

# Uncertainty Detection in Hungarian Texts

Veronika Vincze<sup>1,2</sup>

<sup>1</sup>University of Szeged

Department of Informatics

<sup>2</sup>MTA-SZTE Research Group on Artificial Intelligence

vinczev@inf.u-szeged.hu

## Abstract

Uncertainty detection is essential for many NLP applications. For instance, in information retrieval, it is of primary importance to distinguish among factual, negated and uncertain information. Current research on uncertainty detection has mostly focused on the English language, in contrast, here we present the first machine learning algorithm that aims at identifying linguistic markers of uncertainty in Hungarian texts from two domains: Wikipedia and news media. The system is based on sequence labeling and makes use of a rich feature set including orthographic, lexical, morphological, syntactic and semantic features as well. Having access to annotated data from two domains, we also focus on the domain specificities of uncertainty detection by comparing results obtained in indomain and cross-domain settings. Our results show that the domain of the text has significant influence on uncertainty detection.

## 1 Introduction

Uncertainty detection has become one of the most intensively studied problems of natural language processing (NLP) in these days (Morante and Sporleder, 2012). For several NLP applications, it is essential to distinguish between factual and nonfactual, i.e. negated or uncertain information: for instance, in medical information retrieval, it must be known whether the patient definitely suffers, probably suffers or does not suffer from an illness. This type of information can only be revealed from the texts of the documents if reliable uncertainty detectors are available, which are able to identify linguistic markers of uncertainty, i.e. cues within the text. To the best of our knowledge, uncertainty detectors have been mostly developed for the English language (Morante and Sporleder, 2012; Farkas et al., 2010). Here, we present our machine learning based uncertainty detector developed for Hungarian, a morphologically rich language, and report our results on 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.

The main contributions of this paper are the following:

- it presents the first uncertainty corpus for Hungarian;
- it reports the first results on uncertainty detection in Hungarian texts;
- it introduces new features in the machine learning setting like semantic and pragmatic features;
- we show that there are domain specificities in the distribution of uncertainty cues in Hungarian texts;
- we show that domain specificities have a considerable effect on the efficiency of machine learning.

The structure of the paper is the following. First, related work on uncertainty detection is presented. Then our corpus is described in detail, which is followed by the elaboration of machine learning methods and results on uncertainty detection. The paper concludes with a discussion of results and possible ways for future work are also outlined.

This work is licenced under a Creative Commons Attribution 4.0 International License. Page numbers and proceedings footer are added by the organizers. License details: <http://creativecommons.org/licenses/by/4.0/>

## 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 indicated above, most 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 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., 2008; Nawaz et al., 2010), 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 the first Hungarian uncertainty corpus created for the purpose of this study.

## 3 Experiments

In this section, we present our methodology to detect uncertainty cues in Hungarian. We first describe the uncertainty categories applied and report some statistics on the corpus. Then we describe our machine learning approach based on a rich feature set.

### 3.1 The hUnCertainty Corpus

For the purpose of this study, we manually annotated texts from two domains. First, we randomly selected 1,081 paragraphs from the Hungarian Wikipedia dump. This selection contains 9,722 sentences and 180,000 tokens. Second, we downloaded 300 pieces of criminal news from a Hungarian news portal (<http://www.hvg.hu>), which altogether consist of 5,481 sentences and 94,000 tokens. In total, the hUnCertainty corpus consists of 15,203 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*). Here we just briefly summarize uncertainty categories that were annotated – for a detailed discussion, please refer to Szarvas et al. (2012) and Vincze (2013).

Linguistic uncertainty is traditionally connected to modality and the semantics of the sentence. For instance, the sentence *It may be raining* does not contain enough information to determine whether it is really raining (semantic uncertainty). There are several phenomena that are categorized as semantic uncertainty. A proposition is **epistemically** uncertain if its truth value cannot be determined on the basis of world knowledge. **Conditionals** and **investigations** also belong to this group – the latter is especially frequent in research papers, where authors usually formulate the research question with the help of linguistic devices expressing this type of uncertainty. Non-epistemic types of modality may also be listed here such as **doxastic** uncertainty, which is related to beliefs.

However, there are other uncertainty phenomena that only become uncertain within the context of communication. For instance, the sentence *Many researchers think that COLING will be the best conference of the year* does not reveal how many (and which) researchers think that, hence the source of the proposition about COLING remains uncertain. This is a type of discourse-level uncertainty, more specifically, it is called **weasel** (Ganter and Strube, 2009). On the other hand, **hedges** make the meaning of words fuzzy: they blur the exact meaning of some quality/quantity. Finally, **peacock** cues express unprovable (or unproven) evaluations, qualifications, understatements and exaggerations.

