Detecting Uncertainty Cues in Hungarian Social Media Texts

Veronika Vincze[1,2]
[1]Institute of Informatics, University of Szeged
Árpád tér 2., 6720 Szeged, Hungary
[2]MTA-SZTE Research Group on Artificial Intelligence
Tisza Lajos krt. 103., 6720 Szeged, Hungary
vinczev@inf.u-szeged.hu

Abstract

In this paper, we aim at identifying uncertainty cues in Hungarian social media texts. We present our machine learning based uncertainty detector which is based on a rich features set including lexical, morphological, syntactic, semantic and discourse-based features, and we evaluate our system on a small set of manually annotated social media texts. We also carry out cross-domain and domain adaptation experiments using an annotated corpus of standard Hungarian texts and show that domain differences significantly affect machine learning. Furthermore, we argue that differences among uncertainty cue types may also affect the efficiency of uncertainty detection.

1 Introduction

In several fields of natural language processing, the factuality of information plays an important role (Morante and Sporleder, 2012). Factual and non-factual information should be treated separately, more precisely, negated or speculative/uncertain information should not be mixed up with factual information. For instance, search engines should not retrieve documents where the information in question is negated or unreliable. Uncertainty detectors can help select reliable (certain) and unreliable (uncertain) parts of documents. Thus, developing uncertainty detectors is highly desirable for many fields of NLP (Morante and Sporleder, 2012; Farkas et al., 2010).

With the advent of Web2.0, many social media platforms have become widely popular, which means that a huge amount of user generated textual content appears on the web on a daily basis in the form of weblog posts, Facebook posts and comments, tweets etc. The majority of these contributions is published freely, i.e. without moderation, and even if they are moderated, moderators usually seek for utterances that violate the norms of the given page by using bad language or words that might hurt others' feelings. However, the reliability of the content of user generated data has hardly been investigated, in other words, social media users can publish whatever they want to and the factuality and (un)certainty of these contents may be an issue for those in need of collecting information from the web.

In this paper, we aim at identifying uncertainty cues in social media texts. We focus on Hungarian, a morphologically rich language. Later, we present our machine learning based uncertainty detector. We evaluate our system on a small set of manually annotated social media texts and we compare our results with those obtained by earlier experiments on Hungarian (Vincze, 2014). Finally, we also carry out some cross domain and domain adaptation experiments and we argue that data sparsity may be overcome by simple domain adaptation techniques.

The main contributions of this paper are the following:

- we report the first results on uncertainty detection in Hungarian social media texts;
- we introduce new features in the machine learning setting developed for the linguistic characteristics of social media texts;

Proceedings of the Workshop on Extra-Propositional Aspects of Meaning in Computational Linguistics, pages 11–21, Osaka, Japan, December 12 2016.

- we carry out cross domain and domain adaptation experiments and show that domain differences significantly affect machine learning;
- we argue that linguistic features of uncertainty cue types may also affect the efficiency of uncertainty detection;
- we argue that the efficiency of machine learning can be improved by adding out-domain data to the training.

2 Related Work

Uncertainty detection has recently gained popularity in the NLP literature. The CoNLL-2010 Shared Task aimed at detecting uncertainty cues in biological papers and Wikipedia articles written in English (Farkas et al., 2010). More recently, a special issue of the journal Computational Linguistics (Vol. 38, No. 2) was dedicated to detecting modality and negation in natural language texts (Morante and Sporleder, 2012).

Among the systems for uncertainty detection we can find rule-based ones (Light et al., 2004; Chapman et al., 2007) but also those based on machine learning methods, usually applying a supervised approach. Some of them used 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). 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. Most of these systems focus on the English language, however, we are aware of a study aiming at detecting uncertainty in Hungarian texts (Vincze, 2014).

Supervised machine learning methods were carried out on corpora from 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 (Ganter and Strube, 2009; Farkas et al., 2010; Szarvas et al., 2012), reviews (Konstantinova et al., 2012; Cruz Díaz, 2013) and social media (Wei et al., 2013).

Although most of the earlier studies experimented with indomain data, there are a few approaches that investigated domain differences. For instance, Szarvas et al. (2012) carried out domain adaptation for biological texts, news media and encyclopedia texts and Vincze (2014) experimented with pieces of news and Wikipedia texts.

