in WikiWeasel, represent a different type of uncertainty. The origin of the term can be traced back to Steward Chaplin[2], who used this term metaphorically due to the reason that weasels suck the inner parts of an egg without hurting its shell hence it is an empty shell that the weasel leaves behind but it looks like a normal egg (it creates the impression of an egg full of content, however, it is empty). In a similar way, weasel words suck the life from other words. In this case, it is the exact source of the opinion that is missing rather than the factuality of the event: it is known that some hold this opinion but it is unknown who they are: in news media, this is called the "unnamed source" (Bell, 1991). It is a kind of uncertainty expressed at the discourse level as opposed to uncertainty at the semantic level. In this way, the concept of *source* plays an important role in detecting weasels on the one hand and in detecting the uncertainty status of FactBank events (where events and sources are paired by definition). However, WikiWeasel contains annotation for speculation as well, that is, semantic uncertainty is also marked in the database.

Table 3.1 summarizes the features of each corpus on the basis of their original annotation and terms used.

The above corpora use various terms to name the phenomenon uncertainty, for instance: *hedge*, *speculation*, *factuality*, *polarity*, *weasel* while propositions can be *uncertain*, *speculative*, *probable*, *possible* or *doubtful*. As it can be judged from publicly available annotation guidelines, there are many overlaps but differences as well in the understanding of the above terms, which may be sometimes connected to domain- and genre-specific features of the texts. However, for our purposes, it would be preferable to find the common ground among the different terms and concepts of uncertainty. With

[2] http://en.wikipedia.org/wiki/Weasel_word

regard to the above, we aim at sketching a domain- and genre-independent classification of several types of uncertainty, which is inspired by both theoretical and computational linguistic considerations.

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. In the following, we offer a linguistic background for categorizing uncertainty phenomena, describe each type of uncertainty in detail and analyze the annotation schemes of the corpora from the viewpoint of uncertainty categories. Our classification is assumed to be language-independent, but our examples presented here come from the English language, to keep matters simple.

3.3 Classification of Linguistic Uncertainty

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

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

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

3.3.1 Semantic Uncertainty

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.

Here we subcategorize semantic level uncertainty into *epistemic* and *hypothetical* types (Szarvas et al., 2012). The main difference between epistemic and hypothetical uncertainty is that while 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: they describe 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.

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

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 to non-epistemic types of modality as well. Doxastic modality expresses the speaker's beliefs and hypotheses – which may be known as 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 whereas 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).

(3.2) HYPOTHETICAL:

DYNAMIC: I **have to** go.

DOXASTIC: He **believes** that the Earth is flat.

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

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

To sum up, instances of hypothetical uncertainty are:

- doxastic modality (hypotheses and beliefs, i.e. propositions that are assumed but not (yet) confirmed)

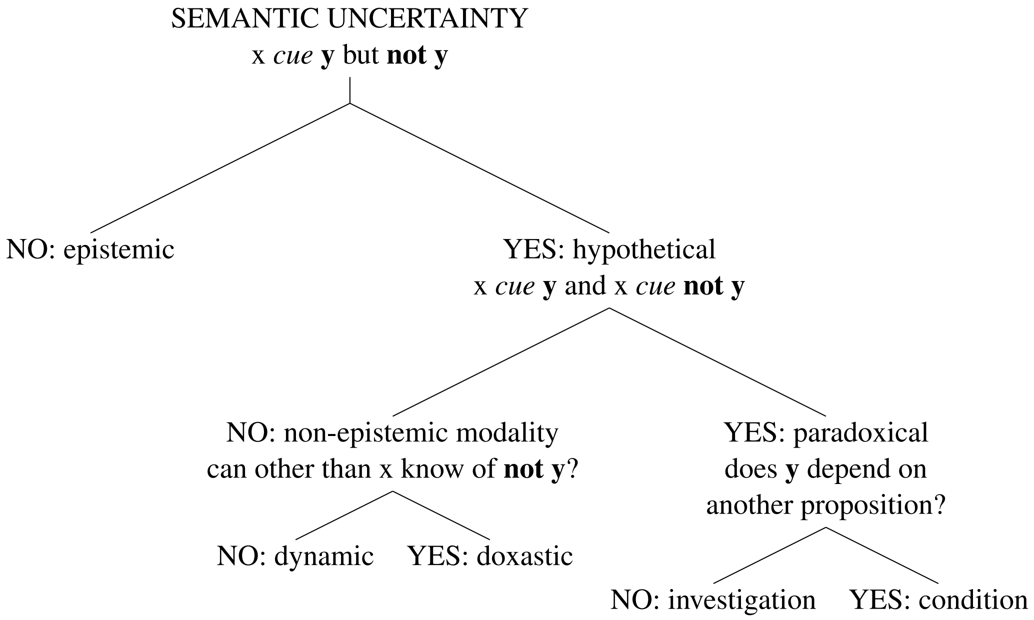

Figure 3.1: Tests for determining semantic uncertainty types.

- propositions under investigation
- conditions (1st and 2nd conditionals (*if*... *then*...), *until*/*unless* clauses)
- dynamic modality:
  - deontic modality (events related to duties, obligations, orders...)
  - buletic modality (plans, intentions, desires...)
  - circumstantial modality (related to external circumstances)
  - dispositional modality (related to physical abilities)

Conditions and instances of dynamic modality are related to 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.

The following test battery is designed to decide what type of semantic uncertainty is present in a sentence under investigation. First, sentences should be normalized, that is, all suspicious uncertainty cue candidates should be removed from the sentence. Nominalized events (e.g. *investigation* or *regulation*) should be transformed into a clause or should be verbalized (*investigate* or *regulate*). The test battery is represented in the form of a decision tree with test questions in each node (see Figure 3.1).

Examples illustrating the test-based classification of propositions are shown below. Cues are in bold.

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

##It **may** be raining but it is not raining.

HYPOTHETICAL:

DYNAMIC: I **have to** go.

I **have to** go but I won't go.

##I **have to** go but I don't **have to** go.

##I **have to** go but others are sure that I won't go.
DOXASTIC: He **believes** that the Earth is flat.

He **believes** that the Earth is flat but the Earth is not flat.

##He **believes** that the Earth is flat and he **believes** that the Earth is not flat.

He **believes** that the Earth is flat but others are sure that the Earth is not flat.
INVESTIGATION: We **examined** the role of NF-kappa B in protein activation.

We **examined** the role of NF-kappa B in protein activation but NF-kappa B has no role in protein activation.

We **examined** the role of NF-kappa B in protein activation and we **examined** the role of NF-kappa B in protein inhibition.

(It does not depend on any other proposition.)
CONDITION: **If** it rains, we**'ll** stay in.

If it rains, we**'ll** stay in but it is sunny, so we'll go out.

If it rains, we**'ll** stay in and **if** it does not rain, we'll go out.

(The propositions depend on each other.)

3.3.2 Discourse-level Uncertainty

We will carefully analyze discourse-level uncertainty phenomena below: we will present the most typical cues and offer examples of propositions considered uncertain at the discourse level. In the following, a detailed presentation of discourse-level uncertainty phenomena is given, which are named after their most typical linguistic markers, i.e. cues. We will concentrate on three key aspects of discourse-level uncertainty, namely, sources, fuzziness and subjectivity.

Weasels

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

Events with no obvious sources are called *weasels* in Wikipedia[3] (Ganter and Strube, 2009): their source is missing or is specified only vaguely or too generally, hence, it cannot be exactly determined who the holder of the opinion is (undetermined source) as it is either not expressed or expressed by an indefinite noun phrase. Weasel sentences usually invoke questions like *Who said that?* and *Who thinks that?* The following sentence illustrates this:

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

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

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

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

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

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

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

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

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

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

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

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

Hedges

Another type of discourse-level uncertainty that will be discussed later on is called a hedge. Although a lot of studies used the term *hedge*, it may denote different linguistic phenomena for different authors. For instance, *hedge* means mostly *speculation* in the biomedical domain (see e.g. Medlock and Briscoe (2007), Vincze et al. (2008b), and Farkas et al. (2010)). When contrasting epistemic modality and hedging, Rizomilioti (2006) categorizes approximators, passive voice and attribution to unnamed sources, among others, as instances of hedging and Hyland (1998) also cites them among common hedging devices.

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

(3.8) Specialized services will **very often** provide a **much** more reliable service based on trusted publications and human reading.

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

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

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

(3.9) This may explain why it has a lower than average estimated albedo of **~0.03**.

(3.10) The duration of attacks averages **3-7** days.

It is interesting to note that in such cases not only cue words but also cue characters are responsible for uncertainty: the tilde and hyphen in these specific cases. Moreover, there are cue words that function as circumscribers as well like *approximately* and another use of *some*:

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

Thus, each type of hedge denotes a set in which the exact amount, quality or frequency of the relevant event or entity is situated but where it is exactly, it remains unclear.

Peacocks

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

Words that express unprovable qualifications or exaggerations are called *peacock* by Wikipedia editors.[4] Their meaning often inherently contains positive or negative subjective judgments, that is, they are polar expressions. Peacock terms include *brilliant*, *excellent* and *best-known*. Although their usage may be acceptable in other contexts, the objective style of Wikipedia editing requires that peacocks should be avoided.

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

Here is a sentence with some peacock terms:

(3.12) Through the **ardent** efforts of Rozsnyai and honorary president Antal Doráti, the Philharmonia Hungarica quickly matured into one of Europe's **most distinguished** orchestras.

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

All peacock terms are similar to hedges to some extent: as they are polar expressions, we should again know from the context what is considered to be the point above or below which something can called be *excellent* (i.e. very positive) or *poor* (i.e. very negative). Thus, the phenomena *hedge* and *peacock* can be called scalar uncertainties since in both cases, a scale is involved in the interpretation of the uncertain term. In the case of peacock, there is a scale of polarity on which phrases can be judged as positive or negative whereas in the case of hedges, there is a scale on which there is a reference

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

point, on the basis of which the uncertain part of the utterance is placed. Although they are similar, we suggest that peacocks and hedges be differentiated in our classification because peacocks are related to subjectivity while hedges are more neutral, hence they can be relevant for different NLP applications (e.g. in opinion mining, which seeks to collect subjective opinions on different topics, peacocks may prove more useful than hedges). Still, hedges shift the value of the quantity / quality mentioned in the text while peacocks denote a specific point on the scale, without modifying it, which again suggests that they should not be lumped in the same class.

3.3.3 Pragmatic Considerations

Discourse level uncertainty can be also analyzed in terms of pragmatics. Since the speaker may not reveal his source or evidence and gives only vague and ambiguous hints on the subject, he does not provide evidence for what is being said, which otherwise would be expected in the conversational context. Therefore, he violates the second part of the Maxim of Quality (Grice, 1975): "Do not say that for which you lack adequate evidence".

By saying something without giving an identifiable source, the speaker may want to imply that he is telling one something very important, something that is true and unquestionable. The semantic content of the proposition is true (there is someone who really says/believes the given piece of information), but the evidentiality and the credibility of the proposition decrease. The implicature behind a weasel sentence is that whatever is said is a general truth, which can be easily refuted in most of the cases, hence the speaker "will be liable to mislead" (Grice, 1975, p.49).

Hedges may also be connected to Gricean maxims. The first part of the Maxim of Quantity ("Make your contribution as informative as required") is violated by hedge sentences. Sentences containing phrases such as *many people*, *approximately 50*, *more than 60%* – in their context – may not be informative enough, thus lack of information is present, giving rise to uncertainty at the discourse level.

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

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

3.4 Comparison with Existing Corpora

The feasibility of the classification proposed in this thesis can be justified by mapping the annotation schemes used in other existing corpora to our categorization 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 (see Chapter 6).

Most of the annotations found in biomedical corpora (Medlock and Briscoe, 2007; Settles et al., 2008; Shatkay et al., 2008; Nawaz et al., 2010a; Thompson et al., 2008) fall into the epistemic uncertainty class. BioScope (Vincze et al., 2008b) 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. As Konstantinova et al. (2012) and Cruz Díaz (2013) followed the BioScope annotation principles, the same applies for their corpus as well. The *probable* class of Genia Event (Kim et al., 2008) is of the epistemically uncertain type while the *doubtful* class belongs to the investigation class. Rubin et al. (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, while 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 3.2 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, while 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.

Although some authors have called attention to the fact that the progressive nature of discourse and dimensions of time should be also taken into account (de Marneffe et al., 2012; Saurí and Pustejovsky, 2012), as can be judged on the basis of available guidelines, most of these corpora make use of semantic uncertainty, with some exceptions that take into account pragmatic or discourse-level information as well that are to be discussed below.

The concept of source has played a significant role in the literature. FactBank (Saurí and Pustejovsky, 2009) explicitly annotates the factuality of events according to their

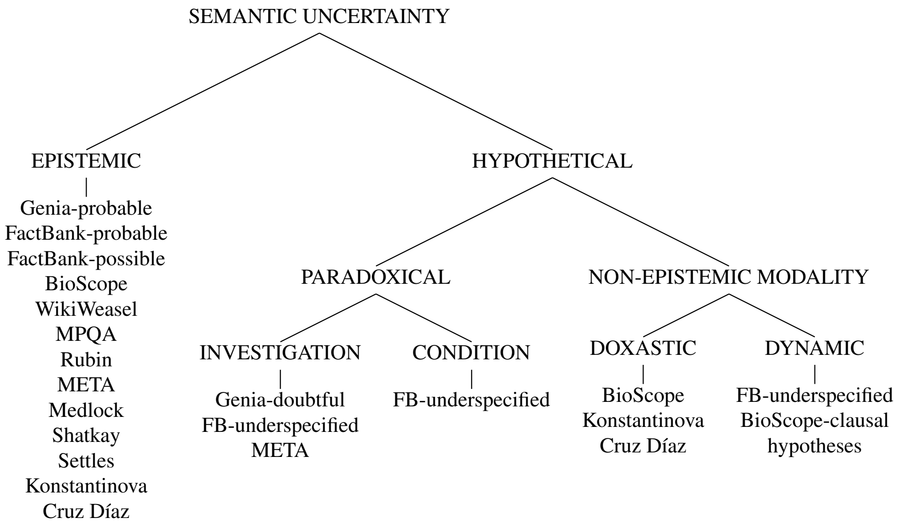

Figure 3.2: Types of uncertainty. FB: FactBank, Genia: Genia Event, Rubin: the dataset described in Rubin et al. (2005), META: the dataset described in Nawaz et al. (2010), Medlock: the dataset described in Medlock and Briscoe (2007), Shatkay: the dataset described in Shatkay et al. (2008), Settles: the dataset described in Settles et al. (2008), Konstantinova: the dataset described in Konstantinova et al. (2012), Cruz Díaz: the dataset described in Cruz Díaz (2013).

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

The lack of source characteristics to weasels can be paired with a certain strategy that Hyland (1996) calls impersonal constructions. It is a type of writer-oriented hedges[5] in his system. It is interesting to note that in his system, the opposite of this strategy can also be found, which could be called *anti-weasel*: the writer emphasizes his responsibility by using first person pronouns. However, this latter strategy does not represent any form of uncertainty in our view.

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

Subjectivity is also related to uncertainty. There is a great diversity among individual

[5] In our classification, however, it should be called a weasel.

views and opinions: a feature of a product may be appreciated by some customers but it might be considered intolerable for others. Thus, what should be considered positive or negative seems subjective.

Many approaches to subjectivity or sentiment analysis rely on lexicons and databases of subjective terms (Wiebe, 2012). The database SentiWordNet (Baccianella et al., 2010) contains a subset of the synsets of the Princeton Wordnet with positivity, negativity and neutrality scores assigned to each concept, depending on the use of its sentiment orientation, thus it is a lexicon where subjective terms are listed and ranked. Wilson (2008) defines subjectivity clues as words and phrases that express private states, that is, individual opinions. She distinguishes lexical cues and syntactic cues that are responsible for subjectivity. She lists several modifiers among her syntactic clues of subjectivity like *quite* and *really*. However, in contrast with other subjective elements, we do not regard them as peacock cues since – as Wilson (2008) herself states – they "work to intensify", so in our system such intensifiers are classified as hedge cues.

3.5 Summary of Results

In this chapter, we offered a classification of uncertainty phenomena on the basis of computational linguistic background, which will be indispensable for creating annotated databases which can serve as a base for training and evaluation datasets in supervised uncertainty detection (our machine learning methods are to be described in Chapters 6 and 7).

The results of this chapter include:

- a language-independent classification of semantic uncertainty;
- a language-independent classification of discourse-level uncertainty;
- a comparison of the annotation principles of existing corpora annotated for uncertainty;
- a unified framework in which all the uncertainty phenomena touched upon in earlier studies can be adequately placed.

These results are described in Szarvas et al. (2012) and Vincze (2013), and they are solely the author's work.

The classification presented here will serve as a theoretical background for the annotation principles of the corpora used for supervised machine learning experiments on uncertainty detection. The corpora will be described in detail in Chapter 4 and our experiments on detecting semantic uncertainty and discourse-level uncertainty in English and Hungarian will be presented in Chapters 6 and 7, respectively.

Chapter 4

Corpora for Uncertainty Detection

4.1 Introduction

In this chapter, corpora which were manually annotated for uncertainty cues and will be exploited in our experiments for uncertainty detection (see Chapters 6 and 7) will be presented. As the training and evaluation of uncertainty detectors require the existence of annotated corpora, we created the benchmark datasets BioScope 1.0 (Vincze et al., 2008b), BioScope 1.5 (Farkas et al., 2010) and BioScope 2.0 (Szarvas et al., 2012), WikiWeasel 1.0 (Farkas et al., 2010), WikiWeasel 2.0 (Szarvas et al., 2012) and WikiWeasel 3.0 (Vincze, 2013) and hUnCertainty and we also reannotated FactBank (Saurí and Pustejovsky, 2009; Szarvas et al., 2012) in harmony with the principles presented in Chapter 3.

When selecting texts for (re)annotation, we paid special attention to the following issues:

- **variety of domains**

  In order to examine linguistic uncertainty from a perspective as wide as possible, texts from multiple domains are required to be annotated. On the other hand, the performance of uncertainty detectors obtained on different domains can be also contrasted.

- **variety of genres**

  Texts belonging to different genres may also differ in the types and cues of uncertainty they contain (e.g. Wikipedia texts are claimed to contain more instances of weasels than scientific papers), which should be paid attention to when developing uncertainty detectors for these genres.

- **multilinguality**

  If texts to be annotated are available in several languages, interlingual comparisons can be easily carried out. Moreover, the universality of annotation principles may also be tested in this way and the adaptability of uncertainty detectors to other languages may also be easily investigated.

In the following, we will present the annotated corpora and we also offer some statistical data on the uncertainty cue distribution in the corpora. With the exception of FactBank 1.0, all of the corpora were created by us.

4.2 The BioScope Corpus

The BioScope corpus (Vincze et al., 2008b) is – to our best knowledge – the largest corpus available that is annotated for both negation and hedge keywords and the first one that contains annotation for linguistic scopes. It includes three types of texts from the biomedical domain – namely, radiological reports, biological full papers (5 full articles from the functional genomics literature (related to the fruit fly) and 4 articles from the open access BMC Bioinformatics website) and abstracts from the GENIA corpus (Kim et al., 2008).

4.2.1 BioScope 1.0

The annotation in BioScope is based on linguistic principles, i.e. parts of sentences which do not contain any biomedical term are also annotated if they assert the non-existence/uncertainty of something. The annotation was carried out by two students of linguistics supervised by a linguist. Problematic cases were continuously discussed among the annotators and dissimilar annotations were later resolved by the linguist. Inter-annotator agreement rates are available at `http://www.inf.u-szeged.hu/rgai/bioscope`.

Speculation is understood in BioScope in the following way: sentences that state the possible existence of a thing, i.e. neither its existence nor its non-existence is unequivocally stated, are considered speculative sentences. Only one level of uncertainty is marked (as opposed to the GENIA corpus (Kim et al., 2008) or Shatkay et al. (2008)). A sentence is considered a statement if it does not include any speculative element that suggests uncertainty.

Negation is seen in BioScope as the implication of the non-existence of something. However, the presence of a word with negative content does not necessarily imply that the sentence should be annotated as negative, since there are sentences that include grammatically negative words but have a speculative meaning or are actually regular assertions.

As for annotating, the most important thing to consider is that hedging or negation is determined not just by the presence of an apparent cue: it is rather an issue of the keyword, the context and the syntactic structure of the sentence taken together. The annotation was based on four basic principles:

- Each keyword has a scope.
- The scope must include its keyword.
- Min-max strategy:
  - The minimal unit expressing hedge/negation is marked as keyword.

  - The scope is extended to the maximal syntactic unit in terms of constituency grammar. The scope of verbs, auxiliaries, adjectives and adverbs usually extends to the right of the keyword. In the case of verbal elements, i.e. verbs and auxiliaries, it ends at the end of the clause (if the verbal element is within a relative clause or a coordinated clause) or the sentence, hence all complements and adjuncts are included.

- No intersecting scopes are allowed.

These principles were determined at the very beginning of the annotation process and they were strictly followed throughout the corpus building.

However, in some cases, some language phenomena seemed to contradict the above principles. These issues required a thorough consideration of the possible solutions in accordance with the basic principles in order to keep the annotation of the corpus as consistent as possible. The most notable examples include the following:

- Negative keywords without scope

  (4.1) [Negative] chest radiograph.

  In this case, the scope contains only the keyword since the scope of negation (i.e. the disease which the patient does not suffer from) is not expressed in the sentence.

- Elliptic sentences

  (4.2) Moreover, ANG II stimulated NF-kappaB activation in human monocytes, but [not] in lymphocytes from the same preparation.

  With the present encoding scheme of scopes, there is no way to signal that the negation should be extended to the verb and the object as well.

- Nested scopes

  One scope includes another one:

  (4.3) These observations (suggest that TNF and PMA do (not lead to NF-kappa B activation through induction of changes in the cell redox status)).

  The semantic interpretation of such nested scopes should be understood as “it is possible that there is no such an event that. . . ”.

- Elements in between keyword and target word

  Although *however* is not affected by the hedge cue in the following example, it is included in the scope since consecutive text spans are annotated as scopes:

  (4.4) (Atelectasis in the right mid zone is, however, <possible>).

- Complex keywords

  Sometimes a hedge / negation is expressed via a phrase rather than a single word: these are marked as complex keywords.

- Inclusion of modifiers and adjuncts

  It is often hard to decide whether a modifier or adjunct belongs to the scope or not. In order not to lose potentially important information, the widest scope possible is marked in each case.

