

ARE YOU FOR REAL?

LEARNING EVENT FACTUALITY IN CROATIAN TEXTS

Goran Glavaš, Jan Šnajder, Bojana Dalbelo Bašić
University of Zagreb
Faculty of Electrical Engineering and Computing
Unska 3, 10000 Zagreb, Croatia

ABSTRACT

There is a certain discrepancy between *real-world events* and their representations in text (*linguistic events* or *event mentions*). The event mentions often refer to future or hypothetical events that have not actually occurred or whose occurrence is uncertain. In this paper we address the problem of predicting event *factuality* in Croatian texts using supervised machine learning. For each event mention, we aim to predict its *polarity* (whether the denoted event has actually happened) and its *certainty* (the level of confidence that the denoted event has happened). We use only lexically-based features, in order to investigate how well this problem may be addressed for a resource-poor language such as Croatian. Our preliminary results suggest that while predicting event polarity using only lexically-based features is feasible, predicting event certainty mandates the use of more sophisticated features.

1 INTRODUCTION

In natural language texts (e.g., news articles) events from the real world (*extralinguistic events*) are represented by means of *linguistic events* or *event mentions*. However, there is a discrepancy between real-world events and their linguistic representations. Not all event mentions in text denote real-world events that actually occurred. Some event mentions indicate the absence of an event happening in the real world (e.g., “The president *didn’t* visit Cuba last month”) or the uncertainty of an event happening (e.g., “He *suspected* the plane *crashed*”). *Event factuality* has recently been defined as the level of information expressing the factual nature of eventualities mentioned in text [8]. The factuality of an event may be defined in terms of its *polarity* (not to be confused with sentiment polarity) and *certainty*, as illustrated by the two previous examples. Recognizing the *polarity* of event mentions aims to distinguish between event mentions describing an action or occurrence in the real world from those describing the lack of it. *Certainty*, on the other hand, refers to the level of confidence (*certain*, *probable*, or *possible*) expressed about the occurrence of a denoted real-world event.

Recognizing that an event is being reported as a fact rather than just a possibility, or that an event mention is referring to something that in reality never happened, may be important for many NLP applications, such as question answering, information extraction [5], and textual entailment [4]. In temporal reasoning [10], for example, events are usually placed on a timeline; information about event factuality is important

for deciding which mentions may be temporally grounded.

In general, the factuality of events is the result of an interaction of multiple linguistic elements at lexical, syntactical, and discourse levels [8]. In this paper we focus only on lexical sources of factuality (polarity and certainty clues), investigating the feasibility of factuality prediction for resource-poor languages, such as Croatian. We present a supervised machine learning model that combines multiple lexically-based features for polarity and certainty prediction, framing the polarity prediction as a binary classification task (*positive* vs. *negative* events) and certainty prediction as a ternary classification task (*certain*, *probable*, and *possible* events).

The remainder of the paper is organized as follows. In the following section we discuss the related work. In Section 3 we describe lexical features used for detection of polarity and certainty of events. We present experimental results in Section 4 and we conclude in Section 5.

2 RELATED WORK

Event extraction has received a lot of attention in the last couple of years. The interest was sparked by two evaluation campaigns specifically focusing on events: ACE [1] and TempEval [11]. In addition to extraction of event mentions themselves, the TempEval campaign also addressed the problem of automated extraction of event properties, such as tense, aspect, modality, and polarity. Although the notion of event polarity was already considered in the TempEval event extraction task, it was reduced to cases in which the event negation is expressed explicitly (e.g., “She did *not* teach for a year”). In the context of factuality detection, the polarity is considered more broadly and includes the cases in which the absence of an action is expressed implicitly (e.g., “The government *failed* to increase stability in the region.”).

Karttunen and Zaenen [5] discuss the factuality (or *veridicity*, as they call it) of events from a linguistic point of view, suggesting how information extraction approaches could benefit from their observations. They emphasize the importance of assigning factuality statements and expressions to their sources. E.g., in “The president said that the police *may* have *failed* yesterday”, the uncertainty of “police failing”, introduced by the factuality marker “*may*”, should be credited to “The president” rather than to the author of the text. Karttunen and Zaenen, however, do not consider polarity and certainty as two separate aspects of event factuality.

Saurí and Pustejovsky [8] extend the linguistic observations made by Karttunen and Zaenen and dissect event factuality

into polarity and certainty. Based on a rich set of linguistic observations they build a rule-based system for identifying event mention’s polarity (*positive* or *negative*) and certainty (*certain*, *probable*, and *possible*). As they observed that syntactic subordination is directly involved in the factual characterization of events, their rule-based computational model traverses the tree of syntactic dependencies in a top-down fashion, adjusting the factuality values for events according to polarity and certainty clues found in the nodes of the dependency tree. In this paper, we embrace the notions of event polarity and certainty as defined in [8], classifying event polarity as either *positive* or *negative*, and event certainty as either *possible*, *probable*, or *certain*. Unlike [8], we use supervised machine learning instead of a rule-based computational model. As our goal is to determine whether event polarity and certainty can be efficiently predicted for resource-poor languages, we use only lexically-based features for building our models.

Factuality statements for events mentioned in text always have a source, whether it is implicit (the author of the text) or explicit (usually a subject of predicates such as *say*, *know*, *think*, *believe*, etc.). As observed in [5] and [8], a single event mention can be assigned multiple (possibly even conflicting) factuality assessments coming from different sources (e.g., “I *doubt* that Greece *may* believe Germany *would save* its economy”). Identifying factuality sources and attaching factuality statements to them is out of the scope of this paper. We focus on determining the dominant polarity and certainty value for each event, regardless of factuality source.