Our system described in this paper is also based on supervised machine learning techniques, namely, sequence labeling. The system relies on a rich feature set of lexical, morphological, syntactic, semantic and discourse-based features and also exploits contextual features. To the best of our knowledge, ours is the first system that applies uncertainty detection for Hungarian social media texts.

Besides automatic uncertainty recognition, several studies investigated the distribution of uncertainty cues in different domains (Rizomilioti, 2006; Hyland, 1998; Falahati, 2006). Some of their findings revealed that papers belonging to the humanities contain significantly more cues than papers in sciences. Differences among domains also concern vocabulary items as well as the frequency of certain and uncertain usage of particular uncertainty cues. These findings highlight the practical importance of the domain adaptation of uncertainty detectors.

3 Experiments

In this section, we present our methodology to detect uncertainty cues in Hungarian social media texts. We first describe the corpora used together with the uncertainty categories applied and report some statistics on the corpus. Then our machine learning approach is presented in detail, together with its rich feature set.

3.1 Corpora

In this study, we made use of texts from two social media sources (Vincze et al., 2014). In the first phase of data preparation, we randomly collected, filtered and cleaned texts from Hungarian social media sites. On the one hand, public Facebook posts and comments were collected and on the other hand, questions and answers from a Hungarian FAQ portal[1] were also collected. This latter source of data was employed as it is supposed to be an authentic resource of the language use of young people in Hungary. The Facebook subcorpus of the data contains 1208 sentences and 8615 tokens whereas the FAQ subcorpus contains 728 sentences and 9702 tokens. Altogether, it makes up 1936 sentences and 18,317 tokens.

Although social media texts are written, their nature is rather similar to oral communication. Speed dominates this kind of communication, causing a number of possibilities for error. Quick typing leads to typos, abbreviations and lack of capitalization, punctuation and accentuated letters in these texts. Accentuated and unaccentuated vowels represent different sounds in Hungarian that can change the meaning of words (compare *szél* "wind" and *szel* "cut"), which may lead to ambiguities. Other types of linguistic creativity are also common, such as the use of smileys and English words and abbreviations in Hungarian texts. These characteristics should be considered when processing Hungarian social media texts.

In the second phase of data preparation, sentences were manually annotated for uncertainty cues (Vincze et al., 2014). Here we just provide a brief summary of uncertainty categories, for a more elaborated version, please refer to Szarvas et al. (2012) and Vincze (2013).

There are several different linguistic 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 or on the basis of the speaker's current mental state, e.g. *Steve may have failed at the exam*. **Conditionals** (*If it rains, we won't go to the party*) and **investigations** also belong to semantic uncertainty – the latter is especially frequent in research papers, where it is used to formulate research questions (*Here we aim at investigating whether domain specificities affect our results*). **Doxastic** uncertainty is related to beliefs (*I think Steve failed at the exam*).

Some sentences only become uncertain within the context of the discourse. For instance, the sentence *Many studies claim that the population of Cuba has increased in the past 10 years* does not reveal how many (and which) studies claim that, hence the source of the statement on Cuban population remains unclear. This is a type of **weasel** (Ganter and Strube, 2009). Furthermore, **hedges** blur the exact meaning of some quality/quantity as in *Approximately ten people can be admitted to the company*. Lastly, **peacock** cues express unprovable (or unproven) evaluations, qualifications, understatements and exaggerations like *This was the most gorgeous meal I've ever had in this fascinating restaurant*.

Some examples of uncertain sentences are offered here from the corpus, with the original spelling:

(1) Doxastic uncertainty:

ugy	*érzem*	*a*	*denver*	*ki*	*fog*	*kapni*	.
so	feel-1SG-OBJ	the	Denver	out	will	lose-INF	.

I think Denver will lose the game.

(2) Epistemic uncertainty:

De	*nem*	*biztos*	*hogy*	*mindenkinek*	*telik*	*1000Ft*	*/*	*fő*	*/*	*nap*	*kajára*	*!*
but	not	certain	that	everyone-DAT	afford-3SG	1000Ft	/	person	/	day	food-SUB	!

It is not certain that everyone can afford 1000 Ft per day per person for food.

(3) Condition:

Megint	*egy*	*reklám*	*hogy*	*ha*	*nincs*	*samsung*	*telód*	*nem*	*vagy*
again	an	advertisement	that	if	not.have-3SG	Samsung	phone-2SGPOSS	not	are
ember	*?*								
human	?								

[1] http://www.gyakorikerdesek.hu

Yet another advertisement that says that if you don't have a Samsung mobile, you are not a human?