- Intersecting scopes

  When two keywords occur within one sentence, their scopes might intersect, yielding one apparently empty scope (i.e. scope without keyword) and a scope with two keywords:

  (4.5) (Repression did ([not] <seem> to involve another factor whose activity is affected by the NSAIDs)).

  In such cases, one of the scopes (usually the negative one) was extended:

  (4.6) ((Repression did [not] <seem> to involve another factor whose activity is affected by the NSAIDs)).

On the other hand, there were some cases where the difficulty of annotation could be traced back to lexical issues. Some of the keyword candidates have several senses (e.g. *if*) or can be used in different grammatical structures (e.g. *indicate* vs. *indicate that*) and not all of them are to be marked as a keyword in the corpus. Thus, senses / uses to be annotated and those not to be annotated had to be determined precisely.

Finally, sometimes an apparently negative keyword formed part of a complex hedge keyword (e.g. *cannot be excluded*), which refers to the fact that speculation can be expressed also by a negated word, thus, the presence of a negative word does not automatically entail that the sentence is negated.

BioScope 1.0 is freely available for research and educational purposes at `http://www.inf.u-szeged.hu/rgai/bioscope`.

4.2.2 BioScope 1.5

The abstracts and papers from BioScope 1.0 were used as the training dataset of the CoNLL-2010 Shared Task *Learning to Detect Hedges and their Scope in Natural Language Text* (Farkas et al., 2010). The evaluation dataset of the shared task was based on 15 biomedical articles downloaded from the publicly available PubMedCentral database, including 5 random articles taken from the *BMC Bioinformatics* journal in October 2009, 5 random articles to which the *drosophila* MeSH term was assigned and 5 random articles having the MeSH terms *human, blood cells* and *transcription factor* (the same terms which were used to create the Genia corpus). These latter ten articles were also published in 2009. These new texts were manually annotated for uncertainty cues and their scope. To annotate the training and the evaluation datasets, the same annotation principles were applied, thus they were annotated manually for uncertainty cues and their scope by two independent linguist annotators. Any differences between the two annotations were later resolved by the chief annotator, who was also responsible for creating the annotation guidelines and training the two annotators.

In our experiments, BioScope 1.5 will denote the union of papers and abstracts taken from BioScope 1.5, enhanced with the biological evaluation dataset of the CoNLL-2010 Shared Task (i.e. 15 additional papers). BioScope 1.5 is freely available for research and educational purposes at `http://www.inf.u-szeged.hu/rgai/conll2010st`.

4.2.3 BioScope 2.0

In our experiments we wanted to investigate domain and genre differences in uncertainty detection since 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). In order to uniformly evaluate our methods in several domains and genres, the evaluation datasets were normalized. This meant that cues had to be annotated in each dataset and differentiated for types of semantic uncertainty. This resulted in the reannotation of BioScope, WikiWeasel and FactBank. It must be noted that one class of hypothetical uncertainty – namely, 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, texts in BioScope come from the scientific domain, where it is more important to distinguish facts from hypotheses and propositions under investigation (which can be later confirmed or rejected, compare the meta-knowledge annotation scheme developed for biological events (Nawaz et al., 2010a)), and from propositions that depend on each other (conditions).

The originally annotated cues in BioScope 1.5 were separated into epistemic cues and subtypes of hypothetical cues and instances of hypothetical uncertainty not yet marked were also annotated. In this way, BioScope 2.0 was produced. The dataset is freely available for research and educational purposes at `http://www.inf.u-szeged.hu/rgai/uncertainty`.

4.3 The FactBank Corpus

The FactBank corpus contains texts from the newswire domain (Saurí and Pustejovsky, 2009): there are broadcast news and newswire as well. Their topics include financial, political and criminal news as well.

4.3.1 FactBank 1.0

Events are annotated in FactBank 1.0 and they are evaluated on the basis of their factuality from the viewpoints of their sources. Thus, a single predicate can denote several events since their sources may be nested as in:

(4.7) The newspaper discovered that AT&T said it would double its assets.

The event of doubling is annotated from the viewpoint of AT&T, the newspaper and the author of the sentence as well. This reflects the fact that the genre of news is seen as embedded talk where there are multiple sources of the news (Bell, 1991).

Corpus texts contain annotations for polarity (i.e. a sentence is affirmative or negative) and certainty. Three levels of certainty are distinguished in the database: *certain*, *probable* and *possible*. Events for which there is not enough evidence to attribute any of the former labels to the source are marked as underspecified (uncommitted source). Certain events are those that the source thinks they took place / will take place:

(4.8) John knew that Mary came.

Probable events are those that are likely to happen according to the source:

(4.9) Mary will most probably come to the party.

Possible events may or may not happen according to the source:

(4.10) Mary may come to the party.

Conditionals and imperatives are understood by definition as underspecified events, so are prospective events and volitional predicates (*want*, *plan*, *order*. . .).

The corpus originally does not contain annotation for keywords, it is only predicates denoting events that are marked.

FactBank 1.0 is freely available through the Linguistic Data Consortium (http://www.ldc.upenn.edu/Catalog/catalogEntry.jsp?catalogId=LDC2009T23).

4.3.2 FactBank 2.0

FactBank was also reannotated as part of the normalization process described in Section 4.2.3. 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 (Szarvas et al., 2012). FactBank 2.0 is freely available at http://www.inf.u-szeged.hu/rgai/uncertainty.

4.4 The WikiWeasel Corpus

The chief editors of Wikipedia have drawn the attention of the public to uncertainty issues they call *weasel*. A word is considered to be a weasel word if it creates an impression that something important has been said, but what is really communicated is vague, misleading, evasive or ambiguous. Weasel words do not give a neutral account of facts, rather, they offer an opinion without any backup or source. WikiWeasel contains paragraphs taken from Wikipedia, which are manually annotated for uncertainty cues.

4.4.1 WikiWeasel 1.0

WikiWeasel 1.0 was created for the CoNLL-2010 Shared Task (Farkas et al., 2010). For the selection of the Wikipedia paragraphs used to construct the corpus, we exploited the weasel tags added by the editors of the encyclopedia (marking unsupported opinions or expressions of a non-neutral point of view). Each paragraph containing weasel tags (5,874 different ones) was extracted from the history dump of English Wikipedia. First, 438 randomly selected paragraphs were manually annotated from this pool then the most frequent cue phrases were collected. Later on, two sets of Wikipedia paragraphs were gathered on the basis of whether they contained such cue phrases or not. The aim of this sampling procedure was to provide enough training and evaluation samples containing weasel words and also occurrences of typical weasel words in non-weasel contexts.

As the main application goal of weasel detection is to highlight articles which should be improved (by reformulating or adding factual issues), the creators of the corpus annotated only weasel cues in Wikipedia articles, but no scopes were marked.

During the manual annotation process, the following cue marking principles were employed. Complex verb phrases were annotated as weasel cues since in some cases, both the passive construction and the verb itself are responsible for the phenomenon of weasel. In passive forms with dummy subjects and *there is / there are* constructions, the weasel cue included the grammatical subject (i.e. *it* and *there*) as well. As for numerically vague expressions, the noun phrase containing a quantifier was marked as a weasel cue. If there was no quantifier (in the case of a bare plural), the noun was annotated as a weasel cue. Comparatives and superlatives were annotated together with their article. Anaphoric pronouns referring to a weasel word were also annotated as weasel cues.

WikiWeasel 1.0 is freely available for research and educational purposes at `http://www.inf.u-szeged.hu/rgai/conll2010st`.

4.4.2 WikiWeasel 2.0

The normalization process described in Section 4.2.3 also manifested in the reannotation of WikiWeasel 1.0. Epistemic and hypothetical cues were separated from discourse-level cues. WikiWeasel 2.0 is freely available at `http://www.inf.u-szeged.hu/rgai/uncertainty`.

4.4.3 WikiWeasel 3.0

In order to test the practical applicability of the classification of discourse-level uncertainty phenomena described in Chapter 3, and to investigate the frequency of each uncertainty type, we also created an annotated corpus (Vincze, 2013). We selected the texts of WikiWeasel for annotation.

Texts were manually annotated by two linguists for linguistic cues denoting all types of discourse-level uncertainty, i.e. weasel, peacock and hedge. 200 articles were annotated by both linguists and the inter-annotator agreement rate for the categories weasel, peacock and hedge were 0.4837, 0.4512 and 0.4606, respectively (in terms of κ-measure), which reflects that identifying discourse-level phenomena is not straightforward, however, it can be reasonably well solved considering the subjective nature of the task.

During the annotation, special emphasis was laid on the discourse structure of the text. For instance, weasel cue candidates do not denote uncertainty when the sentence is enhanced with citations. Also, a weasel-like element may be elaborated on in the next sentence, thus it is not to be marked as weasel as in:

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

In order to attain the gold standard for the commonly annotated parts, the two annotators discussed problematic cases and reached a consensus for each case. The annotated corpus is available free of charge for research purposes at `http://www.inf.u-szeged.hu/rgai/uncertainty`.

4.5 hUnCertainty 1.0

In order to test the applicability of uncertainty classification to another language and thus to test the language independence of categories, we created a Hungarian corpus as well, which contains manual annotation for epistemic, hypothetical and discourse-level cues as well.

We exploited the characteristics of English uncertainty corpora when preparing the corpus. Hence, the corpus contains 1,091 randomly selected paragraphs from the Hungarian Wikipedia because among the English corpora, it was WikiWeasel that seemed to contain a considerable amount of both semantic and discourse-level uncertainty. Hungarian equivalents of typical uncertainty cues in English were collected and paragraphs containing them were randomly sampled from the Hungarian Wikipedia dump. Besides, paragraphs which did not contain such words were also included in the corpus so as to avoid biased data. Furthermore, we also downloaded 300 pieces of criminal news from a Hungarian news portal (`http://www.hvg.hu`), which makes it possible to compare the distribution of cues in FactBank 2.0 and this dataset from a cross-lingual perspective on the one hand, and, those in Hungarian Wikipedia and news texts from a cross-domain perspective, on the other hand.

Corpus	Sent.	UC sent.	%	Hedge	Negation	Sem.	Disc.
BioScope 1.0 papers	2,670	519	19.44	•	•		
BioScope 1.0 clinical	6,383	855	13.39	•	•		
BioScope 1.0 abstracts	11,797	2,088	17.70	•	•		
BioScope 1.5 papers	7,676	1,309	17.05	•			
BioScope 1.5 abstracts	11,797	2,088	17.70	•			
BioScope 2.0 papers	7,676	1,493	19.45			•	
BioScope 2.0 abstracts	11,797	2,697	22.86			•	
FactBank 1.0	3,123			•	•		
FactBank 2.0	3,123	554	17.74			•	
WikiWeasel 1.0	20,756	4,718	22.73	•			
WikiWeasel 2.0	20,756	2,606	12.56			•	
WikiWeasel 3.0	20,756	7,336	35.34			•	•
hUnCertainty WP	9,678	2,632	27.2			•	•
hUnCertainty news	5,491	1,414	25.75			•	•
hUnCertainty 1.0	15,169	4,046	26.67			•	•

Table 4.1: Number of (uncertain) sentences and features of the corpora.

The Wikipedia subcorpus contains 9,678 sentences and 180,000 tokens. The news subcorpus consists of 5,491 sentences and 94,000 tokens. In total, the hUnCertainty corpus consists of 15,169 sentences and 274,000 tokens.

During annotation, we followed the categorization of uncertainty phenomena as described in Szarvas et al. (2012) and Vincze (2013) with some slight modifications, due to the morphologically rich nature of Hungarian (for instance, modal auxiliaries like *may* correspond to a derivational suffix in Hungarian, which required that in the case of *jöhet* “may come” the whole word was annotated as uncertain, not just the suffix *-het*).

Table 4.1 offers a comprehensive picture on all the corpora presented.

4.6 Uncertainty Cues in the Corpora

In this section, we present some statistical data on cue distribution in the corpora and we also compare differences between genres and domains. These differences may have a considerable influence on the efficiency of our machine learning methods for uncertainty detection, besides, we will carry out cross-domain and cross-genre experiments on uncertainty detection as well (see Chapters 6 and 7 for a more detailed discussion). Hence, it seems plausible to analyze corpus data and cue distribution from these aspects as well.

4.6.1 Genres and Domains

Texts found in the corpora can be categorized into three genres, which can be further divided to subgenres at a finer level of distinction. Figure 4.1 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 whereas

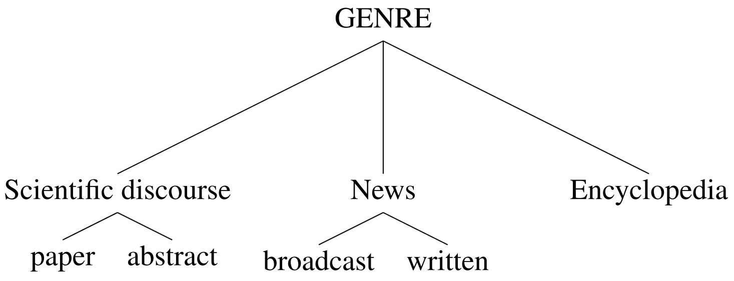

Figure 4.1: Genres of texts.

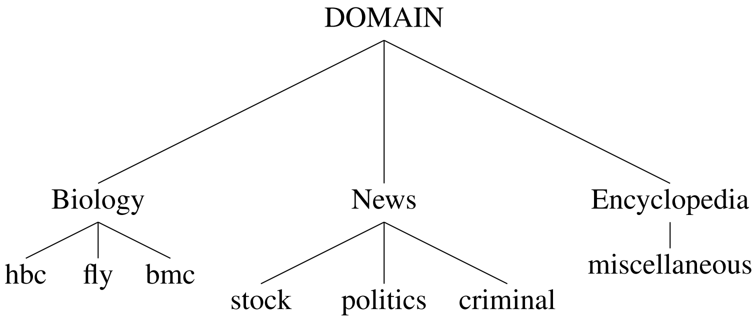

Figure 4.2: Domains of texts.

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 topics at a fine-grained level, which is shown in Figure 4.2. All abstracts and five papers in BioScope are related to the MeSH terms *human*, *blood cell* and *transcription factor* (hbc in Figure 4.2). Nine BMC Bioinformatics papers come from the bioinformatics domain (bmc in Figure 4.2) whereas 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.

4.6.2 Semantic Uncertainty Cues

Three corpora, namely, BioScope 2.0, FactBank 2.0 and WikiWeasel 2.0 contain annotation for semantic uncertainty cues. Table 4.2 provides statistical data on the three corpora. The distribution of different types of semantic uncertainty cues is significant ($p = 0.0042$, ANOVA).

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 4.3 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*,

Dataset	#sent.	#epist.	#dox.	#inv.	#cond.	Total
BioScope 2.0 papers	7676	1373	220	295	187	2075
BioScope 2.0 abstracts	11797	2478	200	784	24	3486
BioScope 2.0 total	19473	3851	420	1079	211	5561
WikiWeasel 2.0	20756	1171	909	94	491	3265
FactBank 2.0	3123	305	201	36	178	720
Total	43352	5927	1530	1209	880	9546

Table 4.2: Data on the English corpora. sent.: sentence, 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

Table 4.3: The most frequent semantic cues in the English corpora. epist.: epistemic cue, dox.: doxastic cue, inv.: investigation cue, cond.: condition cue.

possible, *might* and *suggest* also occur frequently in our dataset.

The distribution of the uncertainty cues was also analyzed from the perspective of uncertainty classes in each corpus, which is presented in Figure 4.3. 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 while in FactBank and WikiWeasel, they cover about 27% of the data. However, the most frequent doxastic keywords exhibit some domain-specific differences: 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 since the aim of scientific publications is to examine whether a hypothesized phenomenon occurs. Among the most frequent cues, *investigate*, *examine* and *study* belong to this group. These data reveal that the frequency

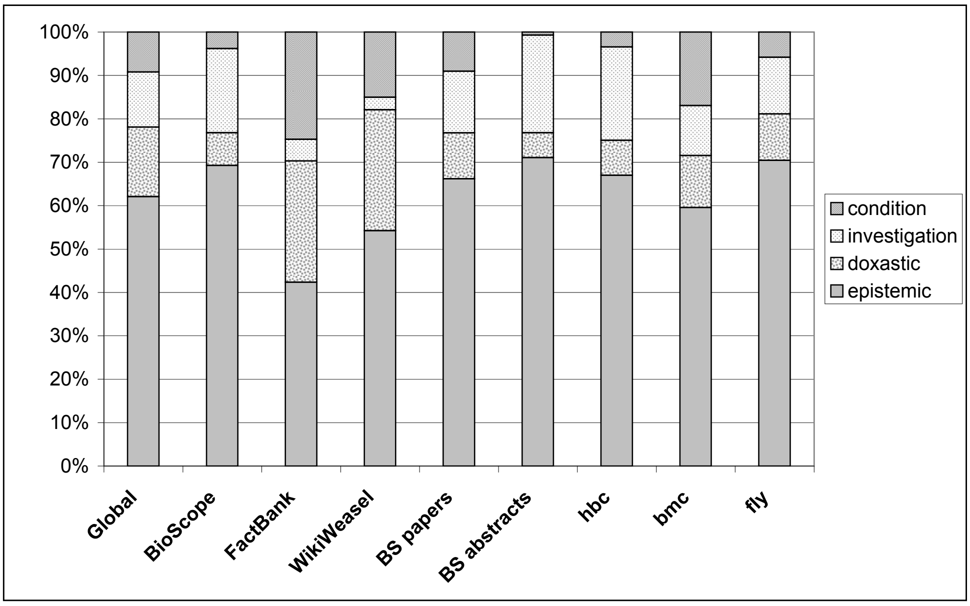

Figure 4.3: Cue type distributions in the English corpora.

of doxastic and investigation cues is strongly domain-dependent, which 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, while *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 could be 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:

(4.12) 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, however, 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 have used. 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 datasets make clear, there were some candidates of investigation verbs which occurred in the investigation sense mostly in abstracts but in another sense in papers, especially in the `bmc` dataset (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` dataset, 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 since 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 some 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 in papers is higher than in abstracts (8.1% and 5.7%, respectively). Besides these differences, the two datasets are quite similar. Results are significant (p = 1.3887E-05, ANOVA).

Domain-related differences can be analyzed when the three subdomains of biological papers are contrasted. As stressed above, `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 – differences are significant (p = 3.04879E-06, ANOVA). As regards lexical issues, the non-cue usage of *possible* in comparative constructions is more frequent in the bmc dataset 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 the above, the fly and the hbc domains seem to be more similar to each other than to the bmc dataset 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 while 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 while in papers, cues belonging to other uncertainty classes can also be found. Thus, when applying techniques developed e.g. for biological texts or abstracts to news texts, 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. However, the adaptation of an uncertainty cue detector constructed for encyclopedia texts requires the special treatment of investigation cues if, for instance, scientific discourse is the target genre since they are underrepresented in the source genre. All these facts will be kept in mind when constructing our uncertainty detector in Chapter 6.

4.6.3 Discourse-level Uncertainty Cues

As it is only WikiWeasel 3.0 that contains annotation for discourse-level uncertainty, we present statistical data on the basis of this corpus. The dataset contains 10,794 discourse-level uncertainty cues[1], which occur in 7,336 uncertain sentences. A sentence was considered to be uncertain if it contained at least one uncertainty cue. But, as the results show, many sentences include more than one uncertainty cue. Statistical data on all the uncertainty cues found in the WikiWeasel corpus are listed in Table 4.4.

Uncertainty cue	#	%	Diff. cues
Hedge	4,743	35.24	260
Weasel	4,138	30.75	99
Peacock	1,913	14.21	540
Discourse-level total	10,794	80.2	899
Epistemic	1,171	8.7	114
Doxastic	909	6.75	36
Conditional	491	3.65	15
Investigation	94	0.7	12
Semantic level total	2,665	19.8	166
Total	13,459	100	1,065

Table 4.4: Uncertainty cues in WikiWeasel 3.0.

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

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

On the number of different cues, Table 4.4 tells us that the set of linguistic cues expressing weasels are the most limited, with almost 100 cues. In contrast, peacock cues vary the most with 540 cues. This suggests that weasels have the most restricted vocabulary in contrast to peacocks, and hedges being in the middle. This also means that the average frequency of a weasel cue is much higher than that of a peacock cue: the average frequency of occurrence of weasel, hedge and peacock cues is 41.8, 18.24 and 3.54, respectively.

We did a more detailed analysis on the lexical distribution of the cues as well. The ten most frequent cues for each type are listed in Table 4.5. These are responsible for about 86%, 45% and 42% of the occurrences of weasel, hedge and peacock cues, respectively. Thus, a limited vocabulary can account for over 85% of weasels.

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

Table 4.5: The most frequent discourse-level uncertainty cues in the WikiWeasel 3.0 corpus.

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

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

4.6.4 Uncertainty Cues in Hungarian

Here we present statistical data on Hungarian uncertainty cues gathered from the hUnCertainty 1.0 corpus. The corpus contains 7,985 uncertainty cues, out of which 5,837 (73.1%) are discourse-level cues and 2,148 (26.9%) are semantic uncertainty cues.

Table 4.6 reports some statistics on the frequency of uncertainty cues in Hungarian and it is also visualized in Figure 4.4. It is revealed that the domain of the texts has a strong effect on the distribution of uncertainty cues: the distribution of semantic uncertainty cues and discourse-level uncertainty cues is balanced in the news subcorpus but in the Wikipedia corpus, about 83% of the cues belong to the discourse-level uncertainty type. These latter data are comparable to the English ones found in WikiWeasel 3.0, where the ratio of discourse-level and semantic uncertainty cues is 1:4 (see Table 4.4).

The distribution of uncertainty cues differs in the two subcorpora, weasels being more frequent in Wikipedia whereas doxastic cues are more probable to occur in the news subcorpus. In the news media, pieces of news are usually reported with their source provided, hence propositions with no explicit source (i.e. weasels) occur rarely in the news subcorpus. Moreover, doxastic cues are related to beliefs and the news subcorpus consists of criminal news (mostly related to murders). When describing the possible reasons behind each criminal act, phrases that refer to beliefs and mental states are often used and thus this type of uncertainty is more likely to be present in news media than in Wikipedia articles.

