Annotating Uncertainty in Hungarian Webtext

Veronika Vincze[1,2], **Katalin Ilona Simkó**[1], **Viktor Varga**[1]
[1]University of Szeged
Department of Informatics
[2]MTA-SZTE Research Group on Artificial Intelligence
vinczev@inf.u-szeged.hu
{kata.simko,viktor.varga.1991}@gmail.com

Abstract

Uncertainty detection has been a popular topic in natural language processing, which manifested in the creation of several corpora for English. Here we show how the annotation guidelines originally developed for English standard texts can be adapted to Hungarian webtext. We annotated a small corpus of Facebook posts for uncertainty phenomena and we illustrate the main characteristics of such texts, with special regard to uncertainty annotation. Our results may be exploited in adapting the guidelines to other languages or domains and later on, in the construction of automatic uncertainty detectors.

1 Background

Detecting uncertainty in natural language texts has received a considerable amount of attention in the last decade (Farkas et al., 2010; Morante and Sporleder, 2012). Several manually annotated corpora have been created, which serve as training and test databases of state-of-the-art uncertainty detectors based on supervised machine learning techniques. Most of these corpora are constructed for English, however, their domains and genres are diverse: biological texts (Medlock and Briscoe, 2007; Kim et al., 2008; Settles et al., 2008; Shatkay et al., 2008; Vincze et al., 2008; Nawaz et al., 2010), clinical texts (Uzuner et al., 2009), pieces of news (Saurí and Pustejovsky, 2009; Wilson, 2008; Rubin et al., 2005; Rubin, 2010), encyclopedia texts (Ganter and Strube, 2009; Farkas et al., 2010; Szarvas et al., 2012; Vincze, 2013), reviews (Konstantinova et al., 2012; Cruz Díaz, 2013) and tweets (Wei et al., 2013) have been annotated for uncertainty, just to mention a few examples.

The diversity of the resources also manifests in the fact that the annotation principles behind the corpora might slightly differ, which led Szarvas et al. (2012) to compare the annotation schemes of three corpora (BioScope, FactBank and WikiWeasel) and they offered a unified classification of semantic uncertainty phenomena, on the basis of which these corpora were reannotated, using uniform guidelines. Some other uncertainty-related linguistic phenomena are described as discourse-level uncertainty in Vincze (2013). As a first objective of our paper, we will carry out a pilot study and investigate how these unified guidelines can be adapted to texts written in a language that is typologically different from English, namely, Hungarian.

As a second goal, we will also focus on annotating texts in a new domain: social media texts – apart from Wei et al. (2013) – have not been extensively investigated from the uncertainty detection perspective. As the use and communication through the internet is becoming more and more important in people's lives, the huge amount of data available from this domain is a valuable source of information for computation linguistics. However, processing texts from the web – especially social media texts from blogs, status updates, chat logs and comments – revealed that they are very challenging for applications trained on standard texts. Most studies in this area focus on English, for instance, sentiment analysis from tweets has been the focus of recent challenges (Wilson et al., 2013) and Facebook posts have been analysed from the perspective of computational psychology (Celli et al., 2013). A syntactically annotated treebank of webtext has been also created for English (Bies et al., 2012). However, methods developed for processing English webtext require serious alterations to be applicable to other languages, for example Hungarian, which is very different from English syntactically and morphologically. Thus, in our pilot study we will annotate Hungarian webtext for uncertainty and examine the possible effects of the domain and the language on uncertainty detection.

In the following, we will present the uncertainty categories that were annotated in Hungarian webtext and we will illustrate the difficulties of both annotating Hungarian webtext and annotating uncertainty phenomena in them.

2 Uncertainty Categories

Here we just briefly summarize uncertainty categories that we applied in the annotation, based on Szarvas et al. (2012) and Vincze (2013).

Linguistic uncertainty is related 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 characteristic of research papers, where research questions usually express this type of uncertainty. Non-epistemic types of modality are also be listed here such as **doxastic** uncertainty, which is related to beliefs.