Some examples of uncertainty cues are offered here (in English, for the sake of simplicity):

EPISTEMIC: It **may** be raining.

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.

WEASEL: **Some** note that the number of deaths during confrontations with police is relatively proportional for a city the size of Cincinnati.

HEDGE: Magdalene Asylums were a **generally** accepted social institution until well into the second half of the 20th century.

PEACOCK: The main source of their inspiration was native Georgia, with its **rich** and **complex** history and culture, its **brehtaking** landscapes and its **courageous** and **hardworking** people.

Table 1 reports some statistics on the frequency of uncertainty cues in Hungarian and it is also visualized in Figure 1. 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 85% of the cues belong to the discourse-level uncertainty type.

Regarding different classes of uncertainty, we should mention that while weasels constitute the most frequent cue category in Wikipedia texts, they occur less frequently in the news corpus. On the other hand, doxastic cues are frequent in the news corpus but in Wikipedia texts, their number is considerably smaller.

Uncertainty cue	Wikipedia		News		Total	
	#	%	#	%	#	%
Weasel	2150	35.95	258	10.93	2408	28.87
Hedge	2100	35.12	800	33.88	2900	34.77
Peacock	788	13.18	94	3.98	882	10.57
Discourse-level total	5038	84.25	1152	48.79	6190	74.21
Epistemic	441	7.37	358	15.16	799	9.58
Doxastic	316	5.28	710	30.07	1026	13.30
Conditional	154	2.58	128	5.42	282	3.38
Investigation	31	0.52	13	0.55	44	0.53
Semantic total	942	15.75	1209	51.21	2151	25.79
Total	5980	100	2361	100	8341	100

Table 1: Uncertainty cues.

### 3.2 Machine Learning Methods

In order to automatically identify uncertainty cues, we developed 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 *magyarlanlc* (Zsibrita et al., 2013).

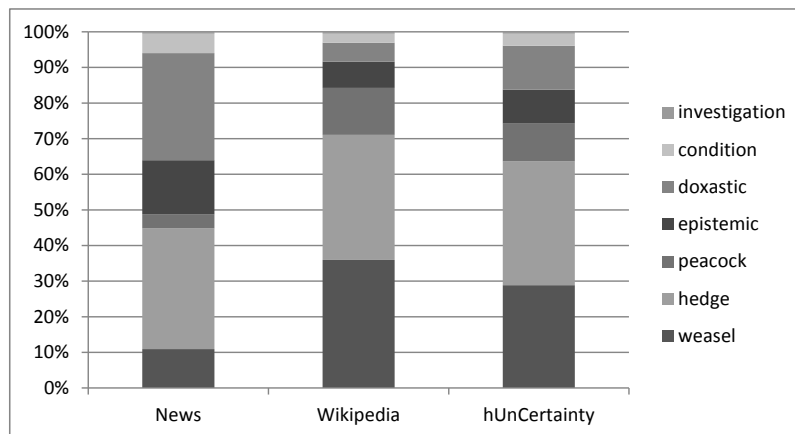


Figure 1: Distribution of cues across domains.

On the basis of results reported in earlier literature, 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 automatically collected uncertainty cues from the English corpora annotated on the basis of similar linguistic principles and manually 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. As mentioned before, modality and mood are morphologically expressed in Hungarian (e.g. in *csinálhatnánk* do-MOD-COND-1PL “we could do”, the suffix *-hat* refers to modality and the suffix *-ná* refers to conditional) hence 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 as these two latter verbal forms are typical instances of expressing generic phrases or generalizations in Hungarian, which are related to weasels. For each noun, its number (i.e. singular/plural) was marked as feature. For each pronoun, we checked whether it was an indefinite one since indefinite pronouns like *valaki* “someone” or *valamilyen* “some” are often used as weasel cues. For each adjective, we marked whether it was comparative or superlative as they can occur as peacock cues.
- **Syntactic features:** for each word, its dependency label was marked. For each noun, it was checked whether it had a determiner as determinerless nouns may be used as weasels in Hungarian. For each verb, it was checked whether it had a subject<sup>1</sup>.
- **Semantic/pragmatic features:** we manually 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

<sup>1</sup>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.

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 4.

### 3.3 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.

## 4 Results

Table 2 shows the results of the baseline and machine learning experiments on the hUnCertainty corpus, obtained by ten-fold cross validation.

Type	Dictionary lookup			Machine learning			Difference		
	Precision	Recall	F-score	Precision	Recall	F-score	Precision	Recall	F-score
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 2: 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.