Uncertainty cue	**Wikipedia**		**News**		**Total**	
	#	%	#	%	#	%
Weasel	1,801	32.02	258	10.93	2,059	25.79
Hedge	2,098	37.3	799	33.86	2,897	36.28
Peacock	787	14	94	3.98	881	11.03
Discourse-level total	4,686	83.3	1,151	48.77	5,837	73.1
Epistemic	439	7.8	358	15.16	797	9.98
Doxastic	315	5.6	710	30.08	1,025	12.84
Conditional	154	2.74	128	5.42	282	3.53
Investigation	31	0.55	13	0.55	44	0.55
Semantic total	939	16.69	1,209	51.22	2,148	26.9
Total	5,625	100	2,360	100	7,985	100

Table 4.6: Uncertainty cues.

If uncertainty is examined at the sentence level, there are 4,046 (26.67%) uncertain sentences in the Hungarian dataset (i.e. they contain at least one cue). More precisely, 3,099 sentences are uncertain at the discourse level (20.44%) and 1,367 sentences are uncertain at the semantic level (9.02%).

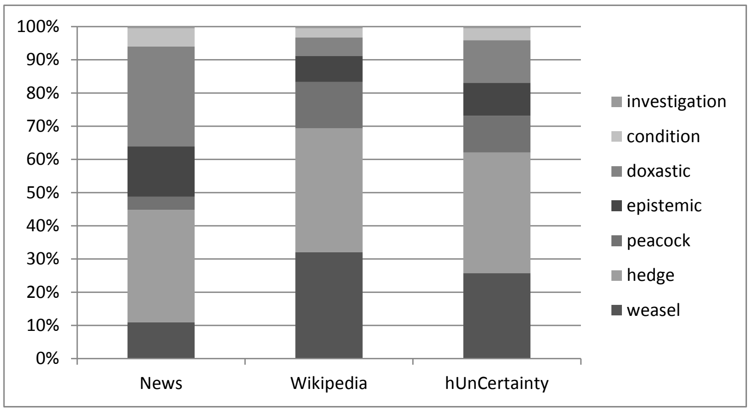

Figure 4.4: Distribution of cues across domains in hUnCertainty 1.0.

Table 4.7 lists the most frequent epistemic and doxastic cues in Hungarian. If they are compared to the corresponding English data, similarities may be found (e.g. *probable*, *consider* and *believe* are among the most frequent cues in both languages), which again underlines the language independence of our categories. The first ten cues are responsible for 45.2% and 71% of the epistemic and doxastic cues, respectively. It is also interesting to note that *szerint* "according to" itself constitutes 43% of the doxastic cues. As conditional and investigation cues do not exhibit a great variety in the corpus, here we just list those cues that occur at least three times: *vizsgál* "examine", *vizsgálat* "examination" and *tanulmányoz* "study" are responsible for 45.5% of investigation cues and *ha* "if", *akkor* "then", *amennyiben* "in case", *volna* "would be", *kell* "must" and *lesz* "will be" represent 71.3% of conditional cue occurrences.

It is also salient from the data that the news subcorpus contains more terms from the criminal domain like *gyanúsít* "accuse". Such difference in terminology may have an influence on machine learning experiments as well, which will be later discussed in Chapter 7.

In Table 4.8, the most frequent discourse-level uncertainty cues are presented.

If compared to the data found in English (see Table 4.5), it can be concluded that there are many overlaps in the discourse-level vocabulary of the two languages (e.g. *some*, *often*, *very*, *many*, *important*, *famous*), which again argues for the language independent applicability of the uncertainty categories described in Chapter 3. The ten most frequent cues cover 41.4%, 34.3% and 28.4% of the weasel, hedge and peacock cues, respectively in Hungarian, which numbers are considerably lower than their English equivalents, especially for weasels.

Some of the Hungarian uncertainty cues are also ambiguous among being uncertain or not on the one hand and belonging to two different uncertainty classes on the other

	Wikipedia	#	News	#	Total	#
Epistemic cues	valószínűleg “probably”	79	lehet “may be”	30	valószínűleg “probably”	101
	talán “maybe”	28	gyanú szerint “as suspected”	26	lehet “may be”	40
	feltehetőleg “presumably”	15	valószínűleg “probably”	22	szerint “according to”	37
	állítólag “supposedly”	14	állítólag “supposedly”	22	állítólag “supposedly”	36
	feltehető “presumable”	11	szerint “according to”	17	talán “maybe”	35
	lehet “may be”	10	gyanúsít “accuse”	16	feltehető “presumable”	27
	lehetséges “possible”	10	feltehető “presumable”	14	gyanú szerint “as suspected”	26
	feltételez “assume”	7	feltételezett “supposed”	13	nem “not”	25
	tekinthető “can be considered”	7	információ szerint “as informed”	11	vélhető “thinkable”	17
	lehetőség “possibility”	6	vélhető “thinkable”	11	feltételezett “supposed”	16
Doxastic cues	szerint “according to”	151	szerint “according to”	298	szerint “according to”	443
	tart “believe”	25	mond “say”	61	mond “say”	62
	tekint “consider”	19	úgy “in such a way”	55	úgy “in such a way”	59
	állít “claim”	18	állít “claim”	23	állít “claim”	41
	vél “think”	10	ír “write”	22	tart “believe”	30
	tulajdonít “attribute”	7	elmond “tell”	19	ír “write”	24
	gondol “think”	6	tagad “deny”	16	elmond “tell”	19
	tesz “assume”	5	beismer “admit”	15	tekint “consider”	19
	hisz “believe”	4	állítás “claim”	12	tagad “deny”	16
	vall “acclaim”	4	vall “acclaim”	11	beismer “admit”	15

Table 4.7: The most frequent epistemic and doxastic uncertainty cues in the hUnCertainty corpus.

	Wikipedia	**#**	**News**	**#**	**Total**	**#**
Weasel cues	számos “several”	151	több “more”	34	számos “several”	160
	egyes “some”	133	információ “information”	15	egyes “some”	137
	egyik “one of”	119	más “other”	15	egyik “one of”	131
	más “other”	106	ismerős “friend”	13	más “other”	121
	néhány “some”	67	egyik “one of”	12	több “more”	74
	stb. “etc.”	48	részlet “detail”	9	néhány “some”	69
	több “more”	40	számos “several”	9	stb. “etc.”	48
	egy “a”	39	szakértő “expert”	7	egy “a”	39
	különböző “different”	37	adat “data”	6	különböző “different”	39
	egyéb “other”	33	sok “many, much”	5	egyéb “other”	35
Hedge cues	általában “generally”	127	több “more”	185	több “more”	201
	gyakran “often”	127	korábban “previously”	52	gyakran “often”	145
	későn “later”	102	sok “many, much”	45	általában “generally”	135
	nagy “big”	92	később “later”	41	későn “late”	102
	jelentős “significant”	56	néhány “some”	38	nagy “big”	93
	nagyon “very”	50	nagy “big”	24	néhány “some”	78
	főleg “mostly”	47	nagyon “very”	22	nagyon “very”	72
	igen “very”	43	gyakran “often”	18	jelentős “significant”	59
	néhány “some”	40	rövid “short”	16	sok “many, much”	57
	főként “mostly”	37	rendszeres “regular”	15	korábban “previously”	52
Peacock cues	fontos “important”	50	fontos “important”	10	fontos “important”	60
	jelentős “significant”	40	ismert “well-known”	6	jelentős “significant”	41
	nagy “great”	29	különös “strange”	4	nagy “great”	33
	ismert “well-known”	24	megrázó “shocking”	4	ismert “well-known”	30
	híres “famous”	22	nagy “great”	4	híres “famous”	25
	kiemelkedő “outstanding”	16	híres “famous”	3	kiemelkedő “outstanding”	18
	komoly “serious”	11	jó “good”	3	erős “strong”	12
	erős “strong”	10	aberrált “aberrated”	2	komoly “serious”	11
	népszerű “popular”	10	agresszív “aggressive”	2	jó “good”	10
	szép “beautiful”	10	erős “strong”	2	népszerű “popular”	10

Table 4.8: The most frequent discourse-level uncertainty cues in the hUnCertainty corpus.

hand. As for the first case, *igen* functions as a hedge when it means “very” but in the “yes” sense, it does not denote uncertainty at all. As for the second case, *nagy* “big, great” occurs among both the most frequent hedge and peacock cues as well: when it refers to the physical size of something, it is a hedge but when it refers to a superior quality, it may function as a peacock.

4.7 Summary of Results

In this chapter, we presented several corpora annotated for uncertainty cues. These corpora will be used in the supervised learning process of our machine learning experiments on uncertainty detection: our algorithms will be trained and evaluated on these datasets (see Chapters 6 and 7). The variety of the corpora makes it also possible to experiment in cross-domain, cross-genre and domain adaptation settings too, hence different machine learning settings can be compared and evaluated in a new applicational field.

The results of this chapter include:

- the English corpora BioScope, FactBank and WikiWeasel were annotated for semantic uncertainty cues;
- WikiWeasel was also annotated for discourse-level uncertainty cues;
- the Hungarian corpus hUnCertainty was annotated for semantic and discourse-level uncertainty cues;
- hUnCertainty and WikiWeasel 3.0 are annotated on the basis of the same principles, and their cue distribution exhibit similarities, which proves the language independence of our classification of uncertainty phenomena presented in Chapter 3;
- statistical data were presented on the frequency of uncertainty cues;
- based on corpus data, the distribution of semantic uncertainty cues was compared across genres and domains, which revealed the domain- and genre-dependency of uncertainty detection.

Vincze et al. (2008b), Vincze (2010b), Farkas et al. (2010), Szarvas et al. (2012), Vincze (2013) and Vincze (2014) introduce these corpora. The author was responsible for designing the methodology of corpus building, preparing the annotation guidelines, supervising the annotation process, moreover, she also participated in annotating and checking the data. She also carried out a statistical analysis of cue distribution in each corpus, thus proving the domain specificity of uncertainty phenomena. The co-authors of the above papers made only marginal contributions to these results like statistical analysis of data in BioScope 1.0 and defining some of the general annotation principles in BioScope 1.0.

In the following chapters, these corpora will serve as a base for uncertainty detection. The annotated corpora are freely available for research purposes at http://www.inf.u-szeged.hu/rgai/uncertainty.

Chapter 5

Scope-Based and Event-Based Uncertainty Annotations and the Strength of Uncertainty

5.1 Introduction

As it was shown in Chapters 3 and 4, there are many corpora that contain annotation for uncertainty. It is not only the concept of uncertainty that might differ from corpus to corpus but the linguistic unit that is marked as uncertain or not can also be different. Moreover, some corpora distinguish levels of uncertainty, i.e. more or less probable statements are separately annotated. However, when there is need for an uncertainty detector, it is essential to know what the exact task and the main goal of uncertainty detection are. That is, in each use case, the end user must specify whether the uncertainty detector should identify:

- uncertain sentences;
- uncertainty cues;
- uncertain text spans (e.g. phrases or clauses, speaking in terms of syntax);
- uncertain events.

Most typically, the following approaches are used to fulfil the above goals. Identifying uncertain text spans is usually preceded by identifying uncertainty cues in the text (Szarvas et al., 2012), and later on, for each cue, its linguistic scope is determined. On the other hand, if uncertain events are to be detected, the identification of events is required and then they are classified whether they are uncertain or not.

To illustrate the differences among these approaches, in this chapter, we aim at comparing the event-based and scope-based methods of uncertainty annotation by contrasting the Genia Event and BioScope 1.0 corpora. We also touch upon the question how levels of uncertainty are distinguished in several corpora and we also pay attention to the implications of the above distinctions in practical applications.

5.2 Scope-Based and Event-Based Uncertainty Annotations

Here we quantitatively compare the negation and speculation annotations of the BioScope 1.0 (Vincze et al., 2008b) and Genia Event (Kim et al., 2008) corpora. As BioScope 1.0 has already been presented in detail in Chapter 4, here we restrict ourselves only to the description of Genia Event annotation principles.

The Genia Event corpus was primarily designed for (biological) event annotation (Kim et al., 2008) and the database contains annotation for uncertainty and negation at the level of events. The annotation scheme focuses on events, and arguments of events can occasionally be found across clause boundaries, typically due to anaphora or coreference (out of 35,419 Genia events used in our experiment, 1,127 referred to an external event and 2,076 clues are arguments of an event expressed in another sentence (mostly cluetypes *theme* (1,447 instances, 70%) and *cause* (619 instances, 29.8%)).

As for uncertainty, events can have three labels in the corpus: `certain`, `probable` and `doubtful`. Events are marked as doubtful if they are under investigation or they form part of a hypothesis, etc. An example (event arguments are underlined in our examples) for a doubtful event is provided here:

(5.1) We then investigated if HCMV binding also resulted in the translation and
secretion of cytokines.

Events are considered probable if their existence cannot be stated for certain. An example of a probable event is shown here:

(5.2) Together, this evidence strongly implicates BSAP in the regulation of the
CD19 gene.

The attribute `certain` is chosen by default if none of the two others hold: an event the existence of which cannot be questioned in any way.

As for negation, events are marked with the labels `exist` or `non-exist`. An example for a negated event is shown below:

(5.3) Analysis of Tax mutants showed that two mutants, IEXC29S and IEXL320G,
were unable to significantly transactivate the c-sis/PDGF-B promoter.

In the corpus, no explicit marking of either the keywords or the scope of negation and hedging can be found.

5.2.1 Methodology

We investigated sentences that occur in both corpora, i.e. the intersection of the two corpora containing 958 abstracts and 8,942 sentences (abstracts that were not segmented in the same way on the sentence level in the two corpora were neglected) was used. This corpus contains 1,287 `negation` and 1,980 `speculation` BioScope 1.0 scopes (376 nested

	TP	B+G-	G+B-
negation	1,554	1,484	569
probable	1,295	3,761	180

Table 5.1: Numbers of agreement and disagreement between BioScope and Genia Event.

scopes) while 2,123 `non-exist` and 1,475 `probable` Genia events (200 events have both labels).

As for negation, events with at least one clue occurring within a negative scope in BioScope 1.0 and being annotated as `non-exist` in Genia Event were considered as cases of agreement. With regard to speculation, events with at least one clue within a speculative scope in BioScope 1.0 and being marked as `probable` in Genia Event were accepted as cases of agreement. Mismatches included events with different labels in the two corpora (e.g. an event labeled as negative in Genia Event and speculative in BioScope 1.0) on the one hand, and events annotated only in one of the corpora on the other hand.

In order to understand the differences between the annotation principles and to investigate the possible contribution of the BioScope annotation to Genia event modality detectors, we randomly sampled 200 sentences from the intersection of the two corpora. This sampling consists of 50 sentences where events are marked to be negated by Genia and none of its arguments was included in a BioScope negation scope and 50 sentences where at least one of the arguments of an event was under a BioScope negation scope and marked as existing by Genia (50+50 sentences were selected for speculation analogously). By manual inspection of this sample we thematically categorized these differences.

5.2.2 Number of Disagreements

Table 5.1 shows the number of cases of agreement and disagreement between the two corpora (agreement rate: 48%). The numbers in column TP (true positive) denote instances which are considered in the same way in both corpora. The numbers in column B+G- refer to cases where in BioScope any clue of a Genia event is under a negative / speculative scope, however, in Genia Event, it is not. As opposed to this, in column G+B- , the numbers show cases where Genia contains some speculative / negative annotation for any argument of the event but BioScope does not.

5.2.3 Categorization of Differences

In this section, mismatches in annotation between the Genia Event and the BioScope corpora are presented. Systematic differences are categorized on the basis of a possible solution aiming to resolve the mismatch, and subtypes of these categories are illustrated with examples along with their estimated frequencies based on a random sample of 200 annotation differences (see Table 5.2).

	B+G-	G+B-
event within event	68%	–
syntax	7%	3%
lexical semantics	20%	72%
morphological negation	–	14%
annotation error	5%	11%
TOTAL	100%	100%

Table 5.2: Frequency of mismatch categories.

Event-centered vs. linguistic annotation

An essential difference in annotation principles between the two corpora is that Genia Event follows the principles of event-centered annotation (Kim et al., 2008) while BioScope annotation does not put special emphasis on events as it aims a task-independent modeling of speculation and negation. Event-centered annotation means that annotators are required to identify as many biological events as possible within the sentence then label each separately for negation and speculation. Events are usually expressed by verbs, however, (deverbal) adjectives and nouns can also refer to events. Consider the following example:

(5.4) Calcineurin acts in synergy with PMA to inactivate I kappa B/MAD3, an inhibitor of NF-kappa B.

This sentence describes two events, the *inactivation of I kappa B/MAD3 by Calcineurin* and the *inhibition of NF-kappa B by I kappa B/MAD3*.

From a linguistic point of view, an event is understood as a predicate together with its arguments and the role of the predicate can be fulfilled by a verb, a noun, or an adjective in the text. In contrast to this, BioScope is not event-oriented in the above sense. Instead, verbs play a central role, i.e. a verb and its arguments form one event in BioScope as well. Accordingly, the above sentence refers to one event in BioScope and *inhibitor* is not considered as a predicate.

As a consequence, there are much more events in Genia than in BioScope. The multiplicity of events in Genia Event and the maximum scope principle exploited in BioScope taken together often yields that a Genia event falls within the scope of a BioScope keyword, however, it should not be seen as a speculated or negated event on its own. Here we provide an illustrative example:

(5.5) In summary, our data [***suggest*** that changes in the composition of transcription factor AP-1 is a key molecular mechanism for increasing IL-2 transcription and may underlie the phenomenon of costimulation by EC].

According to the BioScope analysis of the sentence, the scope of *suggest* extends to the end of the sentence. It entails that although in Genia it is only the events *is a key molecular mechanism* and *underlie the phenomenon* that are marked as probable, the events *changes*, *increasing*, *transcription* and *costimulation* are also included in the BioScope speculative scope. Thus, within this sentence, there are six Genia events out of which two are labeled as probable, however, in BioScope, all six are within a speculative scope, resulting in two cases of agreement and four cases of disagreement. Concerning the whole corpora, the large number of B+G- cases (see Tables 1 and 2) can be explained in a similar way.

Syntactic issues

Some of the mismatches in annotation can be traced back to syntax. For instance, the treatment of subjects remains problematic since in BioScope it is only the complements that are usually included within the scope of a keyword (that is, subjects are not with the exception of passive constructions and raising verbs) in contrast to Genia where events are argument-centered (i.e. complements and subject are considered) as in:

(5.6) Both c-Rel and RelA induced jagged1 gene expression, whereas
a mutant defective for transactivation did [***not***].

In this example, no argument of the event denoted by *induced* is under the BioScope scope, which yields a case of disagreement.

With regard to the problem concerning the treatment of subjects, the dependency parse of the sentence/clause might help the correct identification of the modality of the events. We can apply the following rule: if a verb that functions as the trigger word for an event is negated or hedged, all its children in the dependency tree (including the subject as well) are to be included in the scope of the modifier. In this way, instances of mismatch when it is only the subject that is within the scope of the modifier (e.g. in the case of elliptic sentences) can be eliminated from the G+B- set.

Semantic issues

There are some cases where the difference in annotations originates from conceptual discrepancies. These differences can hardly be resolved without harmonizing the annotation principles behind the corpora and re-annotating the data, however, the most typical cases are presented here.

Events labeled as doubtful in Genia Event are rarely annotated as speculative in BioScope 1.0. In Genia Event, the investigation, examination, study, etc. of a phenomenon does not necessarily mean that the phenomenon exists. Although in BioScope 2.0, cues denoting the investigation type of semantic uncertainty are annotated, in BioScope 1.0 this aspect is neglected and phenomena being under investigation, examination, etc. are only marked as instances of speculation if they are within the scope of a speculative keyword (e.g. *whether*). As only 17% of doubtful Genia event clues is under speculation scope, we focus just on the `probable` class during our comparison.

There are some examples of mismatch where a generalization or a widely accepted claim is stated. Grammatically, these sentences usually occur in the passive voice without explicitly marking the agent (i.e. the one whom the claim originates from). Such sentences are instances of weaseling (see Section 3.3.2), and are annotated as probable events in Genia, however, in BioScope 1.0 they are not as they express a discourse-level type of uncertainty (see Chapter 3). An example for a weasel sentence is shown below:

(5.7) Receptors for leukocyte chemoattractants, including chemokines, are traditionally considered to be responsible for the activation of special leukocyte functions such as chemotaxis, degranulation, and the release of superoxide anions.

Sometimes an event is marked as negation in BioScope but not in Genia:

(5.8) [***Lack of*** full activation of NF-AT] could be correlated to a dramatically reduced capacity to induce calcium flux and could be complemented with a calcium ionophore.

As *lack* is understood as "the state of not having something", it denotes negation, i.e. the non-existence of the following NP complement, that is why it is marked as a negative keyword in BioScope 1.0. However, in Genia, "lack of something" is understood as negation of status, not negation of an event. Hence here the class type of the event is `negative regulation` but the event itself is assertive (out of 4,347 negative regulations in Genia 4,164 are assertive, some of which are annotated as negative in BioScope 1.0 due to semantically negative keywords).

Another case of conceptual discrepancy is morphological negation, i.e. at the morphological level, the cue contains a negative prefix such as *in-* or *un-*. Here is a typical example:

(5.9) In monocytic cells, IL-1beta treatment led to a production of ROIs which is independent of the 5-LOX enzyme but requires the NADPH oxidase activity.

The event denoted by *production* is not triggered by the presence of the 5-LOX enzyme, thus, there is no regulation event here and this is expressed in Genia by marking the regulation event with the attribute `non-exist` while in BioScope 1.0 its meaning is considered to be lexicalized and not necessarily negative.

Mismatches originating from morphological negation mostly include the adjective *independent*. We argue that although this word contains a negative prefix at the level of morphology, its meaning is lexicalized and not necessarily negative: it rather describes a state or a lack of relation between its arguments. In this way, it could be treated similarly to *lack*, that is, not the event itself but its state should be negated. On the other hand, cluewords including morphological negation can be easily identified by automatic methods (segmenting the word into a negative prefix and an existing (adjectival) morpheme) and these can be automatically tagged as negative cues.

The interpretation of some speculative keywords too seems to vary in BioScope 1.0 and Genia Event. The most striking example is the case of events modified by other words or phrases expressing ability (e.g. *be able to*, *ability* etc.), which belong to dynamic modality (see Chapter 3) and are annotated for probability in Genia but not in BioScope 1.0. An example is offered here:

(5.10) NF-kappa B activation correlated with the ability of CD40 to induce Ab secretion and the up-regulation of ICAM-1 and LFA-1.