(4) Weasel:

Na ez olyan , de mégis más .
well this such , but still different .
Well this is the same but somehow different.

(5) Hedge:

Elég nagy probléma .
enough big problem .
This is such a big problem.

(6) Peacock:

legeslegjobb vagy Magyarorszagon !
good-SUPERSUPERLATIVE are Hungary-SUP !
You are the best in Hungary!

In our experiments, we will also make use of the hUnCertainty corpus, which contains 1,091 randomly selected paragraphs from the Hungarian Wikipedia and 300 pieces of criminal news from a Hungarian news portal (`http://www.hvg.hu`) (Vincze, 2014).

Table 1 reports some statistics on the frequency of uncertainty cues in Hungarian (adapted from Vincze et al. (2014)). The annotation principles of the corpora were the same, hence cue distributions in the three domains are comparable. Based on Vincze et al. (2014), we can conclude that the domain of the texts affects the distribution of uncertainty cues: semantic uncertainty cues and discourse-level uncertainty cues are balanced in the news subcorpus but in the Wikipedia and social media corpora, more than 75% of the cues belong to the discourse-level uncertainty type.

Uncertainty cue	**hUnCertainty Wiki**		**hUnCertainty news**		**Social media**	
	#	%	#	%	#	%
Epistemic	439	7.8	358	15.16	21	4.08
Conditional	154	2.74	128	5.42	59	11.47
Doxastic	315	5.6	710	30.08	44	8.56
Investigation	31	0.55	13	0.55	1	0.19
Semantic total	939	16.69	1209	51.22	125	24.3
Peacock	787	14	94	3.98	192	37.35
Weasel	1801	32.02	258	10.93	50	9.72
Hedge	2098	37.3	799	33.86	147	28.59
Discourse-level total	4686	83.3	1151	48.77	389	75.6
Total	5625	100	2360	100	514	100

Table 1: Uncertainty cues in three domains.

The most obvious difference among the corpora is the presence of peacocks: their frequency is much higher in social media than in the other datasets. On the other hand, news tend to contain several instances of doxastic cues and Wikipedia has many weasels. In our experiments, we will demonstrate that such differences may strongly affect the performance of uncertainty detectors.

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 our social media corpus as well as the HunCertainty corpus and morphologically and syntactically parsed them with the help of the toolkit `magyarlanc` (Zsibrita et al., 2013).

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 relied on a method based on conditional random fields (CRF) (Lafferty et al., 2001) in our experiments. We used the MALLET implementation (McCallum, 2002) of CRF. Our feature set is constructed on the basis of earlier uncertainty detectors for Hungarian (Vincze, 2014), however, we added several new features, namely, discourse related features and social media features, due to the specialties of Hungarian social media texts.

- **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 (see Section 2) annotated for uncertainty 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 used as a feature. As Hungarian is a morphologically rich language, modality and mood are morphologically expressed (e.g. in *mehetnénk* go-MOD-COND-1PL "we could go", the suffix *-het* refers to modality and the suffix *-né* refers to conditional). Thus each verb was investigated whether it had a modal suffix and whether it was in the conditional mood. Also, we checked 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 a feature. Since indefinite pronouns like *valaki* "someone" or *valamilyen* "some" are often used as weasel cues, we checked whether the word was an indefinite pronoun. For each adjective, we marked whether it was comparative or superlative as they can often 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. Hungarian is a pro-drop language, which means that the pronominal subject is not obligatorily present in the clause. Furthermore, a common way to express generalization in Hungarian is to use a third person plural verb without a subject, which is one typical strategy of weasels. Thus, for each verb, it was checked whether it had a subject.

- **Semantic 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 positive and negative content developed for sentiment analysis (Liu, 2012) and checked whether the lemma of the given word occurred in these lists.

- **Discourse related features:** Hungarian is a discourse configurational language, which means that word order is determined by the information structure of the sentence. For instance, the preverbal (focus) position is preserved for the most important (novel) information within the sentence. Thus, for each word we noted its position within the sentence, its relative position to the verb and whether it occurred in the focus position.

- **Social media features:** In Hungarian, accentuated letters denote different phonemes, which might have an effect on word meaning as mentioned above. However, users tend to write without using accents in social media, so in order to simulate this scenario, we removed all accents from the texts

and also from the lists applied as lexical features. Smileys and character runs are also typical of social media texts, thus they were marked as features if the word contained or consisted of one.