However, there are other linguistic phenomena that only become uncertain within the context of communication. For instance, the sentence *Many people think that Dublin is the best city in the world* does not reveal who exactly think that, hence the source of the proposition about Dublin 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 evaluations, qualifications, understatements and exaggerations.

The above categories proved to be applicable to Hungarian texts as well. However, the morphologically rich nature of Hungarian required some slight changes in the annotation process. 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*.

3 Annotating Hungarian Webtext

Annotating uncertainty in webtexts comes with the usual difficulties of working with this domain. We annotated Hungarian posts and comments from Facebook, which made the uncertainty annotation more challenging than on standard texts. Texts were randomly selected from the public posts available at the Facebook-sites of some well-known brands (like mobile companies, electronic devices, nutrition expert companies etc.) and from the comments that users made on these posts. For our pilot annotation, we used 1373 sentences and 18,327 tokens (as provided by magyarlanc, a linguistic preprocessing toolkit developed for standard Hungarian texts (Zsibrita et al., 2013)).

One fundamental property of social media texts is their similarity to oral communication despite their written form. The communication is online and multimodal; its speed causing a number of possibilities for error. The quick typing makes typos, abbreviations and lack of capitalization, punctuation and accentuated letters more common in these texts. Accentuated and unaccentuated vowels represent different sounds in Hungarian that can change the meaning of words (*kerek* "round", *kerék* "wheel" and *kérek* "I want"). Other types of linguistic creativity are also common, such as the use of emoticons and English words and abbreviations in Hungarian texts. However, these attributes do not characterize social media texts homogeneously. For instance, blog posts are closer to standard texts since they are usually written by a PR expert from the side of the brand, who presumably spends more time with elaborating on the text of the posts than an average user. On the other hand, comments and chat texts are closer to oral communication because users here want to react as quickly as possible, making them harder to analyze.

Our corpus of Facebook posts and comments exhibited a number of these properties. It contained a lot of typos, abbreviations and letters that should have been accentuated. These sometimes caused interpretation problems even for the human annotators; especially as these posts and comments were annotated without broader context. Lack of capitalization and punctuation was more common in the comment section of the corpus than in the posts. Emoticons were also frequent in both parts of the corpus.

Example 1: Typos in our corpus.

ugya ilynem *van csak fekete* ***elől*** *és szürke* ***hátúl*** – original

ugyanilyenem van csak fekete **elöl** és szürke **hátul** – standardized

(same.kind-POSS1SG have but black front and grey back)

“I have the same kind but its front is black and its back is grey.”

Example 2: Abbreviation in our corpus.

Amúgy meg ***sztem*** *Nektek nem kellene a Saját oldalatokon magyarázkodni!* – original

Amúgy meg **szerintem** Nektek nem kellene a saját oldalatokon magyarázkodni! – standardized

(by.the.way PART according.to-POSS1SG you-DAT not should the own site-POSS3PL-SUP explain.yourselves-INF)

“By the way I think you should not be explaining yourselves on your own site.”

Example 3: Lack of accentuation in our corpus.

es Marai Sandornak is ma van a szuletesnapja. – original

és Márai Sándornak is ma van a születésnapja. – standardized

(and Márai Sándor-GEN also today has the birthday-POSS3SG)

“And today is also Márai Sándor’s birthday”

4 Uncertainty in Hungarian Webtext

Apart from the above mentioned usual problems when dealing with webtext, other difficulties emerged during their uncertainty annotation. Uncertainty is often related to opinions, but writers of these texts do not usually express these as opinions, but as factual elements. Linguistic uncertainty is not annotated in these cases, as these sentences do not hold uncertain meanings semantically, even if certain facts in them are clearly not true or at least the writers obviously lack evidence to back them up.

Example 4: Information without evidence in our corpus.

Új megfigyelés, hogy az elektronok úgy viselkednek, mint az antioxidánsok.

(new observation that the electrons that.way behave as the antioxidants)

“It is a new observation that electrons behave as antioxidants.”