A highly interesting subclass of words expressing ability is when the derivational suffix conveys the "ability" meaning as in *inducible* or *inhibitable*. Take the following sentence:

(5.11) Despite stimulation with LPS, disruption of the NF-kappaB signaling pathway in precursor B cells led to the loss of inducible Oct-2 DNA binding activity in vitro and the suppression of Oct-2-directed transcription in vivo.

The event described by *inducible* can be paraphrased as *Oct-2 DNA binding activity can be induced in vitro*, which is an 'ability' usage of the auxiliary *can*, thus, it is annotated for probability in Genia but not in BioScope 1.0.

The lexical semantic-related differences originate from conceptual discrepancies of the two corpora. Although some of them have been eliminated due to the annotation scheme used for BioScope 2.0, in general these mismatches can hardly be resolved without totally harmonizing the annotation principles behind the corpora. As one of the chief design goals of BioScope 1.0 annotation was to be task-independent and the modality annotation of Genia is fine-tuned to biological event extraction, biological information extractors may incorporate the modality principles of Genia while BioScope 1.0 annotations may be followed when the target domain differs from the biomedical one.

Lastly we note that few differences (about 5.7%) in the annotation can be obviously traced back to annotation errors.

5.2.4 The Usability of Different Annotation Schemes

As discussed earlier, the annotation scheme of BioScope relies on linguistic principles while Genia Event is based on a more detailed annotation system specifically tailored to biological event annotation, where several complex relations are encoded between participants of the events – often across clause boundaries. In this way, the annotation scheme of Genia Event is highly domain-specific and the corpus can be fruitfully utilized in biomedical information extraction, resulting in a deep and precise analysis of biological events though it might require a lot of additional work to adapt the system to other domains. On the other hand, as the BioScope annotation scheme is linguistic-based, scope- and cue-marking rules extracted from the corpus data can be more easily exploited when developing negation/hedge detectors in other domains as well. The latter has been empirically supported by experiments on a tweet corpus annotated on the basis of the principles described in Chapter 3 (Wei et al., 2013).

Detailed event annotations

Table 1 and 2 reveal that the biggest subset of the differences (60%) came from the issue that Genia handles events within events as individual information sources while BioScope deals with constituent-based text spans. An interesting question for consideration is whether the expected output of an information extraction system consists of facts solely on the basis of this textual evidence, where the trigger for the event does not belong to the main statement of the sentence/document. Note that the information content of these events within events is usually introduced and discussed in detail in other parts of the document or in other publications or belongs to the trivial domain knowledge.

Similar considerations implied the design of the "Meta-Knowledge Annotation Scheme for Bio-Events" (Nawaz et al., 2010b). It introduces dedicated labeling dimensions of events about

- New Knowledge (yes/no), the motivation of which is that events "...could correspond to new knowledge, but only if they represent observations from the current study, rather than observations cited from elsewhere. In a similar way, an analysis drawn from experimental results in the current study could be treated as new knowledge, but generally only if it represents a straightforward interpretation of results, rather than something more speculative."
- Knowledge type (investigation / observation / analysis / general) whose "...purpose is to form the basis of distinguishing between the most critical types of rhetorical/pragmatic intent, according to the needs of biologists."

Krallinger (2010) also argues that from a biologist point of view only the events supported by experimental evidence are interesting. This entails that trivial domain knowledge and assertions without empirical evidence (i.e. weasels) should be treated distinctively. As the BioScope corpus is designed to be task-independent, its scopes could not be applied directly for the deep and detailed (sub)event annotation of Genia since many subevents that belong to trivial domain knowledge fall under scope. However, it can recognize the negation and hedge state of chief statements by exploiting syntactic relations (dependency links) between the keyword marked in BioScope and its trigger word (denoting the chief event): in this way it is possible to determine whether they represent new knowledge or not. Note that there are in-sentence scope detectors published and weasel detectors have been also created recently (see Farkas et al. (2010) and also Chapters 6 and 7).

BioScope for event modality detection

We discussed in the previous section that the scopes of BioScope are not useful directly to the detection of assertion and certainty state of Genia events, however, we believe that using cue phrases in event modality detection can yield significant contribution. For instance, Kilicoglu and Bergler (2009) constructed lexicons for speculation and negation keywords and introduced rules for recognizing the modality state of an event by utilizing the dependency path between the event clue phrase and the speculation/negation cue.

Kilicoglu and Bergler employed hand-crafted lexicons for cue recognition, however, keywords are ambiguous, i.e. they express speculation and negation just in certain contexts. Hence a cue phrase detection system is needed which classifies tokens based on their local context then the dependency paths between these predicted speculation/negation evidences and event triggers should be analyzed. The BioScope corpus can be employed as a training dataset for general speculation/negation cue classifiers. The state-of-the-art modifier cue detectors achieve strict phrase-level F-measures over 80% (Farkas et al., 2010). Dependency-based rules defined for each (sub)type of keywords can be also added to the system in order to determine the negative/speculative status of the event.

Automatically recognized in-sentence scopes (i.e. the negated or hedged text spans) are important for many natural language processing applications. For instance

- in clinical document classification tasks (Pestian et al., 2007; Uzuner, 2009), the goal is to assign labels to medical documents according to factual statements about the patient in question. Here the removal (or separate handling) of hedged or negated text spans has a great contribution in the training and prediction phases as well.
- In information retrieval the query mentions under hedging can be ranked lower,
- in machine translation the extension of negation or speculation scopes has to be precisely known in order to translate meaning adequately.

Although the BioScope corpus consists of clinical and biological documents, its annotation guidelines do not contain any domain-specific instruction. Councill et al. (2010) employed BioScope as training corpus for detecting negated scopes for opinion mining from product reviews, Konstantinova et al. (2012) and Cruz Díaz (2013) used the guidelines of BioScope for constructing their corpora on product reviews and Wei et al. (2013) reports a tweet corpus based on the annotation principles described in Chapter 3, which nicely illustrates the applicability of BioScope's annotations in different tasks and domains.

5.3 Strength of Uncertainty

A further dimension of uncertainty – besides the two main types (semantic and discourse level) – is the degree to which a proposition is uncertain. It is notable that many of the available corpora distinguish between levels of epistemic uncertainty: uncertainty categories differ from each other as the strength of speculation is concerned, e.g. Shatkay et al. (2008) apply three levels of uncertainty, the MPQA corpus exploits four levels of uncertainty (Wilson, 2008) and FactBank (Saurí and Pustejovsky, 2009) differentiates between the more certain *probable* and the less certain *possible* classes. Multiple levels of hypothetical uncertainty can also be distinguished: for instance, there is a difference of strength of uncertainty between the future tense (dynamic modality) expressed by *will* or the *going to* construction: *going to* refers to a proposition which is more likely to happen than the one expressed by *will* (Swan, 1995).

An example of gradually decreasing certainty (i.e. increasing uncertainty) can be observed in the following lyrics:

(5.12) It hurts so bad that I'm never **gonna** drink again

I'll **probably** never drink again
I **may** not ever drink again
At least not 'til next weekend
I'm never **gonna** drink again[1]

It illustrates that the *going to* construction is almost certain, *probably* is somewhat less certain while *may* expresses the lowest level of certainty.

Based on the above, it can be argued that uncertainty can be seen as a scale which can be divided into levels with different uncertainty strength: some corpora make use of such distinction while others only use the scale as a whole dimension without partitioning it. The scalability of uncertainty can be formalized in terms of possible worlds as well: the more probable a proposition is, the higher number of possible worlds belong to it where the proposition holds true.

As for discourse level uncertainty, the credibility or reliability of the source may also influence the strength of uncertainty. Experts, scientists, ministers etc. are considered as credible sources (cf. Katsos and Breheny (2010) and Bell (1991)) while unnamed or unidentifiable sources are less reliable. Certain predicates (e.g. *claim* or *state*) may also undermine the reliability of the information. Compare:

(5.13) **Experts** say that a typhoon will soon arise in this part of the ocean.

(5.14) **People** say that a typhoon will soon arise in this part of the ocean.

(5.15) Cartman **said** he was not guilty.

(5.16) Cartman **claimed** he was not guilty.

Although further investigations are needed concerning this matter, 5.13 is probably judged more reliable than 5.14 and Cartman's innocence is more readily questioned in 5.16 than in 5.15.

The strength of uncertainty is a dimension which is worth examining in detail from both theoretical and computational linguistics aspects. However, its detailed analysis falls outside the scope of this thesis and it remains as a field to be explored in future work.

5.4 Summary of Results

In this chapter, we discussed the differences between the linguistic-based and event-oriented annotation of negation and speculation in biological documents.

The main results of this chapter include:

- categorizing the differences between the linguistic-based and event-oriented annotation of negation and speculation in the intersection of the BioScope 1.0 and Genia Event corpora;

[1] The Offspring: The Worst Hangover Ever

- estimating the frequency of mismatch categories;
- resolution strategies were offered for mismatch categories: syntactic mismatches can be solved by methods based on dependency parsing and the unified annotation scheme can offer solutions for some semantic issues;
- the scope-oriented annotation system is more adaptable to non-biomedical applications because of the high level of domain specificity in the event-oriented annotation system;
- we argued that the strength of uncertainty can manifest at the levels of both semantic and discourse-level uncertainty.

In Vincze et al. (2011c), the principles behind scope-based and event-based uncertainty detection are compared on the basis of two corpora. The author's main contributions to this paper were categorizing and analyzing the mismatches between the corpora, providing the principles behind scope-based annotation, offering resolution strategies for mismatches and discussing some of the practical implications of the annotation methodology on uncertainty detection. The co-authors of the paper were responsible for principles behind event-based annotation and statistical analysis of the mismatches.

Here we showed what kind of uncertainty phenomena can be identified in corpora annotated on the basis of scopes or events. These results may determine the use of existing corpora in information extraction tasks, e.g. if the goal is to identify the assertion and certainty status of events in a text, the implementation of the uncertainty detector may heavily rely on the GENIA Event corpus. Moreover, the analysis of the mismatches in annotation may also affect the construction process of future corpora annotated for uncertainty, more specifically, our results may be useful in determining whether the cue-based or event-based annotation is more suitable for the given application at hand.

Part III

Uncertainty Detection

Chapter 6

Uncertainty Detection in English Texts

6.1 Introduction

In Chapters 3, 4 and 5, we presented the theoretical background for uncertainty annotation and we also described corpora annotated by us, which may be used in supervised machine learning settings that aim the automatic detection of uncertainty phenomena in language. In this chapter, we present our uncertainty cue detector developed for English texts, which is able to distinguish four fine-grained categories of semantic uncertainty (epistemic, doxastic, investigation and condition types). In our investigations, we will make use of the corpora BioScope 2.0, FactBank 2.0 and WikiWeasel 2.0 and 3.0, which enables us to carry out cross-domain and cross-genre experiments. We will also investigate the effects of exploiting outdomain data in the training process and thus, domain adaptation experiments will be also performed. Finally, a baseline method for detecting discourse-level uncertainty in English Wikipedia texts will also be presented.

6.2 Related Work on Uncertainty Cue Detection

Here we overview 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, we are 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, i.e. how to utilize the recognized cues (see, for instance, Kilicoglu and Bergler (2008), Uzuner et al. (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). However, as the focus of our system is cue recognition, we omit their detailed description here.

6.2.1 In-domain Cue Detection

In-domain uncertainty detectors have been developed since the mid '90s. Most of these systems use hand-crafted lexicons for cue recognition and they treat each occurrence of the lexicon items as a cue, i.e. they do not address the problem of disambiguating cues (Friedman et al., 1994; Light et al., 2004; Farkas and Szarvas, 2008; Saurí, 2008; Conway et al., 2009; Van Landeghem et al., 2009). ConText (Chapman et al., 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 et al. (2009) introduced a system which used non-consecutive tokens as cues too (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 dataset). They employed token classification (Morante and Daelemans, 2009; Sánchez et al., 2010; Fernandes et al., 2010; Clausen, 2010) or sequence labeling approaches (Zhang et al., 2010; Li et al., 2010; Rei and Briscoe, 2010; Tang et al., 2010). In both cases the tokens were labeled according to whether they are part of a cue. The latter assigned a label sequence to a sentence (a sequence of tokens) thus it naturally dealt with the context of a particular word. On the other hand, context information for a token was built into the feature space of the token classification approaches. Özgür and Radev (2009) and Velldal (2010) matched cues from a lexicon then applied a binary classifier based on features describing the context of the cue candidate.

Each of these approaches used a rich feature representation for tokens, which usually included surface-level, part-of-speech and chunk-level features. A few systems have also employed dependency relation types originating at the cue (Velldal et al., 2010; Zhang et al., 2010; Sánchez et al., 2010; Rei and Briscoe, 2010); however, 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). 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 (see Chapter 3) 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. However, all of the above studies focused on only one domain, i.e. in-domain cue detection was carried out, which assumes the availability of a training dataset of sufficient size. The only exception we are aware of was 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 6.3.3 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.

6.2.2 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 was to identify uncertain sentences, which was carried out by semi-automatic construction of cue lexicons. The weakly supervised approaches started with very small seed sets of annotated certain and uncertain sentences, and employ bootstrapping to induce a suitable training corpus in an automatic way. Such approaches collected 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 were then collected (in an iterative fashion) on the basis of their similarity to the current training instances. Based on the observation that uncertain sentences tend to contain more than one uncertainty cue, these models successfully extended the seed sets with automatically labeled sentences, and could 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 shared 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, i.e. 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 was also based on the weakly supervised induction of a corpus of uncertain sentences. It exploited 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.

6.2.3 Cue Distribution Analyses

Besides automatic uncertainty recognition, several studies investigated the distribution of hedge cues in scientific papers from different domains (Rizomilioti, 2006; Hyland, 1998; Falahati, 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) while 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. In Chapter 3.2, we also showed that there are domain differences in the distribution of uncertainty cues.

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.

6.3 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.

6.3.1 Corpora Used in the Experiments

In our experiments, we will make use of three corpora: BioScope 2.0, FactBank 2.0 and WikiWeasel 2.0, which are described in detail in Chapter 3.2.

6.3.2 Evaluation Metrics

As evaluation metrics, we employed 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) if and only if it contains at least one cue. Note that the cue-level metric is quite strict as it is based on recognized phrases, i.e. 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.

6.3.3 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 employ 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 below.

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 learnt 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.

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 2 tokens to the left and right).

- **Surface pattern** of the tokens in a window of size 1 (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 3 to 5 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,[1] 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.

CoNLL-2010 experiments

The CoNLL-2010 shared task *Learning to detect hedges and their scope in natural language text* focused on uncertainty detection (Farkas et al., 2010). Two subtasks were defined at the shared task, where the first task sought to recognize sentences that contain some uncertain language in two different domains while 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[2] is identical to our current aim, 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 datasets (i.e. BioScope 1.5 and WikiWeasel 1.0), 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, i.e. 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 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, i.e. 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 while sentence-level models and token classification approaches gave the best results for Wikipedia texts (see Table 6 in Farkas et al. (2010) and also Table

[1] POS-tagging and chunking were performed on all corpora using the C&C Tools (Curran et al., 2007).

[2] 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.

	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 et al., 2010)	78.7	85.2	–	–
BEST/TOK WIKI (Morante et al., 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

Table 6.1: Results on the original CoNLL-2010 datasets.

6.1 here). 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 datasets can be found in Table 6.1. 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)

A comparison of our models with the CoNLL systems reveals that our uncertainty detection model is very competitive when applied on the biological dataset. Our CRF model trained on the official training dataset 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 dataset. 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 dataset: the Maxent model achieved a cue-level F-score of 22.3 and sentence-level F-score of 58.1 on the Wikipedia evaluation dataset, while 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 dataset are rather long and heterogeneous (due to the annotation principles discussed in Chapter 3.2 and also in Section 4.4.1) 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 since CRF is a state-of-the-art model for chunking / sequence labeling tasks.

We conclude from Table 6.1 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 experiments described below we utilized our CRF model, which performed best in detecting uncertainty *cues* in natural language sentences.

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 learnt 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: $\{x_{i,d}, l_i\}$.[3] 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 learnt 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

[3] 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).

sense distributions of the cues ($p(L|X)$) found in different texts. There is some 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), while 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).[4] In this study we always use the term *domain adaptation* to refer to our problem setting, i.e. where both $p(X)$ and $p(L|X)$ are assumed to change.

In our experiments, we used various datasets taken from multiple genres and domains (see Section 4.6.1 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 datasets 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 $\{f_1, f_2, \dots f_n\}$, 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: $\{f_{1,t}, f_{2,t}, \dots f_{n,t}, f_{1,t+s}, f_{2,t+s}, \dots f_{n,t+s}\}$ while source domain examples are described by only the n shared features: $\{f_{1,t+s}, f_{2,t+s}, \dots f_{n,t+s}\}$. Using the union of the source and target training datasets $\{x_{i,t}, l_i\} \cup \{x_{i,s}, l_i\}$ 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 above 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.

6.3.4 Cross-domain and Genre Experiments

We defined several settings (target and source pairs) with varied domain and genre distances and target dataset 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 its subdomains) were not used as source data because FactBank is significantly smaller than the other corpora (WikiWeasel or scientific texts). As the source dataset is typically larger than the target dataset in practical scenarios, news texts can only be used as target data. Abstracts were only used as source data since 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.

[4] 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.

				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

Table 6.2: Experimental results on different target and source domain pairs.

				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 6.3: The absolute difference between the F-scores of Table 6.2 relative to the baseline TARGET setting, repeated from Table 6.2.

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 datasets were used for training). In the purely in-domain setting (TARGET), we performed 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 dataset (as described in Section 6.3.3). Here, we either used each sentence of the source dataset (DA/ALL) or only those sentences that contained a cue observed in the target train dataset (DA/CUE).

Table 6.2 lists the results obtained on various target and source domains in various machine learning settings. The third column contains the ratio of the target train and source datasets' 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). Table 6.3 contains the absolute differences between a particular result and the in-domain (TARGET) results.

Fine-grained semantic uncertainty classification results are summarized in Tables 6.4 and 6.5. Binary F-score corresponds to coarse-grained classification (uncertain vs. cer-

				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 6.4: Comparison of cue-level binary (F_{bin}) and unlabeled F-scores (F_{unl}).

		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

Table 6.5: The per class cue-level F-scores in fine-grained classification.

tain), while unlabeled F-score is the fine-grained classification converted to binary (disregarding the fine-grained category labels). Table 6.4 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 6.5 shows the per class F_{cue} scores, i.e. how accurately our model recognizes the individual uncertainty types. F_{crs}, F_{tgt} and F_{da} correspond to the CROSS, TARGET and DA/CUE settings, respectively (same as above). The DA/ALL setting is not shown for space reasons and due to its similarity to the DA/CUE results.

The size of the target training datasets proved to be an important factor in these investigations. Hence, we performed experiments with different target dataset 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 dataset as in tables 6.2-6.5. However, for each fold of the cross-validation here we just used N sentences (x-axis of the figures) from the target training dataset and a fixed set of 4000 source sentences to alleviate the effect of varying dataset sizes. Figure 6.1 depicts the learning curves for two target/source dataset pairs. 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 6.2. The lower figures show the per class F_{cue} of the DA/ALL model in the fine-grained classification.

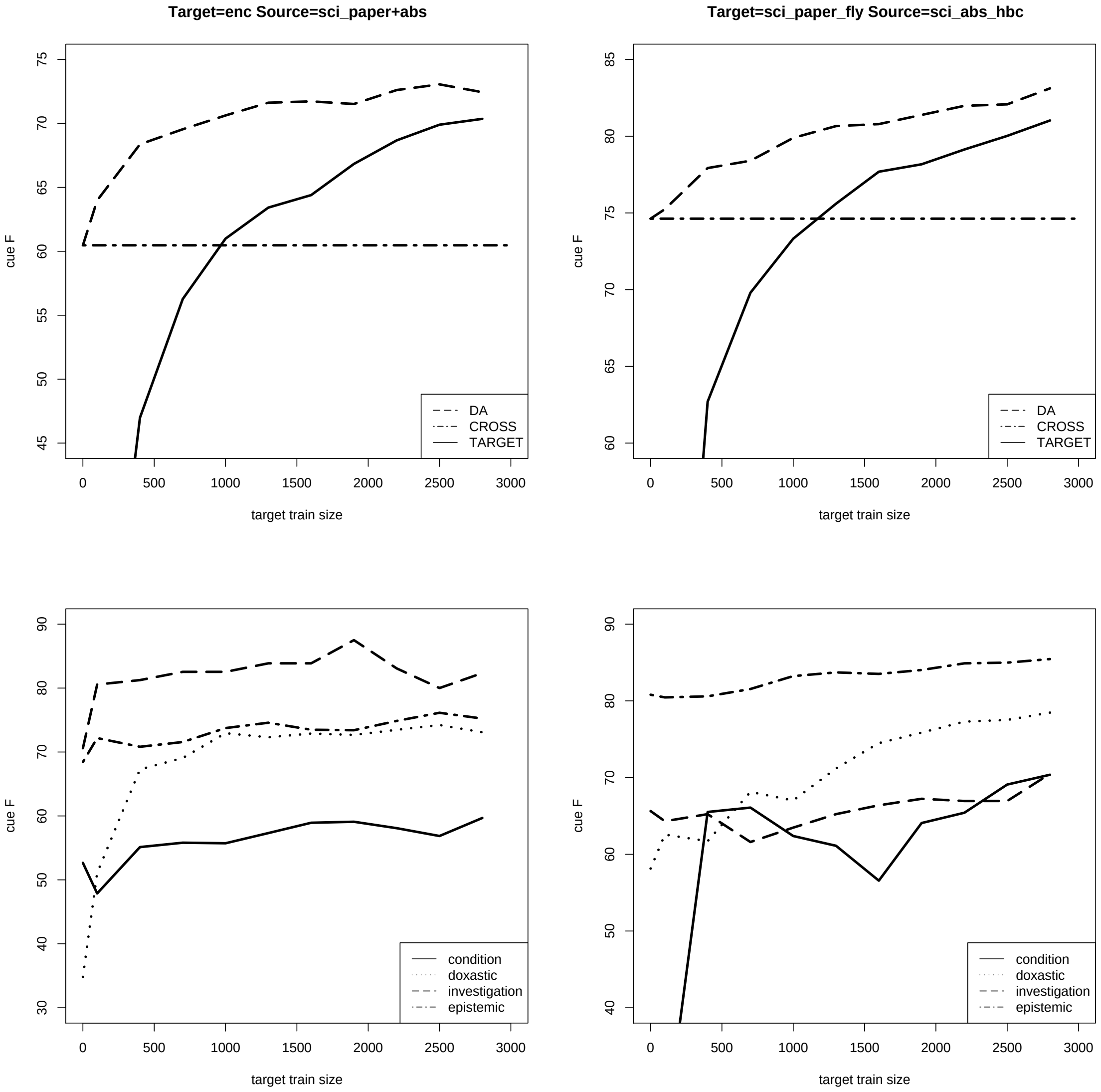

Figure 6.1: Learning curves: results achieved with different target train sizes.