As contextual features for each word, we applied as features the POS tags and dependency labels of words within a window of size two.

Based on this feature set, we carried out our experiments. It should be mentioned that, as there was only 1 investigation cue in the social media corpus, we neglected this class in our experiments due to sparseness problems.

As our main goal was to see how domain differences affect the efficiency of uncertainty detection, we experimented with several methods. First, we applied ten-fold cross validation on the social media corpus in order to check how a small amount of in-domain data can be exploited in uncertainty detection. Since we had the corpus hUnCertainty at hand, we also made use of cross-domain settings, where hUnCertainty was used as the training database but the evaluation was performed on the social media domain.

We also experimented with very simple domain adaptation techniques. We divided our social media corpus into a training and a test part, in a ratio of 80:20 and first trained our system on these splits. Later, we trained the system on hUnCertainty and evaluated it on the test split of the social media corpus. Lastly, we added the training split of the social media corpus to hUnCertainty and retrained the system with this additional in-domain set of texts. Evaluation was again carried out on the test split of social media texts.

For evaluation, we used the metrics precision, recall and F-score for each class and we also calculated a micro F-score to evaluate the performance of the system as a whole. The results of our experiments will be presented in Section 4.

4 Results

The first column of Table 2 represents the results of our in-domain experiments. It is revealed that doxastic cues can be relatively easily identified in social media text, even if only a small dataset is at our disposal. However, the detection of weasels is unsuccessful.

	In-domain			**Cross-domain**			**Difference**		
Cue	Recall	Precision	F-score	Recall	Precision	F-score	Recall	Precision	F-score
epistemic	6.52	60.00	11.76	4.35	18.18	7.02	-2.17	-41.82	-4.75
condition	8.54	21.88	12.28	29.27	36.36	32.43	20.73	14.49	20.15
doxastic	50.56	78.95	61.64	11.24	76.92	19.61	-39.33	-2.02	-42.04
peacock	7.41	25.64	11.49	0.74	12.50	1.40	-6.67	-13.14	-10.10
weasel	0.00	0.00	0.00	9.26	14.71	11.36	9.26	14.71	11.36
hedge	10.80	47.50	17.59	19.89	40.23	26.62	9.09	-7.27	9.02
Micro F	14.43	48.55	22.25	13.23	35.16	19.23	-1.20	-13.39	-3.02

Table 2: In-domain and cross-domain results on social media texts.

The results of our cross-domain experiments using the full amount of data from both corpora (i.e. hUnCertainty as training data and social media texts as test data) are presented in the second column of Table 2 and the relative differences to the in-domain results are shown in the third column. It can be seen that domain differences have mixed results on different classes of uncertainty cues. On the one hand, performance on peacocks, epistemic and doxastic cues is decreased while on the other hand, conditional cues, weasels and hedges can benefit from the out-domain data. All of this might suggest that different types of linguistic uncertainty behave differently in cross-domain context.

The results of our domain adaptation experiments are reported in Table 3 and the relative differences for in-domain, cross-domain and domain adaptation experiments are shown in Table 4. We can see that domain adaptation could outperform simple cross-domain experiments in the case of all of the uncertainty cue types, especially for epistemic and doxastic cues. However, for doxastic cues and peacocks, it can be observed that out-domain data just harmed performance as compared with the in-domain setting while for all the other cue types, out-domain data could improve the results.

	SM 80 → SM 20			hUnCertainty → SM 20			hUnCertainty+SM 80 → SM 20		
Cue	Recall	Precision	F-score	Recall	Precision	F-score	Recall	Precision	F-score
epistemic	0	0	0	11.11	100	20	22.22	100	36.36
condition	10	25	14.29	40	26.67	32	40	33.33	36.36
doxastic	68.18	88.24	76.92	9.09	100	16.67	63.64	93.33	75.68
peacock	3.45	16.67	5.71	0	0	0	3.45	33.33	6.25
weasel	0	0	0	28.57	28.57	28.57	28.57	33.33	30.77
hedge	20	77.78	31.82	28.57	45.45	35.09	31.43	52.38	39.29
Micro F	21.43	64.86	32.21	16.96	39.58	23.75	30.36	57.63	39.77

Table 3: Results of in-domain, cross-domain and domain adaptation experiments.