The uncertainty annotation of this text differed greatly from our corpus of Hungarian Wikipedia articles and news (Vincze, 2014), which domains are much closer to standard language use. Table 1 shows the distribution of the different types of uncertainty cues in these domains. Comparing this new subcorpus with the other two shows certain domain specific characteristics. Unlike Facebook posts and comments, the other two domains should not contain subjective opinions according to the objective nature of news media and encyclopedias. This is consistent with the difference in the proportion of peacock cues in each subcorpus: Facebook posts abound in them but their number is low in the other types of texts.

The relatively small number of hedges and epistemic uncertainty may be attributed to the previously mentioned observation that the writers of these posts and comments often make confident statements, even if these are not actual facts.

The resemblance of Facebook posts and comments to oral communication also means that elements that could also signify uncertainty can have different uses in this context. Certain phrases may indicate politeness or other pragmatic functions that in a different domain would mean and be annotated as linguistic uncertainty.

Example 5: The use of uncertain elements for politeness reasons in our corpus.

sajnos úgy tűnik a futáraink valamiért valóban nem érkeztek meg hozzátok szombaton

(unfortunately that.way seems the carriers-POSS1PL something-CAU really not arrive-PAST-3PL you-ALL Saturday-SUP)

"Unfortunately it seems like our carriers did not get to you on Saturday for some reason."

The phrase *úgy tűnik* "it seems" can express uncertainty in some contexts, but in the above example, it is used as a marker of politeness, in order to apologize for and mitigate the inconvenience they caused to their customers by not delivering some package in time.

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

Table 1: Uncertainty cues.

5 Conclusions

In this paper, we focused on annotating Hungarian Facebook posts and comments for uncertainty phenomena. We adapted guidelines proposed for uncertainty annotation of standard English texts to Hungarian, and we also showed that this domain exhibit certain characteristics which are not present in other domains that are more similar to standard language use. First, users usually express their opinions as facts, thus relatively less markers of hedges or epistemic uncertainty occur in the corpus. Second, uncertainty cue candidates can fulfill politeness functions, and apparently they do not signal uncertainty in these contexts. Third, the characteristics of webtext may cause difficulties in annotation since in some cases, the meaning of the text is vague due to typos or other errors.

Our pilot study of annotating Hungarian webtext for uncertainty leads us to conclude that the annotation guidelines are mostly applicable to Hungarian as well and webtexts also exhibit the same uncertainty categories as more standard texts, although the distribution of uncertainty categories differ among different types of text. Besides, politeness factors should get more attention in this domain. Our results may be employed in adapting annotation guidelines of uncertainty to other languages or domains as well. Later on, we would like to extend our corpus and we would like to implement machine learning methods to automatically detect uncertainty in Hungarian webtext, for which these findings will be most probably fruitfully exploited.

Acknowledgements

This research was funded in part by the European Union and the European Social Fund through the project FuturICT.hu (grant no.: TÁMOP-4.2.2.C-11/1/KONV-2012-0013). Veronika Vincze was partially funded by the National Excellence Program TÁMOP-4.2.4.A/2-11/1-2012-0001 of the State of Hungary, co-financed by the European Social Fund.

References

Ann Bies, Justin Mott, Colin Warner, and Seth Kulick. 2012. English Web Treebank. Linguistic Data Consortium, Philadelphia.

Fabio Celli, Fabio Pianesi, David Stilwell, and Michal Kosinski. 2013. Extracting evaluative conditions from online reviews: Toward enhancing opinion mining. In *Workshop on Computational Personality Recognition*, Boston, July.

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, may. European Language Resources Association (ELRA).

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.

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.

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.

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 Wilson, Zornitsa Kozareva, Preslav Nakov, Sara Rosenthal, Veselin Stoyanov, and Alan Ritter. 2013. SemEval-2013 Task 2: Sentiment Analysis in Twitter. In *Proceedings of the International Workshop on Semantic Evaluation*, SemEval '13, June.

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.