6.4 Discussion

As Table 6.2 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 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 datasets are annotated consistently. This justifies the practical applicability of our unified representation of uncertainty (see Chapter 3) across multiple domains.

6.4.1 Domain Adaptation Results

The size of the target and source datasets 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 below 1%) is where the target dataset 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 datasets: the more distant the domain and genre of the source and target datasets are, the more the CROSS performance – where no labeled target data is used – degrades, compared to the TARGET model. In general, when the distance between both the domain and the genre of texts is substantial (++/++ and +/+ rows in tables 6.2 and 6.3), 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 4.3, for instance). We verified this observation by employing 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`.

However, when there is a domain or genre match between source and target (-/+ and +/- rows in Tables 6.2 and 6.3) 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 learnt uncertainty classifiers can be directly applied to *slightly* different datasets. This suitability is due to the learnt 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 dataset. 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 datasets, 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 6.4 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 is used for the training and thus the datasets 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 6.5) 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 datasets. However, the improvement by domain adaptation is clearly more prominent in fine-grained than in coarse-grained classification: 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 6.5 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 classification and annotation scheme proposed in Chapter 3. 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 datasets: 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 (for some interesting examples, see Section 6.4.3).

6.4.2 The Required Amount of Annotation

Based on our experiments, we may conclude that a manually annotated training dataset 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 6.1) 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 6.1 shows that the size of the target training dataset 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 dataset. Figure 6.1 justifies this observation, as the CROSS and DA settings outperform the TARGET setting on each source-target dataset 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 Chapter 3.2). However, in the news target domain, 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 datasets and hence no reliable observations can be made regarding this subclass (see Table 3.2).

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.

6.4.3 Interesting Examples and Error Analysis

As might be expected, most of the erroneous cue predictions were due to vocabulary differences, e.g. *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. However, many of these anomalies were eliminated by the application of domain adaptation techniques.

Many errors were related to multi-class cues. These cues are especially hard to disambiguate since 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, which are illustrated below:

(6.1) 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.

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

(6.3) 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.

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

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

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

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. However, the future in the past meaning (i.e. past tense of *will*) 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 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. weasel).

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 2.0 and WikiWeasel 2.0. 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:

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

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

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

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

(6.10) 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.

(6.11) VERB: Many of the predictions in these two datasets **concern** protein pairs and proteins that are not present in other datasets.

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 since a lot of 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.

6.5 Discourse-level Uncertainty Detection in English

For a pilot study, we carried out some baseline experiments on the WikiWeasel 3.0 corpus. We divided the corpus into training (80%) and test (20%) sets and applied a simple dictionary-based approach which classified each cue candidate as uncertain if it was tagged as uncertain in at least 50% of its occurrences in the training dataset. For ambiguous cues, the most frequent label was chosen (e.g. *most* was used as a peacock cue). We also examined the effect of merging the labels, i.e. each cue was treated in the same way and no classes were distinguished. Similar to the CoNLL-2010 shared task, we evaluated our results at the cue level as well as at the sentence level. Our results are shown in Table 6.6.

	Cue level			**Sentence level**		
	P	R	F	P	R	F
Weasel	70.88	67.24	69.01	74.43	71.83	73.11
Hedge	87.80	66.16	75.46	91.85	71.93	80.68
Peacock	42.22	47.30	44.62	40.34	53.41	45.97
Micro F	71.96	63.48	67.45	74.58	69.24	71.81

Table 6.6: Baseline results for discourse-level uncertainty detection in terms of precision / recall / F-score.

Table 6.6 shows that the peacock class is the most difficult to detect, which may be due to the fact that this class has the most diverse cues (see Chapter 4 for a detailed analysis of cue distribution) and thus applying a dictionary-based method leads to a lower recall. Still, the lower precision was due to the higher level of ambiguity concerning the most typical peacock cues (like *most*). As for hedges, a simple lexical approach can result in a good precision score, which suggests that hedge cues are less ambiguous than weasel or peacock cues. Merging the labels results in a better performance since cues which are ambiguous among several classes are now treated in a uniform way. It is also seen that sentence-level results are significantly higher than cue-level results (ANOVA, p = 0.0026). Uncertain sentences typically contain more than one cue and in the former scenario, it is sufficient to recognize only one cue in the sentence to regard the sentence as uncertain and false negatives do not affect the performance significantly.

If we compare the data with WikiWeasel 1.0, the CoNLL-2010 version of the corpus, it is seen that the new annotation scheme leads to many more cues (6,725 cue phrases in 4,718 uncertain sentences in the original version vs. 10,794 cues in 7,336 sentences in version 2.0) and – although the datasets are not directly comparable – it gives a much better performance: the best system achieved an F-score of 60.2 on weasel detection at the sentence level and 36.5 at the cue level and no classes of cues were distinguished there (Farkas et al., 2010). This difference may be attributed to several factors. First, not all hedge phenomena (used in the sense introduced here) were systematically annotated in WikiWeasel 1.0. Second, complex syntactic structures that contained several types of uncertainty were annotated as one complex cue (e.g. the phrase *it has been widely suggested*, which contains epistemic uncertainty (*suggested*), weasel (passive sentence with no agent) and hedge (*widely*) as well). Third, WikiWeasel 1.0 did not distinguish subtypes of cues, i.e. semantic uncertainty and weasels were annotated in the same way. It was probably because of this lack of distinction that participants of the shared task got considerably lower results for Wikipedia articles than for biological papers, which contained fewer weasel cues (Farkas et al., 2010). However, the new annotation makes it possible to select those types of uncertainty that are relevant for a given application, e.g. peacocks are important for opinion mining and (numeric) hedges are essential for information retrieval (to find relevant documents for queries that contain numbers). Lastly, weasel detection is of the utmost importance in every information extraction application where it should be known who the author/source is.

6.6 Summary of Results

In this chapter, we presented our uncertainty detector developed for English. The results of this chapter include:

- an accurate semantic uncertainty detector that distinguishes four fine-grained categories of semantic uncertainty (epistemic, doxastic, investigation and condition types);
- our experiments revealed that shallow features provide good results in recognizing semantic uncertainty;
- we achieved successful results for domain adaptation across various domains and genres by applying domain adaptation techniques to fully exploit out-of-domain data and minimize annotation costs to adapt to a new domain;
- a baseline method for detecting discourse-level uncertainty in English Wikipedia texts.

In Szarvas et al. (2012), semantic uncertainty phenomena are identified by a cross-domain uncertainty detector. The author participated in the data preparation and corpus annotation, she designed the uncertainty categories to be identified, she defined some of the features implemented in the machine learning algorithm, she compared the domain- and genre-specific characteristics of the texts concerning uncertainty detection and she carried out the error analysis of the experiments. The co-authors implemented the machine-learning based uncertainty detector and carried out the experiments for English, however, experimental results are considered as a shared contribution of all authors.

In Vincze (2013), the author presents some baseline experiments on identifying discourse-level uncertainty phenomena in English and she also compares her results with those of previous studies.

Chapter 7

Uncertainty Detection in Hungarian Texts

7.1 Introduction

Although uncertainty detection has become one of the most intensively studied problems of natural language processing (NLP) in these days (Morante and Sporleder, 2012), to the best of our knowledge, uncertainty detectors have been mostly developed for the English language (Morante and Sporleder, 2012; Farkas et al., 2010). In this chapter, we present our machine learning based uncertainty detector developed for Hungarian, a morphologically rich language, and report our results on hUnCertainty, a manually annotated uncertainty corpus, which contains texts from two domains: first, Hungarian Wikipedia texts and second, pieces of news from a Hungarian news portal (see Chapter 4). Moreover, we present the first results on applying machine learning techniques to discourse-level uncertainty detection.

7.2 Related Work

In these days, identifying uncertainty cues is one of the popular topics in NLP. This is supported by the CoNLL-2010 Shared Task, which aimed at detecting uncertainty cues in biological papers and Wikipedia articles written in English (Farkas et al., 2010). Moreover, a special issue of the journal Computational Linguistics (Vol. 38, No. 2) was recently dedicated to detecting modality and negation in natural language texts (Morante and Sporleder, 2012). As it is indicated above, most of earlier research on uncertainty detection focused on the English language. As for the domains of the texts, newspapers (Saurí and Pustejovsky, 2009), biological or medical texts (Szarvas et al., 2012; Morante et al., 2009; Farkas et al., 2010; Kim et al., 2008), Wikipedia articles (Ganter and Strube, 2009; Farkas et al., 2010; Szarvas et al., 2012) and most recently social media texts (Wei et al., 2013) have been selected for the experiments.

Systems for uncertainty detection were originally rule-based (Light et al., 2004; Chapman et al., 2007) but recently, they exploit machine learning methods, usually applying a supervised approach (see e.g. Medlock and Briscoe (2007), Morante et al. (2009), Özgür and Radev (2009), Szarvas et al. (2012) and the systems of the CoNLL-2010 Shared Task (Farkas et al., 2010)). In harmony with the latest tendencies, our system here is also based on supervised machine learning techniques, which employs a rich feature set of lexical, morphological, syntactic and semantic features and also exploits contextual features.

Supervised machine learning methods require annotated corpora. There have been several English corpora annotated for uncertainty in different domains such as biology (Medlock and Briscoe, 2007; Kim et al., 2008; Settles et al., 2008; Shatkay et al., 2008; Vincze et al., 2008b; Nawaz et al., 2010a), medicine (Uzuner et al., 2009), news media (Saurí and Pustejovsky, 2009; Wilson, 2008; Rubin et al., 2005; Rubin, 2010), encyclopedia (Farkas et al., 2010), reviews (Konstantinova et al., 2012; Cruz Díaz, 2013) and social media (Wei et al., 2013). For our experiments, however, we make use of hUnCertainty, the first Hungarian uncertainty corpus (Vincze (2014), see also Section 4.5).

7.3 Experiments

In this section, we present our methodology to detect uncertainty cues in Hungarian. We describe our machine learning approach based on a rich feature set. For training and evaluation, we make use of the hUnCertainty corpus (see Chapter 4).

7.3.1 Machine Learning Methods

In order to automatically identify uncertainty cues, we made use of a machine learning method to be discussed below. In our experiments, we used the above-described corpus and morphologically and syntactically parsed it with the help of the toolkit `magyarlanc` (Zsibrita et al., 2013).

On the basis of results reported for English semantic cues (see Szarvas et al. (2012), Farkas et al. (2010) and Chapter 6), sequence labeling proved to be one of the most successful methods on English uncertainty detection (see e.g. Szarvas et al. (2012)), hence we also applied a method based on conditional random fields (CRF) (Lafferty et al., 2001) in our experiments. We used the MALLET implementation (McCallum, 2002) of CRF with the following rich feature set:

- **Orthographic features:** we investigated whether the word contains punctuation marks, digits, uppercase or lowercase letters, the length of the word, consonant bi- and trigrams.

- **Lexical features:** we collected uncertainty cues from the English corpora annotated on the basis of similar linguistic principles and translated these lists into Hungarian. Lists were used as binary features: if the lemma of the given word occurred in one of the lists, the feature was assigned the value *true*, else it was *false*.

- **Morphological features:** for each word, its part of speech and lemma were noted. For each verb, it was investigated whether it had a modal suffix, whether it was in the conditional mood and whether its form was first person plural or third person plural. For each noun, its number was marked as feature. For each pronoun, we checked whether it was an indefinite one. For each adjective, we marked whether it was comparative or superlative.

- **Syntactic features:** for each word, its dependency label was marked. For each noun, it was checked whether it had a determiner and for each verb, whether it had a subject[1].

- **Semantic/pragmatic features:** we compiled a list of speech act verbs in Hungarian and checked whether the given verb was one of them. Besides, we translated lists of English words with positive and negative content developed for sentiment analysis (Liu, 2012) and checked whether the lemma of the given word occurred in these lists.

As contextual features for each word, we applied as features the POS tags and dependency labels of words within a window of size two. Although earlier research on English uncertainty detection mostly made use of orthographical, morphological and syntactic information (see e.g. Szarvas et al. (2012)), here we included some new feature types in our feature set, namely, pragmatic and semantic features.

Based on this feature set, we carried out our experiments. Since only 3% of the tokens in the corpus function as uncertainty cues, it seemed necessary to filter the training database: half of the cueless sentences were randomly selected and deleted from the training dataset. Moreover, as there were only 44 investigation cues in the data, we omitted this class from training and evaluation as well, due to sparseness problems.

First, we applied ten-fold cross validation on the corpus. Since we had two domains of texts at hand, it enabled us to experiment with the two domains separately as well: ten-fold cross validation was carried out for both domains individually and we also made use of cross-domain settings, where one of the domains was used as the training database but the evaluation was performed on the other domain. For evaluation, we used the metrics precision, recall and F-score. The results of our experiments will be presented in Section 7.4.

7.3.2 Baseline Methods

As a baseline, we applied a simple dictionary lookup method. Lists mentioned among the lexical features were utilized here: whenever the lemma of the given word matched one of the words in the list, we tagged it as an uncertainty cue of the type determined by the given list.

[1] Hungarian is a pro-drop language, hence the subject is not obligatorily present in the clause. Moreover, applying a third person plural verb without a subject is a common way to express generalization in Hungarian, which is one typical strategy of weasels.

7.4 Results

Table 7.1 shows the results of the baseline and machine learning experiments on the hUnCertainty corpus, obtained by ten-fold cross validation.

	Dictionary lookup			**Machine learning**			**Difference**		
Type	P	R	F	P	R	F	P	R	F
Weasel	18.12	35.92	24.09	52.48	30.73	38.76	+34.37	-5.19	+14.68
Hedge	55.10	32.42	40.82	61.26	48.94	54.41	+6.17	+16.52	+13.59
Peacock	21.66	30.77	25.42	32.61	11.88	17.41	+10.95	-18.89	-8.01
Epistemic	42.46	30.02	35.18	63.18	34.07	44.27	+20.72	+4.04	+9.09
Doxastic	29.30	46.16	35.85	52.42	46.26	49.15	+23.12	+0.10	+13.30
Condition	31.73	62.90	42.18	51.41	25.80	34.35	+19.68	-37.10	-7.83
Micro P/R/F	29.09	35.74	32.07	55.95	37.46	44.87	+26.86	+1.72	+12.80

Table 7.1: Results on the hUnCertainty corpus.

The results of the machine learning approach have outperformed those achieved by the baseline dictionary lookup method, except for two classes. This is primarily due to better precision, which has grown for each uncertainty category in the case of sequence labeling. However, recall values are more diverse: for hedges and epistemic cues, it has grown, for doxastic cues it has not changed significantly, but for peacocks and conditional cues we can see a serious decrease. The low recall values might be the reason why the F-score obtained by the dictionary lookup method is higher than the one obtained by machine learning in the case of peacocks and conditionals.

	Dictionary lookup			**Machine learning**			**Difference**		
Type	P	R	F	P	R	F	P	R	F
Weasel	3.24	17.83	5.48	37.50	15.12	21.55	+34.26	-2.71	+16.06
Hedge	53.61	39.05	45.18	61.55	49.69	54.99	+7.94	+10.64	+9.80
Peacock	13.82	31.91	19.29	47.06	8.51	14.41	+33.23	-23.40	-4.88
Epistemic	31.90	20.67	25.08	56.63	39.39	46.46	+24.73	+18.72	+21.37
Doxastic	33.50	37.61	35.43	57.05	51.83	54.32	+23.55	+14.23	+18.88
Condition	35.27	57.03	43.58	54.39	24.22	33.51	+19.12	-32.81	-10.07
Micro P/R/F	23.21	34.17	27.65	57.31	41.93	48.43	+34.10	+7.76	+20.78

Table 7.2: Results on the news subcorpus.

We also experimented separately on the two domains. Table 7.2 shows those on the news subcorpus, whereas Table 7.3 shows the results achieved on the Wikipedia subcorpus.

In both domains, we can observe that machine learning methods outperform the baseline dictionary lookup method, except for the peacock and conditional cue classes. However, there are domain differences in the results. First, weasels seem to be much hard to detect in the news subcorpus than in the Wikipedia subcorpus (21.55 vs. 43.8 in terms of F-score). Second, peacocks are also harder to detect in the news subcorpus (F-scores of 14.41 vs. 20.22). Third, there is a considerable gap between the recall scores in the case of doxastic cues: in the Wikipedia subcorpus, the dictionary lookup method outperforms

	Dictionary lookup			**Machine learning**			**Difference**		
Type	P	R	F	P	R	F	P	R	F
Weasel	26.03	38.50	31.06	59.26	34.74	43.80	+33.23	-3.76	+12.74
Hedge	55.86	29.92	38.97	64.59	50.02	56.38	+8.73	+20.10	+17.41
Peacock	23.29	30.63	26.46	37.85	13.8	20.22	+14.56	-16.83	-6.24
Epistemic	49.57	37.34	42.59	63.95	36.03	46.09	+14.38	-1.31	+3.50
Doxastic	25.24	65.20	36.40	54.31	33.54	41.47	+29.07	-31.66	+5.07
Condition	29.66	67.74	41.26	47.12	31.61	37.84	+17.46	-36.13	-3.42
Micro P/R/F	32.28	36.40	34.21	59.70	37.5	46.06	+27.42	+1.10	+11.85

Table 7.3: Results on the Wikipedia subcorpus.

CRF (the difference is 36.13 percentage points) but in the news subcorpus, CRF achieves higher recall with 14.23 percentage points.

To further explore domain differences, we carried out some cross validation experiments. First, we trained our CRF model on the Wikipedia domain and then evaluated it on the news domain. Later, the model was trained on the news domain and evaluated on the Wikipedia domain. Tables 7.4 and 7.5 present the results, respectively, contrasted to the results achieved in the indomain settings. It is revealed that the indomain results almost always outperform the cross-domain results. It is also striking that although the gain in micro F-score is almost the same in the two settings, the biggest difference can be observed for semantic uncertainty classes in the case of the Wikipedia → news setting, while the difference is much bigger for discourse-level uncertainty types in the news → Wikipedia setting.

	Cross validation			**Indomain ten fold**			**Difference**		
Type	P	R	F	P	R	F	P	R	F
Weasel	17.53	19.77	18.58	37.50	15.12	21.55	+19.97	-4.65	+2.97
Hedge	57.40	39.30	46.66	61.55	49.69	54.99	+4.15	+10.39	+8.33
Peacock	22.81	13.83	17.22	47.06	8.51	14.41	+24.25	-5.32	-2.80
Epistemic	50.00	16.76	25.10	56.63	39.39	46.46	+6.63	+22.63	+21.35
Doxastic	46.63	10.70	17.41	57.05	51.83	54.32	+10.43	+41.13	+36.91
Condition	62.96	26.56	37.36	54.39	24.22	33.51	-8.58	-2.34	-3.85
Micro P/R/F	44.48	23.35	30.62	57.31	41.93	48.43	+12.83	+18.58	+17.81

Table 7.4: Cross-domain results: Wikipedia → news.

As some uncertainty detectors aim at identifying uncertain sentences only, that is, they handle the task at the sentence level and do not pay attention to the detection of individual cues (Medlock and Briscoe, 2007), we also applied a more relaxed evaluation metric. If at least one of the tokens within the sentence was labeled as an uncertainty cue – regardless of its type –, the sentence was considered as uncertain. Results on the identification of uncertain sentences are summarized in Table 7.6, in terms of precision, recall and F-score. It is revealed that here there are no sharp differences in performance as far as the indomain settings are concerned since the system can achieve an F-score of about 70 in both domains and on the whole corpus as well. However, in the cross-

	Cross validation			**Indomain ten fold**			**Difference**		
Type	P	R	F	P	R	F	P	R	F
Weasel	71.26	6.87	12.53	59.26	34.74	43.8	-12.00	+27.87	+31.27
Hedge	63.48	26.33	37.22	64.59	50.02	56.38	+1.11	+23.69	+19.16
Peacock	43.14	5.57	9.87	37.85	13.80	20.22	-5.29	+8.23	+10.35
Epistemic	78.65	30.57	44.03	63.95	36.03	46.09	-14.70	+5.46	+2.06
Doxastic	39.55	33.23	36.12	54.31	33.54	41.47	+14.76	+0.31	+5.35
Condition	47.31	28.39	35.48	47.12	31.61	37.84	-0.19	+3.22	+2.36
Micro P/R/F	59.98	18.00	27.68	59.7	37.5	46.06	-0.28	+19.50	+18.38

Table 7.5: Cross-domain results: news → Wikipedia.

domain settings lower precision values and F-scores can be observed, while recall values basically remain the same with regard to the indomain settings.

Evaluation setting	**Precision**	**Recall**	**F-score**
hUnCertainty 10 fold	62.20	78.06	69.23
News 10 fold	67.38	78.01	72.30
Wikipedia 10 fold	60.32	80.05	68.80
Wikipedia → news	45.88	74.21	56.70
News → Wikipedia	35.73	84.61	50.24

Table 7.6: Machine learning results at the sentence level.

7.5 Discussion

Our results prove that a sequence labeling approach can be efficiently used for the automatic identification of uncertainty cues in Hungarian texts. With our baseline dictionary lookup method, the best results were achieved on the epistemic, conditional and hedge cues while the sequence labeling approach was the most successful on the hedge, epistemic and doxastic cues. All of this indicates that hedge and epistemic cues are the easiest to detect. On the other hand, uncertainty types where there was a small difference between the results achieved by the two approaches (for instance, semantic uncertainty cues in the Wikipedia subcorpus) are mostly expressed by lexical means and these cues are less ambiguous. In this setting, the detection of discourse-level uncertainty categories, however, profits more from machine learning, which is most probably due to the fact that here context (discourse) plays a more important rule hence a sequence labeling algorithm is more appropriate for the task, which takes into account contextual information as well.

In the case of peacocks and conditional cues the sequence labeling approach obtained worse results than dictionary lookup: in each case, precision got higher but recall seriously decreased. This suggests that these classes highly rely on lexical features and our machine learning system needs further improvement, with special regard to specific (lexical) features defined for these uncertainty categories.