We also experimented separately on the two domains. Table 3 shows those on the news subcorpus, whereas Table 4 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

Type	Dictionary lookup			Machine learning			Difference		
	Precision	Recall	F-score	Precision	Recall	F-score	Precision	Recall	F-score
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 3: Results on the news subcorpus.

Type	Dictionary lookup			Machine learning			Difference		
	Precision	Recall	F-score	Precision	Recall	F-score	Precision	Recall	F-score
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 4: Results on the Wikipedia subcorpus.

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 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 5 and 6 present the results, respectively, contrasted to the results achieved in the indomain settings. 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  $\rightarrow$  news setting, while the difference is much bigger for discourse-level uncertainty types in the news  $\rightarrow$  Wikipedia setting.

Type	Cross validation			Indomain ten fold			Difference		
	Precision	Recall	F-score	Precision	Recall	F-score	Precision	Recall	F-score
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 5: Cross-domain results: Wikipedia  $\rightarrow$  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.

Type	Cross validation			Indomain ten fold			Difference		
	Precision	Recall	F-score	Precision	Recall	F-score	Precision	Recall	F-score
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 6: Cross-domain results: news  $\rightarrow$  Wikipedia.

Results on the identification of uncertain sentences are summarized in Table 7, 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-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 $\rightarrow$ news	45.88	74.21	56.70
News $\rightarrow$ Wikipedia	35.73	84.61	50.24

Table 7: Machine learning results at the sentence level.

## 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 role 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  $\rightarrow$  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 (Szarvas et al., 2012), it can be seen that the task seems to be somewhat easier in English than in Hungarian: that paper reports F-scores from 0.6 to 0.8. One possible reason for this is 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. This is reflected in the number of different cues as well: in the English dataset, there are 166 different semantic cues while in Hungarian, there are 319 (and note that the Hungarian corpus is about half of the size of the English one). 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. Another language-specific feature is that Hungarian is a pro-drop language, so in some cases, the pronominal subject may be omitted from the sentence. Subjectless sentences are a typical strategy in Hungarian to express sourceless statements (weasels), but the subject can be deleted due to syntactic ellipsis as well, thus distinguishing between subjectless sentences that denote uncertainty and those that do not is a special task in Hungarian uncertainty detection.

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.



## 6 Conclusions

In this paper, we presented the first results on Hungarian uncertainty detection. For this, we made use of a manually annotated corpus, which contains texts from two domains: Wikipedia articles and pieces of news from a news portal. We contrasted the cue distribution in the two domains and we also experimented with uncertainty detection. For this purpose, we applied a supervised machine learning approach, which was based on sequence labeling and exploited a rich feature set. We reported the first results on uncertainty detection for Hungarian, which also prove that the performance on uncertainty detection is influenced by the domain of the texts. We hope that this study will enhance research on uncertainty detection for languages other than English.

In the future, we would like to improve our methods, especially in order to achieve better recall at the cue level. Furthermore, we would like to investigate domain specificities in more detail and we would also like to carry out some domain adaptation experiments as well.

## Acknowledgments

This research was supported by the European Union and the State of Hungary, co-financed by the European Social Fund in the framework of TÁMOP-4.2.4.A/2-11/1-2012-0001 “National Excellence Program”.

## References

- Wendy W. Chapman, David Chu, and John N. Dowling. 2007. Context: An algorithm for identifying contextual features from clinical text. In *Proceedings of the ACL Workshop on BioNLP 2007*, pages 81–88.
- Noa P. Cruz Díaz. 2013. Detecting negated and uncertain information in biomedical and review texts. In *Proceedings of the Student Research Workshop associated with RANLP 2013*, pages 45–50, Hissar, Bulgaria, September. RANLP 2013 Organising Committee.
- Richárd Farkas, Veronika Vincze, György Móra, János Csirik, and György Szarvas. 2010. The CoNLL-2010 Shared Task: Learning to Detect Hedges and their Scope in Natural Language Text. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning (CoNLL-2010): Shared Task*, pages 1–12, Uppsala, Sweden, July. Association for Computational Linguistics.
- Viola Ganter and Michael Strube. 2009. Finding Hedges by Chasing Weasels: Hedge Detection Using Wikipedia Tags and Shallow Linguistic Features. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 173–176, Suntec, Singapore, August. Association for Computational Linguistics.
- Jin-Dong Kim, Tomoko Ohta, and Jun’ichi Tsujii. 2008. Corpus annotation for mining biomedical events from literature. *BMC Bioinformatics*, 9(Suppl 10).
- Natalia Konstantinova, Sheila C.M. de Sousa, Noa P. Cruz, Manuel J. Mana, Maite Taboada, and Ruslan Mitkov. 2012. A review corpus annotated for negation, speculation and their scope. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Ugur Dogan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC’12)*, Istanbul, Turkey. European Language Resources Association (ELRA).
- John Lafferty, Andrew McCallum, and Fernando Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of ICML-01, 18th Int. Conf. on Machine Learning*, pages 282–289. Morgan Kaufmann.
- Marc Light, Xin Ying Qiu, and Padmini Srinivasan. 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*, pages 17–24.
- Bing Liu. 2012. *Sentiment Analysis and Opinion Mining*. Morgan & Claypool Publishers.
- Andrew Kachites McCallum. 2002. *MALLET: A Machine Learning for Language Toolkit*. <http://mallet.cs.umass.edu>.
- Ben Medlock and Ted Briscoe. 2007. Weakly Supervised Learning for Hedge Classification in Scientific Literature. In *Proceedings of the ACL*, pages 992–999, Prague, Czech Republic, June.