	Cross-domain vs. in-domain			DA vs. in-domain			Cross-domain vs. DA		
Cue	Recall	Precision	F-score	Recall	Precision	F-score	Recall	Precision	F-score
epistemic	11.11	100	20	22.22	100	36.36	11.11	0	16.36
condition	30	1.67	17.71	30	8.33	22.07	0	6.66	4.36
doxastic	-59.09	11.76	-60.25	-4.54	5.09	-1.24	54.55	-6.67	59.01
peacock	-3.45	-16.67	-5.71	0	16.66	0.54	3.45	33.33	6.25
weasel	28.57	28.57	28.57	28.57	33.33	30.77	0	4.76	2.2
hedge	8.57	-32.33	3.27	11.43	-25.4	7.47	2.86	6.93	4.2
Micro F	-4.47	-25.28	-8.46	8.93	-7.23	7.56	13.4	18.05	16.02

Table 4: Differences of performance in cross-domain and domain adaptation settings, compared to in-domain settings.

Figure 1 visualizes our cross-domain and domain adaptation results in terms of F-score, as compared to those achieved in the 80:20 in-domain setting.

5 Discussion

Here we experimented with two datasets: one including standard Hungarian texts (approximately 15K sentences) and one including social media texts (less than 2000 sentences). Our results indicated that there are domain differences among social media texts and standard Hungarian texts as uncertainty detection is concerned. Numerical results of cross-domain experiments were in all cases significantly outperformed by domain adaptation (t-test, $p = 0.0434$), hence even a small amount of in-domain data, that is, annotated social media texts (i.e. about 1600 sentences) can be exploited in uncertainty detection across domains. On the other hand, there is a significant difference in between results obtained in the 80:20 split settings and in the domain adaptation setting (t-test, $p = 0.0198$), which indicates that a larger amount of out-domain data can also contribute to better results. Thus, the best results can be achieved on social media texts in a scenario when a large amount of out-domain annotated data and a small amount of annotated in-domain data are jointly used as the training dataset.

Comparing the results of in-domain and cross-domain settings, we can observe that in the 80:20 training/test set scenario, epistemic cues and weasel cues cannot be identified at all, which might be related to the fact that these cues rarely occur in the social media data. However, in hUnCertainty, there are quite a few occurrences of them, hence out-domain data may help in their identification, even in a cross-domain setting.

In addition, more interesting differences can be found if uncertainty classes are contrasted. In the case of peacocks, doxastic cues and epistemic cues, cross-domain experiments clearly harm performance with regard to the in-domain settings, despite the much bigger training data. In the domain adaptation setting, however, the added value of in-domain data is noticeable, which indicates that these types of

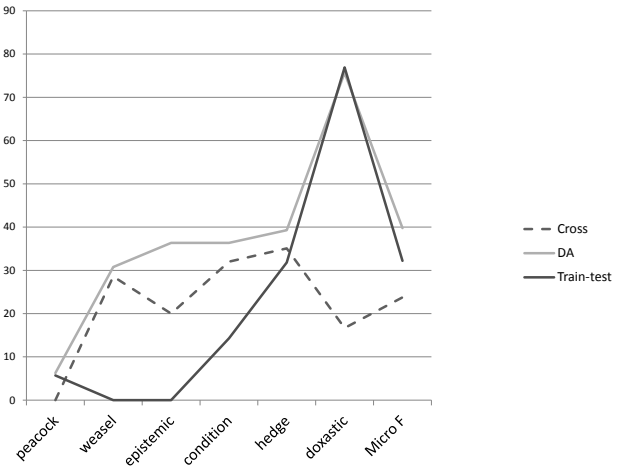

Figure 1: In-domain (Train-test), cross-domain and domain adaptation (DA) results per uncertainty class.

linguistic uncertainty are strongly domain-specific. In other words, the linguistic means to express them may change from domain to domain. For instance, the abbreviated form of *szerintem* "I think", *sztem* is very often used in social media texts as a doxastic cue but it is never used in its short form in standard texts. Thus, adding in-domain data to the training set might provide examples of such cases typical of social media language use. Also, it should be noted that precision values are relatively high for doxastic and epistemic cues even in the cross-domain settings. This might be related to the fact that these types of uncertainty cues occur rarely in social media texts and even if they occur, they are mostly different from the linguistic means used in standard texts. So, the system is unable to identify many of such cues based on the training data but when it marks one cue as doxastic/epistemic, it is most probably a true positive.