As for domain differences, we found that the distribution of uncertainty cues differs in the two subcorpora, weasels being more frequent in Wikipedia whereas doxastic cues are more probable to occur in the news subcorpus. Domain differences concerning weasels and doxastic cues are highlighted in the cross domain experiments as well. When the training dataset contains fewer cues of the given uncertainty type, the performance falls back on the target domain: when trained on the news subcorpus, an F-score of 12.53 can be obtained for weasels in the Wikipedia subcorpus, which is 31.27 points less than the indomain results. Similarly, an F-score of 17.41 can be obtained for doxastic cues in the news domain when Wikipedia is used as the training set but the indomain setting yields an F-score of 54.32.

All of the above facts may be related to the characteristics of the texts. Weasels are sourceless propositions and in the news media, it is indispensable to know who the source of the news is, thus, pieces are usually reported with their source provided and so, propositions with no explicit source (i.e. weasels) occur rarely in the news subcorpus. On the other hand, doxastic cues are related to beliefs and the news subcorpus consists of criminal news (mostly related to murders). When describing the possible reasons behind each criminal act, phrases that refer to beliefs and mental states are often used and thus this type of uncertainty is likely to be present in such pieces of news but not in Wikipedia articles.

In the cross domain experiments, indomain results outperform those obtained by the cross domain models. The difference in performance is significant (t-test, p = 0.042 for the news subcorpus and p = 0.0103 for the Wikipedia subcorpus). That is, the choice of the training dataset significantly affects the results, which indicates that there really are domain differences in uncertainty detection. There are only two exceptions that do not correspond to these tendencies: the peacock and conditional cues in the Wikipedia → news setting. The reason why a model trained on a different domain can perform better might lie in the size of the subcorpora. The Wikipedia domain contains much more peacock cues than the news domain and although the domains are different, training on a dataset with more cue instances seems to be beneficial for the results.

If we evaluate the models' performance at the sentence level rather than at the cue level, it can be observed that better results can be achieved, especially with regard to recall values. One reason for that may be that a single uncertain sentence may include more than one cues and should one of them be missed, it does not seriously harm performance (in case at least one cue per sentence is correctly detected).

If our results are compared to those achieved on semantic uncertainty cues found in English Wikipedia articles (see Chapter 6 and Szarvas et al. (2012) as well), it can be seen that the task seems to be somewhat easier in English than in Hungarian: for English, F-scores from 0.6 to 0.8 are reported. However, it must be mentioned here that there are typological differences between English and Hungarian and so, uncertainty marking is rather lexically determined in English but in Hungarian, morphology also plays an essential role. For instance, the modal suffixes *-hat/-het* correspond to the auxiliaries *may* and *might* and while in English they function as separate lexical items, in Hungarian they are always attached to the verbal stem and never occur on their own. As such, applying the word form or the lemma as features may result in relatively high F-scores in English, where the word form itself denotes uncertainty, but these features are less effective in Hungarian without any morphological features included.

The outputs of the machine learning system were further investigated, in order to find the most typical errors our system made. It was revealed that the most problematic issue was the disambiguation of ambiguous cues. For instance, the words *számos* "several" or *sok* "many" may function as hedges or weasels, or *nagy* "big" may be a hedge or a peacock, depending on the context. Such cues were often misclassified by the system. Another common source of errors was that some cues have non-cue meanings as well, like the verb *tart*, which can be a doxastic cue with the meaning "think" but when it means "keep", it is not uncertain at all. The identification of epistemic cues that include negation words was also not straightforward: multiword cues such as *nem zárható ki* "it cannot be excluded" or *nem tudni* "it is not known" were not marked as cues by the system.

7.6 Summary of Results

In this chapter, we presented the first results on Hungarian uncertainty detection. For this purpose, we applied a supervised machine learning approach, which was based on sequence labeling and exploited a rich feature set.

The main results of this chapter are the following:

- the first results on uncertainty detection in Hungarian texts were reported;
- the first machine learning results on discourse-level uncertainty detection were reported;
- new features were introduced in the machine learning setting for uncertainty detection like semantic and pragmatic features;
- we proved that domain specificities have a considerable effect on the efficiency of machine learning.

Results of this chapter are solely the author's work and they are described in Vincze (2014).

Chapter 8

Uncertainty Detection in the Medical Domain: Identifying Obesity and Related Diseases

8.1 Introduction

Medical institutes usually store considerable amount of valuable information (patient data) as free text. Such information has a great potential in aiding research related to diseases or improving the quality of medical care. The size of document repositories makes automated processing in a cost-efficient and timely manner an increasingly important issue. The intelligent processing of clinical texts is the main goal of Natural Language Processing (Ananiadou and Mcnaught, 2005) for medical texts.

In this chapter, we introduce our automatic system for identifying morbidities in the flow-text parts of clinical discharge summaries and we focus on the applicability of uncertainty detectors in a real-life NLP task. The system was designed and implemented for the Obesity Challenge organized by the Informatics for Integrating Biology and the Bedside (I2B2), a National Center for Biomedical Computing in spring 2008. They asked participants to construct systems that could correctly replicate the textual and intuitive judgments of medical experts on obesity and its co-morbidities based on narrative patient records. This task can be regarded as essentially a document classification problem: systems have to assign class labels (according to each morbidity addressed, separately) based on the information found in the whole discharge summary.

8.2 Background

There were several clinical text processing shared tasks in the past few years that were very similar to the obesity challenge in the sense that a document-level decision had to be made, while only a small portion of the text held relevant information for this decision. The smoker challenge organized by I2B2 in 2006 (Uzuner et al., 2008) targeted the identification of the patient's smoker status (smoker, non-smoker, past-smoker, unknown). The clinical coding challenge (Pestian et al., 2007) organized by the Computational Medicine Center of Cincinatti Children's Hospital in 2007 focused on the assignment of ICD codes to radiology reports to enable automated billing.

8.2.1 The Obesity Challenge

The target diseases of the Obesity Challenge included obesity and its 15 most frequent co-morbidities exhibited by patients, while the target labels corresponded to expert judgments based on textual evidence and intuition. That is, for each patient, both what the text explicitly said about obesity and co-morbidities, and what the text implied about obesity and co-morbidities, were provided as gold standard labels by obesity experts. The development of systems that can successfully replicate the decisions made by obesity experts would be desirable to facilitate large scale research on obesity, one of the leading preventable causes of death (Allison et al., 1999; Mokdad et al., 2004; Barness et al., 2007).

The dataset consisted of 1,237 discharge summaries. Each document had been annotated for obesity and the other 15 diseases. Out of these documents, 730 were made available to the challenge participants for development and the remaining 507 documents constituted the evaluation set. For textual annotation, cases of disagreement in labeling were resolved by a resident doctor. For intuitive annotation, there was no tie-breaking so documents with inconsistent labeling were simply excluded from the training and evaluation data for that particular disease. This meant that the number of training and test examples varied from disease to disease, especially for the intuitive task. The third annotator could not decide on a final textual label for about 1% of the documents – these were also discarded (and received no label for that disease). For a more comprehensive description of the task and the data itself, see www.i2b2.org/NLP/ and Uzuner (2009).

8.2.2 Related Work

Even though several results are reported in peer-reviewed literature on medical text classification (Wilcox and Hripcsak, 2003; Hazlehurst et al., 2005; Pakhomov et al., 2006), the most obvious references to work related to this study are the systems submitted to the same challenge by other participants.

28 teams submitted valid predictions to the challenge. The two main approaches of participants were the construction of rule-based dictionary lookup systems and Bag-of-Words (or bi- and trigram based) statistical classifiers.

The dictionaries of rule-based systems mostly consisted of the names of the diseases, and their various spelling variants, abbreviations, etc. One team also used other related clinical named entities (Savova et al., 2008). The dictionaries used were constructed mainly manually (either by domain experts (Childs et al., 2008) or computer scientists (Solt et al., 2008)), but one team applied fully automatic approach to construct their lexicons (Yang et al., 2008).

Machine learning methods applied by participating systems ranged from Maximum Entropy Classifiers (Peshkin et al., 2008) and Support Vector Machines (Savova et al., 2008) to Bayesian classifiers (Naïve Bayes (Califf, 2008) and Bayesian Network (Matthews, 2008)). These systems showed competitive performance on the frequent classes but had major difficulties in predicting the less represented negative and uncertain information in the texts.

8.2.3 Our Approach

Based on previous studies on similar tasks (Szarvas et al., 2006; Farkas and Szarvas, 2008), we observed that the classic uni-, bi- or trigram (or in general n-gram) of words representation – which was originally used for topic-like document classification problems – is not well suited to specific medical text classification problems like the obesity challenge, regardless of the learning method applied. This is mainly because the target pieces of information are in several sentences (possibly fragmented over the text) and the majority of the sentences are irrelevant to the problem. These irrelevant texts count as noise when training statistical classifiers on the vector space representation of the entire document.

The key issue here is finding those single words or phrases that identify relevant text parts (usually sentences) and decisions should be made based on these relevant excerpts of the text after the careful analysis of the located keywords' contexts. In this sense the obesity challenge is more like an Information Extraction task, which gathers the relevant piece of information from scattered sentences of the document, then makes the document-level decision based on the extracted pieces of information.

These aspects motivated us to develop a rule-based system to the challenge that exploits the lists of keywords that trigger important sentences (that is, the names and various spellings of the actual disease) and to implement a simple context analyser that enabled the correct prediction of negative and uncertain information in text. We applied statistical methods to complement, assist and speed-up manual work wherever it proved to be possible. The system can be tested online at `www.inf.u-szeged.hu/rgai/obesity`.

8.3 Methodology

Our approach focused on the rapid development of dictionary-lookup-based systems, which also took into account the document structure and the context of disease terms for classification. To achieve this, we used statistical methods to pre-select the most common (and most confident) terms and evaluated outlier documents by hand to discover infrequent terms and spelling variants. Uncertainty and negation detection exploited keyword lists to identify negations/hedges and delimiter lists to determine their scope. Terms within the scope of a negation or uncertain cue were handled with respect to this information. We expected a system with dictionaries gathered semi-automatically to show a good performance with moderate development costs (we examined just a small proportion of the patient records manually).

8.3.1 Textual Model

For the challenge we applied a dictionary-lookup-based system. That is, we collected a dictionary of terms and abbreviations for each disease separately, processed each document and collected occurrences of dictionary terms from the text. Sentences containing disease terms were then further evaluated to decide the appropriate class label for the corresponding disease. Further evaluations included a judgment of relevance (information on the patient and not on family members, etc.) and an analysis of context to detect negation and uncertainty.

After locating and evaluating all the relevant pieces of information in the document, the main decision function of our system was based on the following rules (the rules were executed in order, and once a rule was matched, the system assigned the relevant classification):

Classify a document as:

1. YES if any terms were matched in an assertive context
2. NO if any terms were matched in a negative context
3. QUESTIONABLE if any terms were matched in an uncertain context
4. UNMENTIONED if none of the previous steps triggered a different labeling.

8.3.2 Intuitive Model

Our intuitive model was based on the textual model. That is, we attempted to discriminate the documents classified as UNMENTIONED by our textual classifier to intuitive YES or NO classes. When the textual system assigned a label that was different from UNMENTIONED, we accepted that decision as an intuitive judgment as well. Obviously this assumption is somewhat simplistic, but based on our observations on the training dataset, this assumption turned out to be quite reasonable.

In order to classify textual UNMENTIONED documents, we collected phrases and numeric expressions which indicated an intuitive YES label. While the phrases were collected using a semi-automated procedure similar to the one used to set up the disease term dictionaries, the numeric expressions describing relevant biomarkers were constructed by hand (there were only a few such expressions and each had a different local context). Such phrases were typically names of associated drugs and medication, or phrases related to certain social habits of the patients (e.g. a cigarette for hypertension), numeric expressions were tension values, weight, etc. Since these terms usually contained implicit information on the corresponding disease, it made no sense to evaluate their context for hedge cues. That is, the lists gathered specifically for the intuitive task were not used to predict intuitive QUESTIONABLE labels.

After locating and evaluating all relevant pieces of information in the document, the main decision function of our system was based on the following rules:

1. Classify textual YES/NO/QUESTIONABLE accordingly

2. For textual UNMENTIONED documents, execute these rules in order until a rule is matched:
   (a) classify a document as an intuitive YES if any intuitive terms were matched in an assertive context
   (b) classify a document as an intuitive YES if a numeric expression was below/above the predefined threshold
   (c) classify a document as an intuitive NO.

8.3.3 System Components

Keyword / Excluding term selection

The terms included in the dictionary were gathered semi-automatically: we filtered them according to their frequency (infrequent terms were discarded in order to reduce the number of term-candidates and avoid overfitting on the data) and then ranked each term according to their positive class (YES) conditional probability scores $(p(yes|word))$. We evaluated the top ranked terms and added the meaningful ones to the disease-name dictionary manually. This way a 95% complete dictionary could be gathered quite rapidly - only the most frequent and reliable few dozens of keywords had to be evaluated manually for every disease. For each disease there were some outlier YES documents that were not captured this way. We examined these documents manually for potential disease terms that were too infrequent to capture by the statistical method (e.g. freq('hyper tg')=2; freq('gallbladder stone')=1).

Next, we collected pseudo terms (i.e. longer phrases containing a previously added disease term that are irrelevant to the disease) using a similar semi-automated procedure. This step was performed so as to avoid the overfitting of the dictionary lookup system (e.g. 'depression', but not 'st. depression' or 'hypertension' but not 'pulmonary hypertension').

The disease name dictionaries we collected were then extended with a few spelling variants manually, to handle different spellings of the same term.

Irrelevant contexts

We also made use of an UNMENTIONED dictionary that triggered the exclusion of the text from further processing. This way we neglected sections under headings like 'FAMILY HISTORY:' and also phrases like 'son with…', 'family history of…'. Irrelevant headings triggered the exclusion of the whole section from further processing. To define the scope of irrelevant phrases, we used the same context identifier as that for negation and uncertainty detection (see below).

Negation / Uncertainty detection

Here, we briefly discuss how uncertainty detection 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* in Chapter 1), and finally linking them with the textual representation of the propositions in question.

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, i.e. each token is either part of an uncertainty cue, or it is not a cue. 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; while it does not hold and is considered false if the cue does not have any impact on the interpretation of the event.

The linking of uncertainty cues and event markers can be established by using dependency grammar rules, i.e. the problem is mainly syntax driven. 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 et al. (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.

The system with the previously described components was able to tag documents with YES labels or leave them as UNMENTIONED. Doing this, we also extracted sentences with disease names from YES-tagged QUESTIONABLE & NO documents and these sentences served as the basis for implementing a simple negation and uncertainty detection module. This exploited a list of negation / uncertainty cues and a list of delimiters (which triggered the end of scope). Hedge, negation cues and delimiter words/punctuation were chosen so as to provide an optimal performance on the training dataset in terms of a macro-averaged F-measure. That is, we selected each word from the extracted sentences that seemed to be a meaningful delimiter or keyword, and discarded those words that lowered performance scores. This approach is similar to NegEx (Chapman et al., 2001). BioScope (Vincze et al., 2008b) also demonstrates that this simple scope resolution approach works well for clinical texts.

The few QUESTIONABLE and NO cases that were not covered this way were again examined manually to extract such terms as ‘normotensive’ for NO-hypertension for example (or were neglected if we found no clear evidence for the QUESTIONABLE and NO label, and also when extending dictionaries to capture the particular instance actually caused an overfitting and errors on other examples).

Intuitive terms

We extended the system with intuitive dictionaries that triggered intuitive YES labels. These dictionaries were used to classify a document as an intuitive YES when it was judged to be UNMENTIONED by the textual classifier system.

MedLine Plus These terms (typically names of associated drugs and medication, etc) were collected from the MedlinePlus encyclopedia and then filtered for intuitive positive class-conditional probability.

C4.5 We also extracted terms like these by training decision trees to discriminate intuitive YES and NO documents using a vector space model representation of the documents. We made the assumption here that complex rules represented by decision trees of depth greater than 1 were unlikely to provide a meaningful classification result so we extracted the words from the nodes of the learnt decision trees and included them as single terms in our dictionary lookup system.

Biomarker expressions

We also added a model that looked for numeric expressions preceding or following certain keywords (that is, biomarker expressions) in the text to classify intuitive YES documents. These were typically for obesity / weight or body mass index/, hypertension /high tension values/, etc...). Thresholds for the numeric expressions were set to provide the optimal performance on the training dataset. For instance: if the phrase ‘ejection fraction’ is found and the associated value is below 50, predict intuitive YES label for congestive heart failure.

8.4 Results

According to the official evaluation, our system was good for an F-macro score of 84% on the training set for our best model (which degraded to 76% on the test set), and an intuitive F-macro score of 82% on the training set (which degraded to 67% on the test set). This system came sixth in the textual F-macro ranking and second in the intuitive F-macro ranking (third best and second best micro-averaged scores, respectively). The micro-averaged results were in the high 90s as the system was especially accurate on the YES and UNMENTIONED classes (YES and NO for intuitive judgment), and these classes had many more examples than the QUESTIONABLE and textual NO classes.

Details on the performance gain of each component, per class results on the test dataset and confusion matrices of our best submission can be found in Tables 1-5.

System	Training set		Test set	
	F_{micro}	F_{macro}	F_{micro}	F_{macro}
Basic system	97.91	83.94	97.29	76.22
Basic system w/o U-dict	97.57	82.07	96.88	73.10
Basic system w/o neg/unc	97.26	51.23	96.81	51.03
Basic system w/o both	96.93	51.03	96.47	50.82

Table 8.1: Textual results.

System	Training set		Test set	
	F_{micro}	F_{macro}	F_{micro}	F_{macro}
Basic system	97.11	82.32	96.42	67.27
Basic system w/o I-terms	96.21	81.57	95.42	66.42
Basic system w/o numexp	96.90	82.15	96.26	67.13
Basic system w/o both	96.00	81.39	95.26	66.28

Table 8.2: Intuitive results.

Disease	Textual		Intuitive	
	F_{micro}	F_{macro}	F_{micro}	F_{macro}
Asthma	98.81	82.47	98.73	97.42
CAD	91.15	85.13	93.89	62.57
CHF	93.15	77.81	93.84	62.83
Depression	98.42	97.16	97.48	96.61
Diabetes	95.43	81.60	96.66	96.03
Gallstones	98.82	79.06	99.19	98.48
GERD	98.41	73.34	91.78	57.80
Gout	99.21	97.93	99.20	98.18
Hypercholesterolemia	96.61	84.95	91.42	91.47
Hypertension	97.21	80.06	96.19	94.61
Hypertriglyceridemia	98.82	78.27	98.97	94.41
OA	96.81	94.23	93.51	59.93
Obesity	96.96	48.83	97.54	97.49
OSA	99.20	65.87	99.60	88.34
PVD	98.42	96.81	96.99	62.81
Venous Insufficiency	99.01	89.75	96.02	78.22

Table 8.3: Per disease results on test set.

	Training set				Test set			
	Y	N	Q	U	Y	N	Q	U
YES	3,072	6	1	129	2,089	7	2	94
NO	11	66	0	10	9	41	1	14
QUESTIONABLE	8	0	22	9	6	0	8	3
UNMENTIONED	46	18	5	8,227	60	16	6	5,688

Table 8.4: Textual confusion matrix.

	Training set			Test set		
	Y	N	Q	Y	N	Q
YES	3,020	243	4	2,096	186	3
NO	47	7,315	0	61	5,037	2
QUESTIONABLE	5	10	11	3	10	1

Table 8.5: Intuitive confusion matrix.

8.5 Discussion

Our intuitive model was based on the textual model. This is why we got a worse performance in intuitive QUESTIONABLE tagging on the test data: we neglected textual UNMENTIONED documents that had an intuitive QUESTIONABLE label because there were too few of them – only 9 examples for the 16 diseases altogether – in the training data to model this phenomenon, especially without background medical knowledge.

The results confirm that our system enhanced with a simple uncertainty detector is able to adequately detect uncertain (questionable) diseases. However, we suffered greatly from the training/test distribution of QUESTIONABLE examples. Out of the 26 intuitive-QUESTIONABLE examples in the training set only 9 got a textual-UNMENTIONED label and 17 were textual-QUESTIONABLE. As regards the 14 intuitive-QUESTIONABLE examples in the test set, only 1 was textual-QUESTIONABLE. This meant that our approach had no real chance to find the majority of the QUESTIONABLE test examples. There were two test documents that had no textual gold standard label (the third annotator was unable to resolve disagreements on the textual label), but it had an intuitive QUESTIONABLE label, which we found odd. Our system achieved the second best result on the previously unseen test set for both the micro- and macro-averaged evaluation (intuitive task). The good micro ranking tells us that the dictionaries we collected had a good coverage compared to other participants, while our second place in macro ranking confirms that predicting intuitive QUESTIONABLE cases also proved rather difficult (or even impossible) for the other participating systems as well.

The model suffered from a lack of coverage for the NO and QUESTIONABLE classes in textual annotation as well (although not as severely as for the intuitive task – the performance dropped from 84% to 76% in the textual task, mainly due to more NO & QUESTIONABLE documents left as UNMENTIONED than in the training set).

We should add here that the main evaluation metric of the challenge was the macro-averaged F-measure. This metric gave special emphasis to the rare NO & QUESTIONABLE classes, which means that a few dozen examples had a major impact on the results. The results presented in Table 1 demonstrate that negation and uncertainty detection had a major impact on macro-averaged evaluation, while on the other hand it had a negligible effect on micro-averaged scores due to the small number of instances that belonged to these classes.

This explains both the worse results on the test set (it was particularly hard to model these infrequent classes), and some seemingly strong drops (e.g. for osteoarthritis) or increases (e.g. for obesity) in performance for particular diseases. Micro-averaged results, which take all document-label pair into account with a uniform weight, are more stable. Moreover, our third place in the micro ranking surely confirms that our disease term dictionaries had a reasonably good coverage (compared to other systems), while our context analyzer overlooked some NO & QUESTIONABLE cases (sixth place in macro ranking).

We suppose that the relatively good results achieved by our model are due to the high-precision term-dictionaries and context-analysing rules. We argue that such simple solutions are efficient whenever the classification depends on the presence or absence of certain single facts (assertions) in the text. In such problems, usually one sentence (in some cases, 2-3) contains the target information. This means that the information can be extracted using a simple approach based on dictionary lookup and modifier detection; and the recognition of complex dependencies in the document is not necessary.