- Roser Morante and Caroline Sporleder. 2012. Modality and negation: An introduction to the special issue. *Computational Linguistics*, 38:223–260, June.
- Roser Morante, Vincent van Asch, and Antal van den Bosch. 2009. Joint memory-based learning of syntactic and semantic dependencies in multiple languages. In *Proceedings of CoNLL*, pages 25–30.
- Raheel Nawaz, Paul Thompson, and Sophia Ananiadou. 2010. Evaluating a meta-knowledge annotation scheme for bio-events. In *Proceedings of the Workshop on Negation and Speculation in Natural Language Processing*, pages 69–77, Uppsala, Sweden, July. University of Antwerp.
- Arzucan Özgür and Dragomir R. Radev. 2009. Detecting speculations and their scopes in scientific text. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 1398–1407, Singapore, August. Association for Computational Linguistics.
- Victoria L. Rubin, Elizabeth D. Liddy, and Noriko Kando. 2005. Certainty identification in texts: Categorization model and manual tagging results. In J.G. Shanahan, J. Qu, and J. Wiebe, editors, *Computing attitude and affect in text: Theory and applications (the information retrieval series)*, New York. Springer Verlag.
- Victoria L. Rubin. 2010. Epistemic modality: From uncertainty to certainty in the context of information seeking as interactions with texts. *Information Processing & Management*, 46(5):533–540.
- Roser Saurí and James Pustejovsky. 2009. FactBank: a corpus annotated with event factuality. *Language Resources and Evaluation*, 43:227–268.
- Burr Settles, Mark Craven, and Lewis Friedland. 2008. Active learning with real annotation costs. In *Proceedings of the NIPS Workshop on Cost-Sensitive Learning*, pages 1–10.
- Hagit Shatkay, Fengxia Pan, Andrey Rzhetsky, and W. John Wilbur. 2008. Multi-dimensional classification of biomedical text: Toward automated, practical provision of high-utility text to diverse users. *Bioinformatics*, 24(18):2086–2093.
- György Szarvas, Veronika Vincze, Richárd Farkas, György Móra, and Iryna Gurevych. 2012. Cross-genre and cross-domain detection of semantic uncertainty. *Computational Linguistics*, 38:335–367, June.
- Özlem Uzuner, Xiaoran Zhang, and Tawanda Sibanda. 2009. Machine Learning and Rule-based Approaches to Assertion Classification. *Journal of the American Medical Informatics Association*, 16(1):109–115, January.
- Veronika Vincze, György Szarvas, Richárd Farkas, György Móra, and János Csirik. 2008. The BioScope Corpus: Biomedical Texts Annotated for Uncertainty, Negation and their Scopes. *BMC Bioinformatics*, 9(Suppl 11):S9.
- Veronika Vincze. 2013. Weasels, Hedges and Peacocks: Discourse-level Uncertainty in Wikipedia Articles. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 383–391, Nagoya, Japan, October. Asian Federation of Natural Language Processing.
- Zhongyu Wei, Junwen Chen, Wei Gao, Binyang Li, Lanjun Zhou, Yulan He, and Kam-Fai Wong. 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)*, pages 58–62, Sofia, Bulgaria, August. Association for Computational Linguistics.
- Theresa Ann Wilson. 2008. *Fine-grained Subjectivity and Sentiment Analysis: Recognizing the Intensity, Polarity, and Attitudes of Private States*. Ph.D. thesis, University of Pittsburgh, Pittsburgh.
- János Zsibrita, Veronika Vincze, and Richárd Farkas. 2013. magyarlanc: A Toolkit for Morphological and Dependency Parsing of Hungarian. In *Proceedings of RANLP-2013*, pages 763–771, Hissar, Bulgaria.