In contrast, condition cues, weasels and hedges seem to be less domain-specific according to the results: in-domain data also contributes positively to their identification but only to a moderate degree as compared to doxastic cues for instance (see the gaps in between cross-domain and domain adaptation results in Figure 1). Thus, social media users appear to exploit the same set of linguistic tools to express these types of linguistic uncertainty: the conditional mood is mostly used for conditions, indefinite pronouns are used for weasels, and intensifiers for hedges. We should also note, however, that weasels seem to be very difficult to learn only from social media data, which might be related to data sparsity.

The class of peacocks proved to be the most difficult one to detect in our experiment. There might be several reasons for that. First, this is the class which contained the most occurrences of uncertainty cues in social media, and also, this class is very diverse: it contained a lot of different cues with a low number of average occurrences. Thus, data sparsity might have hindered the performance of the system. Second, the usage of peacock cues seem to depend on the domain to a high extent. For instance, some standard expressions are used in their abbreviated forms like *sajna* instead of *sajnos* "unfortunately". Moreover, some vulgar expressions also occur as peacocks in social media like *szar* "shit", which again cannot be found in standard texts, i.e. Wikipedia and news portals. On the other hand, character runs were especially frequent with peacock cues (like *isteniiiii* instead of *isteni* "heavenly"), which may have also decreased the results. Finally, social media users tend to apply a lot of diminutive forms as peacock, even in the form of neologisms, which again are not easy to detect on the basis of the training data, e.g. *fini* and *fincsi* both occurred as diminutive forms of *finom* "fine, tasty".

Our results can also be contrasted to those obtained on standard Hungarian texts reported in Vincze (2014). The micro F-score interpreted for all uncertainty categories was 44.87. Here, our results are somewhat lower (a micro F-score of 39.77 after domain adaptation) but we should mention that processing social media texts is generally considered to be harder than processing standard texts and we had only a small amount of annotated data at our disposal. Also, it is interesting to note that comparing types of uncertainty cues, numerical results achieved on doxastic cues are higher than those achieved on standard corpora in the in-domain setting (F-scores of 61.64 and 49.15, respectively), which might be explained by the fact that the set of lexical items used as doxastic cues is rather limited in social media whereas in standard texts, there is a greater variety of such cues at the lexical level.

Some of our results suggest that a generalized treatment for all types of linguistic uncertainty classes may not be always viable. This is especially true for peacocks: performance on this class was constantly low, independently of the setting and training dataset we made use of. It seems that the treatment of peacocks require a more refined identification strategy, which might include enhancing the system with extended lists of sentiment words (as peacock cues are closely related to sentiment expressions), morphological analysis of diminutives and more sophisticated ways of processing neologisms and typos. Creating specific methods for the identification of such cues might be a possible direction for future research on uncertainty detection in the social media.

6 Conclusions

In this paper, we presented our system for identifying uncertainty cues in Hungarian social media texts. For this purpose, we created a machine learning based uncertainty detector which was based on a rich features set including lexical, morphological, syntactic, semantic and discourse-based features. Our system was evaluated on a small set of manually annotated social media texts. In order to see how domain differences affect machine learning, we also performed cross-domain and domain adaptation experiments using an annotated corpus of standard Hungarian texts. Our results indicated that specialties of social media texts should be accounted for when implementing an uncertainty detector. Also, selecting the training data has a significant effect on learning efficiency, but adding out-domain data to a small set of in-domain data can also contribute to performance. Moreover, differences among uncertainty cue types may also affect the efficiency of uncertainty detection and therefore some types of linguistic uncertainty may require special treatment in uncertainty detection.

In the future, we would like to improve our system by adding more refined techniques for processing Hungarian social media texts. We also intend to experiment with peacocks in more detail, which proved to be the most difficult uncertainty class to detect. Finally, as the majority of studies on uncertainty detection focus on English, it would be interesting to see how our system could perform on social media texts written in English. In this way, interlingual comparisons could also be made, which can be beneficial for both linguistics and natural language processing.

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.

David Clausen. 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, pages 120–125, Uppsala, Sweden. Association for Computational Linguistics.

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.

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

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.

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

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

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, may. 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.

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

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

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

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.

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

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

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.

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

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.

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

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.

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

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, Katalin Ilona Simkó, and Viktor Varga. 2014. Annotating uncertainty in hungarian webtext. In *Proceedings of LAW VIII - The 8th Linguistic Annotation Workshop*, pages 64–69, Dublin, Ireland, August. Association for Computational Linguistics and Dublin City University.

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.

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

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.

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

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.