8.6 Summary of Results

In this chapter, we presented a real-world application of uncertainty detection, namely, we introduced our approach to determine the status of the patient concerning obesity and 15 related diseases from medical documents. Our method was enhanced with an uncertainty detector, which had a significant role in labeling questionable cases, thus it proves that an uncertainty detector can be adequately applied in a real-world information extraction task.

The results of this chapter include:

- our automatic system for identifying morbidities in the flow-text parts of clinical discharge summaries;
- we showed how uncertainty detection may enhance information extraction tasks;
- an uncertainty detector integrated into the system;
- the results demonstrate that a simple approach based on dictionary lookup and uncertainty/negation detection may be successfully applied for the task.

These results are described in Farkas et al. (2009). The author's main contributions to the paper were offering linguistics-based rules for uncertainty and negation detection, collecting uncertainty cues typical of the medical domain, determining the linguistic scope of such cues and collating dictionaries of relevant medical terms and morbidity names. The latter is a shared contribution with another co-author and statistical methods for term identification and context detection and the application of biomarkers in the system were the contributions of other co-authors. Again, the final results of the system are considered as a shared contribution of all authors.

Here, we applied a rule-based model of uncertainty detection, which performed well in the clinical domain. This indicates that even simple methods for uncertainty detection may be fruitfully applied in other information extraction or document classification tasks as well, where it is necessary to distinguish factual and non-factual information. In the future, we would like to test the applicability of our machine learning methods described in Chapter 6 for such tasks too.

Chapter 9

Summary

9.1 Summary in English

In this thesis, we aimed at detecting uncertainty in English and Hungarian natural language texts. This research question can be investigated from a dual perspective since it is situated in the field of natural language processing, i.e. in the intersection of linguistics and computer science. Thus, in our investigations, we also made use of linguistic background but the emphasis was put on computer science. As opposed to earlier studies that focused on specific domains and were English-oriented, we offered here a comprehensive approach to uncertainty detection, which can be easily adapted to the specific needs of many domains and languages. In our investigations, we paid attention to create linguistically plausible models of uncertainty that were exploited in the implementation of our uncertainty detectors for several domains, with the help of supervised machine learning techniques.

Hereby we summarize the most important achievements described in the thesis. The first part of the thesis introduced the background of uncertainty detection and the basics of machine learning. In the second part of the thesis, we presented uncertainty phenomena as they occur in language and annotated corpora, whereas in the third part of the thesis, we demonstrated how linguistic uncertainty can be detected in natural language texts by automatic methods.

9.1.1 Contributions of the Thesis

The main results achieved in this thesis will be summarized in the next sections, listed in the order of relevance for computer science.

Detecting Semantic Uncertainty

We carried out experiments on detecting semantic uncertainty in English and Hungarian – for the latter task, to the best of our knowledge, we reported the first published results. We implemented an accurate semantic uncertainty detector that distinguishes four fine-grained categories of semantic uncertainty (epistemic, doxastic, investigation and condition types) and our experiments revealed that shallow features provide good results in recognizing semantic uncertainty for both English and Hungarian. We also applied domain adaptation techniques and achieved successful results for uncertainty detection across various domains and genres in English, and we extended the feature set with semantic and pragmatic features for Hungarian (**Thesis 1**).

Detecting Discourse-level Uncertainty

We implemented systems for detecting three types of uncertainty at the discourse level (weasels, hedges and peacocks). We introduced a baseline method for detecting discourse-level uncertainty in English Wikipedia texts and we applied a supervised machine learning approach to do the same in Hungarian, which was based on sequence labeling and exploited a rich feature set. We achieved reasonable results for both languages (**Thesis 2**).

Uncertainty Detection in the Medical Domain

We presented a real-world application of uncertainty detection: we introduced our approach to determine the status of the patient concerning obesity and 15 related diseases from clinical discharge summaries. Our uncertainty detector had a significant role in labeling questionable cases, which proves that an uncertainty detector can be adequately applied in a real-world information extraction task (**Thesis 3**).

Classification of Uncertainty Phenomena

We offered a language-independent classification of uncertainty phenomena on the basis of theoretical linguistic and computational linguistic background. We paid attention to both semantic and discourse-level uncertainty, we compared the annotation principles of existing corpora annotated for uncertainty and we also provided a unified framework in which all the uncertainty phenomena touched upon in earlier studies can be adequately placed, which served as a base for manually annotating corpora for linguistic uncertainty cues (**Thesis 4**).

Creating Corpora Annotated for Uncertainty

We created several corpora (BioScope, FactBank, WikiWeasel, hUnCertainty) and annotated them for uncertainty cues, based on the above-mentioned unified framework for uncertainty phenomena. We also presented statistical data on cue distribution in the corpora, which revealed the domain- and genre-dependence of uncertainty detection. These corpora were used in our machine learning experiments on uncertainty detection (**Thesis 5**).

Scope-based and Event-based Annotations

We categorized the differences between the linguistic-based and event-oriented annotation of negation and speculation in the intersection of the BioScope 1.0 and Genia Event corpora. We concluded that the scope-oriented annotation system is more adaptable to non-biomedical applications because of the high level of domain specificity in the event-oriented annotation system. We also argued that the strength of uncertainty can manifest at the levels of both semantic and discourse-level uncertainty (**Thesis 6**).

9.1.2 Conclusions and Future Work

In this thesis, we focused on uncertainty detection in natural language texts. On the basis of the main contributions, we can argue that:

- supervised machine learning methods can be successfully applied for uncertainty detection;
- machine learning-based uncertainty detection can be successfully carried out for English as well as for Hungarian;
- uncertainty detectors can enhance the performance of information extraction systems as illustrated by the example of identifying morbidities in the flow-text parts of clinical discharge summaries;
- there are domain specificities of uncertainty cue distribution;
- domain adaptation techniques may help diminish the distance between domains in uncertainty detection;
- linguistic uncertainty can be modeled in a language- and domain-independent way;
- the annotation scheme may determine the field of usage of the corpora, e.g. corpora with event-based annotation are mostly used in biological information extraction.

Besides the main points described above, the results of the thesis may be applicable in other fields of NLP research as well as in other disciplines. Here we just propose some possible application areas where our uncertainty detectors may be employed, without the intention of giving an exhaustive list.

Information extraction applied for the news media may certainly profit from finding weasels, i.e. missing or undeterminable sources. Pieces of information without an identifiable (and reliable) source require special treatment: they will be excluded from the news or they will be communicated to the public in a special form, using phrases such as *according to unnamed sources* etc. Moreover, information extraction in general should also profit from distinguishing certain and uncertain information and IE systems may offer the user these categories separated from each other.

Information retrieval may also be enhanced by detecting uncertainty. Again, it is essential to distinguish documents that contain certain information related to the query from documents with uncertain information, and if the system can make a distinction between the two types of information, the users can later decide whether to make use of only certain information or they want to take into account uncertain pieces of information too.

There is one specific type of texts, namely, patents, where there is a tendency to generalize over the scope of the patent in order to prevent further abuse (Osenga, 2006). Thus, the scope of the patents can be expanded or other use cases can later be included in the patent. For this purpose, patents abound in hedges, hence any NLP system that is developed for or adapted to the linguistic processing of patents must target hedge detection.

Document classification may also profit from detecting uncertainty since different genres of texts involve different types of uncertainty (Hyland, 1998; Falahati, 2006; Rizomilioti, 2006). Thus, the frequency of uncertainty categories may be indicative of the domain of the text as well, which again may be exploited in document classification.

In sentiment analysis and opinion mining, the identification of subjective terms is essential. Although there are some sentiment lexicons such as SentiWordNet (Baccianella et al., 2010), which help find subjective terms in texts, the problem that these terms are often ambiguous between subjective and objective use is still present, hence a subjectivity word sense disambiguation is needed (Wiebe, 2012). With our tools, the detection of peacock terms is viable, so our system may be fruitfully applied in opinion mining as well.

Furthermore, the annotated corpora may be employed in core linguistic research as well. Real-life linguistic data on types of uncertainty can be gathered from our corpora, which may enhance research on semantics, pragmatics or discourse analysis. Finally, our uncertainty detectors may contribute to the improvement of scientific publications. The automatic detection of discourse-level phenomena in scientific writing makes it easier for the authors (and reviewers) to discover undesirable phenomena like sourceless sentences or exaggerations (cf. Day (1998)). In this way, it may effectively contribute to the writing and rewriting of scientific works, which leads to papers of better quality.

As future work, we would like to detect uncertainty in other types of texts (for a pilot study on detecting uncertainty in Hungarian webtext, see Vincze et al. (2014)) as well as in texts written in other languages. For that purpose, we would like to annotate some data in new domains and languages and we would like to extend our tools to those areas as well. Later on, we would like to integrate our uncertainty detectors into some IE or IR applications. We believe that our research on uncertainty detection can be successfully exploited in the solution of several NLP tasks and so, it will contribute to develop novel approaches in many fields of natural language processing.

9.2 Magyar nyelvű összefoglaló

Az értekezésben angol és magyar nyelvű szövegekben azonosítottuk a nyelvi bizonytalanságot. A kutatás tárgya kettős szempontból is vizsgálható, hiszen a számítógépes nyelvészet területébe tartozik, így mind nyelvészeti, mind informatikai vonatkozásai is vannak. A kutatásban elsődlegesen a kérdés informatikai vonatkozásait helyeztük előtérbe, mindemellett nyelvészeti szempontokat is figyelembe vettünk. A korábbi tanulmányokkal ellentétben, melyek pusztán egyes doménekre, illetve elsődlegesen az angol nyelvre koncentráltak, jelen értekezésben egy átfogó, nyelv- és doménfüggetlen megközelítést nyújtottunk a bizonytalanság azonosítására, mely köny-nyen alkalmazható több doménre és nyelvre is. A kutatás első lépéseként felvázoltuk a bizonytalanság egy nyelvi modelljét elméleti és számítógépes nyelvészeti háttérre támaszkodva, melyet a későbbiekben a bizonytalanság automatikus azonosításában alkalmaztunk több doménen és nyelven, felügyelt tanulási módszereket felhasználva.

Az alábbiakban összegezzük az értekezés legfontosabb eredményeit, az értekezés szerkezetét követve. Az értekezés első részében bemutattuk a bizonytalanság azonosításának alapjait és az értekezés módszertana szempontjából legfontosabb gépi tanuló eljárásokat. Az értekezés második részében a természetes nyelvekben előforduló bizonytalansággal kapcsolatos nyelvi jelenségeket tekintettük át és foglaltuk rendszerbe, illetve ismertettük az általunk létrehozott, bizonytalanságra annotált korpuszokat. Az értekezés harmadik részében végül megmutattuk, hogy a nyelvi bizonytalanság egyes típusai hogyan azonosíthatók automatikus módszerekkel különféle természetes nyelvű szövegekben.

9.2.1 Az értekezés eredményei

Az értekezésben elért főbb eredmények az alábbiakban foglalhatók össze, az informatikai szempontól lényeges eredmények kiemelésével.

A szemantikai bizonytalanság azonosítása

Kísérleteket végeztünk a szemantikai bizonytalanság azonosítására magyar és angol nyelvű szövegekben, tudomásunk szerint a magyar nyelvre ezek az első publikált eredmények a témában. Rendszerünk a szemantikai bizonytalanság négy osztályának (episztemikus, doxasztikus, vizsgálat és feltételes) azonosítására képes. Eredményeink azt igazolják, hogy már egyszerű jellemzők használatával is jó eredményeket lehetséges elérni a szemantikai bizonytalanság azonosításában mind angol, mind magyar nyelvre. Doménadaptációs technikák használatával szintén sikeres eredményeket kaptunk angol nyelvre a doméneken és műfajokon átívelő bizonytalanságazonosításban. Jellemzőkészletünket pedig a magyar nyelv esetében szemantikai és pragmatikai jellemzőkkel is bővítettük (**1. tézispont**).

A diskurzusszintű bizonytalanság azonosítása

A diskurzusszintű bizonytalanság három típusának (weasel, hedge és peacock) automatikus azonosítására szintén kidolgoztunk egy rendszert. Alapmegoldást nyújtottunk az angol Wikipedia-szövegekben rejlő diskurzusszintű bizonytalanság automatikus azonosítására, illetve a magyar nyelvű szövegek esetében szekvenciajelölésre épülő, gazdag jellemzőtérrel rendelkező felügyelt tanulási módszert alkalmaztunk. Megoldásaink jó eredményt nyújtottak mindkét nyelv esetében (**2. tézispont**).

Bizonytalanság azonosítása orvosi szövegekben

Bemutattuk a bizonytalanság azonosításának egy gyakorlati példáját is: klinikai zárójelentések szövege alapján következtettünk arra, hogy a beteg elhízott-e, illetve szenvede 15 másik betegség valamelyikében. A nem egyértelmű esetek felcímkézésében jelentős szerep jutott a bizonytalanságazonosító rendszerünknek, ami igazolja, hogy egy valós életbeli információkinyerési feladatban is sikeresen alkalmazható a bizonytalanságazonosító rendszerünk (**3. tézispont**).

A bizonytalanság típusainak kategorizálása

A nyelvi bizonytalanság különféle típusainak kategorizálására létrehoztunk egy elméleti és számítógépes alapokon nyugvó egységes, nyelvfüggetlen osztályozást. Mind a szemantikai, mind a diskurzusszintű bizonytalanság típusait besoroltuk a rendszerbe, összehasonlítottuk a korábban létrehozott, bizonytalanságra annotált korpuszok irányelveit, majd beillesztettük a korábbi (számítógépes) nyelvészeti tanulmányokban vizsgált nyelvi jelenségeket az általunk definiált keretrendszerbe. E keretrendszer képezi az általunk létrehozott, bizonytalanságra annotált korpuszok elméleti hátterét (**4. tézispont**).

Bizonytalanságra annotált korpuszok létrehozása

Számos korpuszt hoztunk létre (BioScope, FactBank, WikiWeasel, hUnCertainty), melyekben kézzel megjelöltük a bizonytalanságot jelző kulcsszavakat, a fenti egységes osztályozásra alapozva. Ezeket a korpuszokat használtuk a későbbiekben a bizonytalanság automatikus azonosítására irányuló gépi tanuló kísérleteinkben. A kulcsszavak eloszlását statisztikai módszerekkel is megvizsgáltuk a korpuszok alapján, ami során kiderült, hogy az egyes domének és műfajok sajátosságokat mutatnak a bizonytalanságot jelölő kulcsszavak eloszlása terén (**5. tézispont**).

Hatókör alapú és eseményalapú annotációk

Feltérképeztük és osztályokba soroltuk a nyelvi hatókörökön, illetve az eseményeken alapuló, tagadásra és bizonytalanságra irányuló annotációk közti jellegzetes különbségeket a BioScope 1.0 és a Genia Event korpuszok közös halmazának összevetésével. Eredményeink szerint a hatókör alapú annotáció hatékonyabbnak bizonyul a biológiától eltérő területekre fejlesztett alkalmazások esetében, mivel az eseményalapú annotációs rendszer nagymértékben épít a biológiai domén sajátságaira. Megmutattuk azt is, hogy a szemantikai és a diskurzusszintű bizonytalanság szintjén egyaránt megjelenik a bizonytalanság fokozatosságának kérdése (**6. tézispont**).

9.2.2 Összegzés és jövőbeli tervek

Az értekezés fő célja a nyelvi bizonytalanság azonosítása volt természetes nyelvi szövegekben. A legfontosabb eredményeink a következőkben összegezhetők:

- felügyelt tanulási módszerek jól alkalmazhatók a bizonytalanság azonosítására;
- a gépi tanuláson alapuló bizonytalanságazonosító módszerek angolban és magyarban is jól működnek;
- a bizonytalanságazonosító rendszerek képesek javítani az információkinyerő rendszerek hatékonyságán;
- a bizonytalanságot jelölő kulcsszavak doménfüggő eloszlást mutatnak;
- doménadaptációs technikák segítségével csökkenthető a domének közti távolság a bizonytalanság azonosításában;
- a nyelvi bizonytalanság modellezhető nyelv- és doménfüggetlen módon;
- az annotációs elvek meghatározhatják a korpuszok hasznosíthatóságát, például az eseményalapú annotációt tartalmazó korpuszokat leginkább biológiai információkinyerésben alkalmazzák.

A fentieken kívül az értekezés eredményeit a számítógépes nyelvészet más területein is, továbbá más tudományterületeken is lehet hasznosítani. Az alábbiakban a bizonytalanság azonosításának néhány alkalmazási lehetőségét vázoljuk fel, a teljességre való törekvés nélkül.

A médiában szereplő hírekből történő információkinyerés esetében fontos lehet a weasel kifejezések megtalálása, azaz a hiányzó vagy meghatározatlan forrással rendelkező hírek azonosítása. Az azonosítatlan (vagy nem megbízható) forrással rendelkező hírek vagy egyáltalán nem kerülnek bele a hírműsorba, vagy pedig sajátos formában illesztik be őket, bizonyos tipikus kifejezéseket használva (pl. *meg nem nevezett források szerint*). Mindemellett az általános információkinyerés is profitálhat a biztos és bizonytalan információ elkülönítéséből, így az információkinyerő rendszerek ezeket külön kategóriába sorolhatják be a felhasználók számára.

A bizonytalanság azonosítása az információ-visszakeresést is támogathatja. Ez esetben is létfontosságú elkülöníteni a kereséshez kapcsolódó biztos információt tartalmazó dokumentumokat a bizonytalan információt tartalmazó dokumentumoktól. Amennyiben a rendszer képes erre a megkülönböztetésre, a felhasználó a keresést követően eldöntheti, hogy figyelembe veszi-e a bizonytalan információt tartalmazó dokumentumokat is, vagy kizárólag a biztos információt tartalmazókra koncentrál.

A szabadalmak előszeretettel tartalmaznak általánosításokat annak érdekében, hogy a szabadalom a lehető legnagyobb alkalmazási területet fedje le, megelőzendő az esetleges későbbi visszaéléseket (Osenga, 2006). A szabadalmak alkalmazási köre így kiterjeszthető marad, illetve későbbi használati esetek bekerülhetnek a szabadalom hatókörébe. Mindennek köszönhetően a szabadalmakban rendkívül sok hedge szerepel, így a szabadalmak automatikus feldolgozását célzó rendszerek számára nagy szereppel bír a hedgeek automatikus azonosítása.

A bizonytalanság azonosítása a dokumentumosztályozásban is hasznosítható, hiszen az eltérő műfajba vagy doménbe tartozó szövegekben más-más típusú bizonytalanság számít gyakorinak (Hyland, 1998; Falahati, 2006; Rizomilioti, 2006). Így a bizonytalanság típusainak relatív gyakorisága is megbízhatóan jelezheti a szöveg doménjét vagy műfaját, ami szerepet játszhat a dokumentumok osztályba sorolásában.

Lényegi szereppel bír a véleménykinyerésben a szubjektív kifejezések azonosítása. Noha rendelkezésre állnak a SentiWordNet (Baccianella et al., 2010) és más hasonló lexikonok, melyek a szövegekben előforduló szubjektív kifejezések megtalálásában jelentenek segítséget, e kifejezések igen gyakran többértelműek, azaz kontextustól függően lehetnek szubjektívek vagy sem. A szubjektívnek tűnő kifejezések egyértelműsítésében (Wiebe, 2012) szerepe lehet a peacock kifejezéseket azonosító módszerünknek, azaz a bizonytalanság azonosítása a véleménykinyerést is támogatni tudja.

A fentieken kívül a nyelvészeti kutatások is hasznot húzhatnak az annotált korpuszokból. A korpuszok valós nyelvhasználati példák tárházául szolgálnak, így a bizonytalanság nyelvi adatai könnyen kigyűjthetők belőlük, támogatva ezzel a szemantikai, pragmatikai vagy diskurzuselemzési kutatásokat. A bizonytalanság azonosítása a tudományos publikációk írását is segítheti: a diskurzusszintű bizonytalanság automatikus azonosítása során meg lehet találni a szövegben az olyan nemkívánatos jelenségeket mint forrás nélküli mondatok, a pontatlan kifejezések vagy a túlzások (vö. Day (1998)). Mindez megkönnyítheti a tudományos cikkek szerzőinek (és bírálóinak) munkáját, támogatva a cikket írását és átírását, ezáltal jobb minőségű publikációkat eredményezve.

A jövőben más nyelvű, illetve más típusú szövegekben is szeretnénk azonosítani a bizonytalanságot jelző kulcsszavakat. E célból más nyelvű és más doménbe tartozó szövegekből is szeretnénk annotált korpuszt építeni (lásd Vincze et al. (2014) a magyar webes szövegekben fellelhető bizonytalanság azonosításáról), továbbá automatikus módszereinket is szeretnénk az új korpuszokra kiterjeszteni. Tervezzük, hogy a későbbiekben a bizonytalanságazonosító rendszerünket beépítjük különféle információkinyerő, illetve információ-visszakereső alkalmazásokba. Véleményünk szerint az értekezésben ismertetett módszerek jól hasznosíthatók számos számítógépes nyelvészeti feladat megoldásában, valamint újfajta megközelítések kidolgozásához és eddig még feltáratlan alkalmazási területek felfedezéséhez is hozzájárulhatnak.

References

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

Allison, D.B.; Fontaine, K.R.; Manson, J.E.; Stevens, J.; VanItallie, T.B. 1999. Annual deaths attributable to obesity in the United States. *Journal of the American Medical Association*, 282(16):1530–1538.

Alpaydin, Ethem. 2010. *Introduction to Machine Learning*. The MIT Press, 2nd edition.

Ananiadou, Sophia; Mcnaught, John. 2005. *Text Mining for Biology And Biomedicine*. Artech House, Inc., Norwood, MA, USA.

Baccianella, Stefano; Esuli, Andrea; Sebastiani, Fabrizio. 2010. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In Calzolari, Nicoletta; Choukri, Khalid; Maegaard, Bente; Mariani, Joseph; Odijk, Jan; Piperidis, Stelios; Rosner, Mike; Tapias, Daniel (eds.), *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May. European Language Resources Association (ELRA).

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

Barness, Lewis A.; Opitz, John M.; Gilbert-Barness, Enid. 2007. Obesity: genetic, molecular, and environmental aspects. *American Journal of Medical Genetics Part A*, 143A(24):3016–3034.

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

Berger, Adam L.; Pietra, Stephen Della; Pietra, Vincent J. Della. 1996. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–71.

Brown, Penelope; Levinson, Stephen C. 1987. *Politeness: Some universals in language usage*. Cambridge University Press, Cambridge, UK.

Brown, Gillian; Yule, George. 1983. *Discourse Analysis*. Cambridge University Press, Cambridge, UK.

Califf, Mary Elaine. 2008. Combining Rules and Naïve Bayes for Disease Classification. In *Proceedings of the i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data*.

Chapman, Wendy W.; Bridewell, Will; Hanbury, Paul; Cooper, Gregory F.; Buchanan, Bruce G. 2001. A simple algorithm for identifying negated findings and diseases in discharge summaries. *Journal of Biomedical Informatics*, 5:301–310.

Chapman, Wendy W.; Chu, David; Dowling, John N. 2007. Context: An algorithm for identifying contextual features from clinical text. In *Proceedings of the ACL Workshop on BioNLP 2007*, pp. 81–88.

Childs, Lois C.; Taylor, Robert J.; Simonsen, Lone; Heintzelman, Norris H.; Kowalski, Kimberly M.; Enelow, Robert. 2008. Description of the Lockheed Martin / SAGE Analytica System for the i2b2 Challenge in Natural Language Processing for Clinical Data. In *Proceedings of the i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data*.

Clausen, David. 2010. HedgeHunter: a system for hedge detection and uncertainty classification. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning – Shared Task*, CoNLL '10: Shared Task, pp. 120–125, Uppsala, Sweden. Association for Computational Linguistics.

Conway, Mike; Doan, Son; Collier, Nigel. 2009. Using hedges to enhance a disease outbreak report text mining system. In *Proceedings of the BioNLP 2009 Workshop*, pp. 142–143, Boulder, Colorado, June. Association for Computational Linguistics.

Councill, Isaac; McDonald, Ryan; Velikovich, Leonid. 2010. What's great and what's not: learning to classify the scope of negation for improved sentiment analysis. In *Proceedings of the Workshop on Negation and Speculation in Natural Language Processing*, pp. 51–59, Uppsala, Sweden, July.

Cristea, Dan; Webber, Bonnie. 1997. Expectations in incremental discourse processing. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL97/EACL97)*, pp. 88–95. Morgan Kaufmann.

Cruz Díaz, Noa P. 2013. Detecting negated and uncertain information in biomedical and review texts. In *Proceedings of the Student Research Workshop associated with RANLP 2013*, pp. 45–50, Hissar, Bulgaria, September. RANLP 2013 Organising Committee.

Curran, James; Clark, Stephen; Bos, Johan. 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*, pp. 33–36, Prague, Czech Republic, June. Association for Computational Linguistics.

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

Day, Robert A. 1998. *How to write and publish a scientific paper.* Oryx Press, Phoenix.

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

Diab, Mona; Levin, Lori; Mitamura, Teruko; Rambow, Owen; Prabhakaran, Vinodkumar; Guo, Weiwei. 2009. Committed belief annotation and tagging. In *Proceedings of the Third Linguistic Annotation Workshop*, pp. 68–73, Suntec, Singapore, August. Association for Computational Linguistics.

Falahati, Reza. 2006. The use of hedging across different disciplines and rhetorical sections of research articles. In Carter, Nicole; Hadic-Zabala, Loreley; Rimrott, Anne; Storoshenko, Dennis Ryan (eds.), *Proceedings of the 22nd NorthWest Linguistics Conference (NWLC22)*, pp. 99–112, Burnaby, Canada. Simon Fraser University.

Farkas, Richárd; Szarvas, György. 2008. Automatic construction of rule-based ICD-9-CM coding systems. *BMC Bioinformatics*, 9:1–9. 10.1186/1471-2105-9-S3-S10.

Farkas, Richárd; Szarvas, György; Hegedűs, István; Almási, Attila; Vincze, Veronika; Ormándi, Róbert; Busa-Fekete, Róbert. 2009. Semi-automated construction of decision rules to predict morbidities from clinical texts. *Journal of the American Medical Informatics Association*, 16:601–605.

Farkas, Richárd; Vincze, Veronika; Móra, György; Csirik, János; Szarvas, György. 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*, pp. 1–12, Uppsala, Sweden, July. Association for Computational Linguistics.

Fernandes, Eraldo R.; Crestana, Carlos E. M.; Milidiú, Ruy L. 2010. Hedge detection using the RelHunter approach. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning – Shared Task*, CoNLL '10: Shared Task, pp. 64–69, Uppsala, Sweden. Association for Computational Linguistics.

Friedman, Carol; Alderson, Philip O.; Austin, John H. M.; Cimino, James J.; Johnson, Stephen B. 1994. A General Natural-language Text Processor for Clinical Radiology. *Journal of the American Medical Informatics Association*, 1(2):161–174.

Ganter, Viola; Strube, Michael. 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*, pp. 173–176, Suntec, Singapore, August. Association for Computational Linguistics.

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*, pp. 26–31, Uppsala, Sweden, July. Association for Computational Linguistics.

Grice, H. Paul. 1975. Logic and conversation. In Cole, P.; Morgan, J. (eds.), *Syntax and semantics, vol 3.*, New York. Academic Press.

Hazlehurst, Brian; Frost, H. Robert; Sittig, Dean F.; Stevens, Victor J. 2005. Application of information technology: Mediclass: A system for detecting and classifying encounter-based clinical events in any electronic medical record. *JAMIA*, 12(5):517–529.

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.

Katsos, Napoleon; Breheny, Richard. 2010. Two experiments and some suggestions on the meaning of scalars and numerals. In Németh T., Enikő; Bibok, Károly (eds.), *The role of data at the semantics-pragmatics interface*, Berlin, New York. De Gruyter Mouton.

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

Kilicoglu, Halil; Bergler, Sabine. 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*, pp. 46–53, Columbus, Ohio, June. Association for Computational Linguistics.

Kilicoglu, Halil; Bergler, Sabine. 2009. Syntactic dependency based heuristics for biological event extraction. In *Proceedings of the BioNLP 2009 Workshop Companion Volume for Shared Task*, pp. 119–127, Boulder, Colorado, June. Association for Computational Linguistics.

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

Kim, Jin-Dong; Ohta, Tomoko; Pyysalo, Sampo; Kano, Yoshinobu; Tsujii, Jun'ichi. 2009. Overview of BioNLP'09 Shared Task on Event Extraction. In *Proceedings of the BioNLP 2009 Workshop Companion Volume for Shared Task*, pp. 1–9, Boulder, Colorado, June. Association for Computational Linguistics.

Konstantinova, Natalia; de Sousa, Sheila C.M.; Cruz, Noa P.; Mana, Manuel J.; Taboada, Maite; Mitkov, Ruslan. 2012. A review corpus annotated for negation, speculation and their scope. In Calzolari, Nicoletta; Choukri, Khalid; Declerck, Thierry; Dogan, Mehmet Ugur; Maegaard, Bente; Mariani, Joseph; Odijk, Jan; Piperidis, Stelios (eds.), *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, May. European Language Resources Association (ELRA).

Krallinger, Martin. 2010. Importance of negations and experimental qualifiers in biomedical literature. In *Proceedings of the Workshop on Negation and Speculation in Natural Language Processing*, pp. 46–49, Uppsala, Sweden, July.

Lafferty, John; McCallum, Andrew; Pereira, Fernando. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of ICML-01, 18th Int. Conf. on Machine Learning*, pp. 282–289. Morgan Kaufmann.

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

Li, Xinxin; Shen, Jianping; Gao, Xiang; Wang, Xuan. 2010. Exploiting rich features for detecting hedges and their scope. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning – Shared Task*, CoNLL '10: Shared Task, pp. 78–83, Uppsala, Sweden. Association for Computational Linguistics.

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

Liu, Bing. 2012. *Sentiment Analysis and Opinion Mining*. Morgan & Claypool Publishers.

MacKinlay, Andrew; Martinez, David; Baldwin, Timothy. 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, pp. 77–85, Uppsala, Sweden. Association for Computational Linguistics.

Matthews, Michael P. 2008. Bayesian Networks and the i2b2 Obesity Challenge. In *Proceedings of the i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data*.

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

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

Mokdad, A.H.; Marks, J.S.; Stroup, D.F.; Gerberding, J.L. 2004. Actual causes of death in the United States, 2000. *Journal of the American Medical Association*, 291(10):1238–1245.

Morante, Roser; Daelemans, Walter. 2009. Learning the scope of hedge cues in biomedical texts. In *Proceedings of the BioNLP 2009 Workshop*, pp. 28–36, Boulder, Colorado, June. Association for Computational Linguistics.

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

Morante, Roser; Sporleder, Caroline. 2012. Modality and negation: An introduction to the special issue. *Computational Linguistics*, 38:223–260, June.

Morante, Roser; van Asch, Vincent; van den Bosch, Antal. 2009. Joint memory-based learning of syntactic and semantic dependencies in multiple languages. In *Proceedings of CoNLL*, pp. 25–30.

Morante, Roser; Van Asch, Vincent; Daelemans, Walter. 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*, pp. 40–47, Uppsala, Sweden, July. Association for Computational Linguistics.

Nagy T., István; Vincze, Veronika; Berend, Gábor. 2011. Domain-dependent identification of multiword expressions. In *Proceedings of RANLP 2011*, Hissar, Bulgaria.

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

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

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

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

Pakhomov, Serguei V.; Buntrock, James D.; Chute, Christopher G. 2006. Automating the assignment of diagnosis codes to patient encounters using example-based and machine learning techniques. *Journal of the American Medical Informatics Association: JAMIA*, 13(5):516–525.

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

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

Peshkin, Leonid; Cano, Carlos; Carpenter, Bob; Baldwin, Breck. 2008. Regularized Logistic Regression for Clinical Record Processing. In *Proceedings of the i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data*.

Pestian, John P.; Brew, Christopher; Matykiewicz, Paweł; Hovermale, D. J.; Johnson, Neil; Cohen, K. Bretonnel; Duch, Włodzisław. 2007. A shared task involving multi-label classification of clinical free text. In *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, BioNLP '07, pp. 97–104, Stroudsburg, PA, USA. Association for Computational Linguistics.

Rei, Marek; Briscoe, Ted. 2010. Combining manual rules and supervised learning for hedge cue and scope detection. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning – Shared Task*, CoNLL '10: Shared Task, pp. 56–63, Uppsala, Sweden. Association for Computational Linguistics.

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

Rubin, Victoria L.; Liddy, Elizabeth D.; Kando, Noriko. 2005. Certainty identification in texts: Categorization model and manual tagging results. In Shanahan, J.G.; Qu, J.; Wiebe, J. (eds.), *Computing attitude and affect in text: Theory and applications (the information retrieval series)*, New York. Springer Verlag.

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.

Russell, Stuart J.; Norvig, Peter. 2010. *Artificial Intelligence - A Modern Approach (3. internat. ed.).* Pearson Education.

Sánchez, Liliana Mamani; Li, Baoli; Vogel, Carl. 2010. Exploiting CCG structures with tree kernels for speculation detection. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning – Shared Task*, CoNLL '10: Shared Task, pp. 126–131, Uppsala, Sweden. Association for Computational Linguistics.

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

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

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

Savova, Guergana; Clark, Cheryl; Zheng, Jiaping; Cohen, K. Bretonnel; Murphy, Sean; Wellner, Ben; Harris, David; Lazo, Marcia; Aberdeen, John; Hu, Qian; Chute, Christopher; Hirschman, Lynette. 2008. The Mayo/MITRE System for Discovery of Obesity and Its Comorbidities. In *Proceedings of the i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data.*

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

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

Solt, Illés; Tikk, Domonkos; Gál, Viktor; Kardkovács, Zsolt Tivadar. 2008. Context-Aware Rule Based Classifier for Semantic Classification of Diseases in Discharge Summaries. In *Proceedings of the i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data.*

Sun, Chengjie; Lin, Lei; Wang, Xiaolong; Guan, Yi. 2007. Using maximum entropy model to extract Protein-Protein interaction information from biomedical literature. In *Advanced Intelligent Computing Theories and Applications. With Aspects of Theoretical and Methodological Issues*, pp. 730–737.

Swan, Michael. 1995. *Practical English Usage.* Oxford University Press, Oxford.

Szarvas, György; Iván, Szilárd; Bánhalmi, András; Csirik, János. 2006. Automatic Extraction of Semantic Content from Medical Discharge Records. *WSEAS Transaction on Systems and Control*, 1(2):312–317.

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

Szarvas, György. 2008. Hedge classification in biomedical texts with a weakly supervised selection of keywords. In *Proceedings of ACL-08: HLT*, pp. 281–289, Columbus, Ohio, June. Association for Computational Linguistics.

Taboada, Maite; Anthony, Caroline; Voll, Kimberly. 2006. Methods for creating semantic orientation dictionaries. In *Conference on Language Resources and Evaluation (LREC)*, pp. 427–432.

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

Tang, Buzhou; Wang, Xiaolong; Wang, Xuan; Yuan, Bo; Fan, Shixi. 2010. A cascade method for detecting hedges and their scope in natural language text. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning – Shared Task*, CoNLL '10: Shared Task, pp. 13–17, Uppsala, Sweden. Association for Computational Linguistics.

Thompson, Paul; Venturi, Giulia; Mcnaught, John; Montemagni, Simonetta; Ananiadou, Sophia. 2008. Categorising Modality in Biomedical Texts. In *Proceedings of the LREC 2008 Workshop on Building and Evaluating Resources for Biomedical Text Mining*, pp. 27–34.

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*, pp. 148–150, Uppsala, Sweden, July. Association for Computational Linguistics.

Uzuner, Özlem; Goldstein, Ira; Luo, Yuan; Kohane, Isaac. 2008. Identifying patient smoking status from medical discharge records. *Journal of the American Medical Informatics Association: JAMIA*, 15(1):14–24, January.

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

Uzuner, Özlem. 2009. Recognizing Obesity and Comorbidities in Sparse Data. *Journal of the American Medical Informatics Association*, 16(4):561–570, July.

Van Asch, Vincent; Daelemans, Walter. 2010. Using domain similarity for performance estimation. In *Proceedings of the 2010 Workshop on Domain Adaptation for Natural Language Processing*, pp. 31–36, Uppsala, Sweden, July. Association for Computational Linguistics.

Van Landeghem, Sofie; Saeys, Yvan; De Baets, Bernard; Van de Peer, Yves. 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*, pp. 128–136, Boulder, Colorado, June. Association for Computational Linguistics.

Velldal, Erik; Øvrelid, Lilja; Oepen, Stephan. 2010. Resolving speculation: Maxent cue classification and dependency-based scope rules. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning – Shared Task*, CoNLL '10: Shared Task, pp. 48–55, Uppsala, Sweden. Association for Computational Linguistics.

Velldal, Erik; Øvrelid, Lilja; Read, Jonathon; Oepen, Stephan. 2012. Speculation and negation: Rules, rankers, and the role of syntax. *Computational Linguistics*, 38:369–410, June.

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

Vincze, Veronika; Csirik, János. 2010. Hungarian corpus of light verb constructions. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pp. 1110–1118, Beijing, China, August. Coling 2010 Organizing Committee.

Vincze, Veronika; Szarvas, György; Almási, Attila; Szauter, Dóra; Ormándi, Róbert; Farkas, Richárd; Hatvani, Csaba; Csirik, János. 2008a. Hungarian word-sense disambiguated corpus. In Calzolari, Nicoletta; Choukri, Khalid; Maegaard, Bente; Mariani, Joseph; Odijk, Jan; Piperidis, Stelios; Tapias, Daniel (eds.), *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, May. European Language Resources Association (ELRA).

Vincze, Veronika; Szarvas, György; Farkas, Richárd; Móra, György; Csirik, János. 2008b. The BioScope Corpus: Biomedical Texts Annotated for Uncertainty, Negation and their Scopes. *BMC Bioinformatics*, 9(Suppl 11):S9.

Vincze, Veronika; Felvégi, Zsuzsanna; R. Tóth, Krisztina. 2010. Félig kompozicionális szerkezetek a SzegedParalell angol–magyar párhuzamos korpuszban [Semi-compositional constructions in the SzegedParalell English–Hungarian parallel corpus]. In Tanács, Attila; Vincze, Veronika (eds.), *MSzNy 2010 – VII. Magyar Számítógépes Nyelvészeti Konferencia*, pp. 91–101, Szeged, Hungary, December. University of Szeged.

Vincze, Veronika; Nagy T., István; Berend, Gábor. 2011a. Detecting Noun Compounds and Light Verb Constructions: a Contrastive Study. In *Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World*, pp. 116–121, Portland, Oregon, USA, June. ACL.

Vincze, Veronika; Nagy T., István; Berend, Gábor. 2011b. Multiword expressions and named entities in the Wiki50 corpus. In *Proceedings of RANLP 2011*, Hissar, Bulgaria.

Vincze, Veronika; Szarvas, György; Móra, György; Ohta, Tomoko; Farkas, Richárd. 2011c. Linguistic scope-based and biological event-based speculation and negation annotations in the BioScope and Genia Event corpora. *Journal of Biomedical Semantics*, 2(Suppl 5):S8.

Vincze, Veronika; Simkó, Katalin Ilona; Varga, Viktor. 2014. Annotating Uncertainty in Hungarian Webtext. In *Proceedings of LAW VIII*.

Vincze, Veronika. 2007. A félig kompozicionális szerkezetek gépi fordításainak lehetőségéről [On possible ways of automatically translating semi-compositional constructions]. In Váradi, Tamás (ed.), *I. Alkalmazott Nyelvészeti Doktorandusz Konferencia*, pp. 207–218, Budapest. MTA Nyelvtudományi Intézet.

Vincze, Veronika. 2008. A puszta köznév + ige komplexumok státusáról [On the status of bare common noun + verb constructions]. In Sinkovics, Balázs (ed.), *LingDok 7. Nyelvész-doktoranduszok dolgozatai*, pp. 279–297, Szeged, Hungary. University of Szeged.

Vincze, Veronika. 2009a. Angol–magyar főnév + ige szerkezetek és igei párjaik [English–Hungarian noun + verb constructions and their verbal counterparts]. In Váradi, Tamás (ed.), *II. Alkalmazott Nyelvészeti Doktorandusz Konferencia*, pp. 112–122, Budapest. MTA Nyelvtudományi Intézet.

Vincze, Veronika. 2009b. *Előadást tart* vs. *előad*: főnév + ige szerkezetek igei variánsai [*To give a lecture* vs. *to lecture*: verbal counterparts of noun + verb constructions]. In Sinkovics, Balázs (ed.), *LingDok 8. Nyelvész-doktoranduszok dolgozatai*, pp. 265–278, Szeged. Szegedi Tudományegyetem.

Vincze, Veronika. 2009c. Félig kompozicionális szerkezetek a Szeged Korpuszban [Semi-compositional constructions in the Szeged Corpus]. In Tanács, Attila; Szauter, Dóra; Vincze, Veronika (eds.), *VI. Magyar Számítógépes Nyelvészeti Konferencia*, pp. 390–393, Szeged. Szegedi Tudományegyetem.

Vincze, Veronika. 2009d. Főnév + ige szerkezetek a szótárban [Noun + verb constructions in the dictionary]. In Váradi, Tamás (ed.), *III. Alkalmazott Nyelvészeti Doktorandusz Konferencia*, pp. 180–188, Budapest. MTA Nyelvtudományi Intézet.

Vincze, Veronika. 2009e. On the Machine Translatability of Semi-Compositional Constructions. In Váradi, Tamás (ed.), *Válogatás az I. Alkalmazott Nyelvészeti Doktorandusz Konferencia előadásaiból / Selected Papers from the First Applied Linguistics PhD Conference*, pp. 166–178, Budapest. MTA Nyelvtudományi Intézet.

Vincze, Veronika. 2010a. Félig kompozicionális főnév + ige szerkezetek a számítógépes nyelvészetben [Semi-compositional noun + verb constructions in natural language processing]. In Gecső, Tamás; Sárdi, Csilla (eds.), *Új módszerek az alkalmazott nyelvészeti kutatásban*, pp. 327–332, Budapest. Tinta Könyvkiadó.

Vincze, Veronika. 2010b. Speculation and negation annotation in natural language texts: what the case of BioScope might (not) reveal. In *Proceedings of the Workshop on Negation and Speculation in Natural Language Processing*, pp. 28–31, Uppsala, Sweden, July. University of Antwerp.

Vincze, Veronika. 2011. Mi fán terem a főnév + ige szerkezet? [From what tree can you harvest noun + verb constructions?]. In Gécseg, Zsuzsanna (ed.), *LingDok 10. Nyelvész-doktoranduszok dolgozatai*, pp. 225–243, Szeged, Hungary. University of Szeged.

Vincze, Veronika. 2013. Weasels, Hedges and Peacocks: Discourse-level Uncertainty in Wikipedia Articles. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pp. 383–391, Nagoya, Japan, October. Asian Federation of Natural Language Processing.

Vincze, Veronika. 2014. Uncertainty detection in Hungarian texts. In *Proceedings of Coling 2014*.

Wei, Zhongyu; Chen, Junwen; Gao, Wei; Li, Binyang; Zhou, Lanjun; He, Yulan; Wong, Kam-Fai. 2013. An empirical study on uncertainty identification in social media context. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 58–62, Sofia, Bulgaria, August. Association for Computational Linguistics.

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

Wiebe, Janyce. 2012. Subjectivity word sense disambiguation. In *Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis*, p. 2, Jeju, Korea, July. Association for Computational Linguistics.

Wilcox, Adam B.; Hripcsak, George. 2003. The role of domain knowledge in automating medical text report classification. *Journal of the American Medical Informatics Association: JAMIA*, 10(4):330–338.

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, Pittsburgh.

Yang, Hui; Spasic, Irena; Keane, John A.; Nenadic, Goran. 2008. Combining Lexical Profiling, Rules and Machine Learning for Disease Prediction from Hospital Discharge Summaries. In *Proceedings of the i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data.*

Zhang, Shaodian; Zhao, Hai; Zhou, Guodong; Lu, Bao-Liang. 2010. Hedge detection and scope finding by sequence labeling with normalized feature selection. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning – Shared Task*, CoNLL '10: Shared Task, pp. 92–99, Uppsala, Sweden. Association for Computational Linguistics.

Zsibrita, János; Vincze, Veronika; Farkas, Richárd. 2013. magyarlanc: A Toolkit for Morphological and Dependency Parsing of Hungarian. In *Proceedings of RANLP-2013*, pp. 763–771, Hissar, Bulgaria.