3 FACTUALITY PREDICTION WITH LEXICALLY-BASED FEATURES

There are many languages, currently also including Croatian, for which linguistic tools and resources are rather scarce. For such languages, linguistic processing (even at the semantic level) has to rely on low-level, mostly lexically-based features. In this section we describe lexically-based features used for supervised learning of event polarity and certainty.

Polarity and certainty features. The following is a list of features used for both polarity and certainty classification:

1. *Word, lemma, and stem* of the event anchor – An event anchor is the word bearing the meaning of the event. Lemmatization was performed using the semi-automatically acquired morphological lexicon for Croatian [9]. A very simple stemming was employed; we remove the suffix from the last vowel in the word (or the penultimate vowel if the last letter in the word is also vowel). Words shorter than 5 letters were not stemmed;
2. *Ending* of the event anchor – The suffix of the word after the last vowel (or the penultimate vowel, if the last letter is a vowel);
3. *Morphosyntactic descriptor* of the event anchor – The MULTEXT-East morphosyntactic descriptors [2] are also obtained from the semi-automatically acquired morphological lexicon [9];
4. *Bag-of-words (BoW)* of the left and right context of the event anchor – We use two separate feature sets, one for

the left context BoW and one for the right context BoW. We define the context as a token window of size 5;

5. *Lemmas* of the first tokens preceding and following the event anchor;
6. *Event type* – The TimeML-based type of the event (cf. Section 5);
7. *Verbal and deverbal noun* – A binary feature indicating whether the event anchor is a verbal or a deverbal noun (e.g., *trčanje* – *running*). This feature is motivated by the observation that events expressed as (de)verbal nouns tend to be more hypothetical;
8. *Interrogative sentence* – A binary feature indicating whether the sentence containing the event is interrogative. In an interrogative sentence something is unknown (hence the question), thus events in interrogative sentences are more likely to be uncertain;
9. *Argument of another event* – A binary feature indicating whether an event anchor is an argument of another event. Events that take another event as argument are of *intentional action* (LACTION) type [7]. We do not use a syntactic parser, thus we cannot detect event arguments based on syntactic relations. Instead, we consider that event e_1 has another event e_2 as its argument if e_1 is of LACTION type and e_1 occurs in a two-token left context of e_2 . This will be wrong in a small number of cases in which two events are close to each other and there is no syntactic relation between them. Events that are direct arguments of other events tend to be non-factual more often. E.g., in “Napadač je *propustio postići* pogodak” (“The striker *failed* to *score* the goal”), the governing event (*propustio* – *failed*) indicates non-occurrence of “*scoring*”, while in “Kofi Annan je *pokušao uspostaviti* mir u Siriji” (“Kofi Annan *attempted* to *establish* peace in Syria”), the governing event (*pokušao* – *attempted*) indicates the uncertainty of “*establishing*”;

Polarity features. The following is the list of features used only for polarity classification:

1. *Negativity clues* found in the left context – The left context of the event consists of all the sentence tokens preceding the event anchor. We compiled a set of most frequent negativity clues in Croatian (inflectional forms of *not to be* and *not to want*, and additionally the words *no*, *noone*, *nothing*, *nowhere*, *never*, and *neither*): *ne*, *nisam*, *nisi*, *nije*, *nismo*, *niste*, *nisu*, *neću*, *nećeš*, *neće*, *nećemo*, *nećete*, *nikad*, *nigdje*, *nikome*, *ništa*, *ni*, *niti*;
2. *Negativity clues* found in the immediate left context – The same negativity clues as above, but restricted to those occurring closer (within a three-token window) to the event anchor. Considering both the immediate and more distant context, we aim to recognize the influence of both immediate and long-distance polarity modifiers on event polarity;
3. *Distance* between the event anchor and the closest negativity clue.

Certainty features. The following is the list of features used only for certainty classification:

1. *Conditionality clues* found in the left context – A set of conditionality clues found in the event anchor’s left context. We compiled a set of frequent conditionality clues (the words *if* and *whether*, and the inflectional forms of the verb *would*): *ako, ukoliko, bih, bi, bismo, biste*. The conditional clues strongly indicate hypotheticality of events and are therefore potentially important for predicting event possibility;
2. *Conditionality clues* found in the immediate left and right context – Two binary feature sets, one for each context side (left and right) of the event anchor. We look for the same conditional clues as above, but closer (within a three-token window) to the event anchor;
3. *Distance* between the event anchor and the closest conditional clue;
4. *Future tense clues* found in the left context – A set of future tense clues occurring in the left context of the event anchor. Future events have not occurred yet, hence by definition they introduce some uncertainty. We compiled a set of clues used for expressing the future tense (inflectional forms of the verb *will* and the perfective present tense forms of the verb *to be*): *ću, ćeš, će, ćemo, ćete, budem, budeš, bude, budemo, budete, budu*;
5. *Future tense clues* found in the immediate left and right context – Two sets of features, one for each context side (left and right) of the event anchor. We look for the same future tense clues as above, but closer (within a three-token window) to the event anchor;
6. *Distance* between the event anchor and the closest future tense clue;
7. *Possibility clues* found in the left context – A set of clues whose core meaning is closely related to uncertainty and possibility. We compiled a set of clues indicating possibility (inflectional forms of *can/could*, and the words *maybe* and *possible*): *moći, moglo, mogao, mogla, možda, moguće*;
8. *Possibility clues* found in the immediate left and right context – Two sets of features, one for each context side (left and right) of the event anchor. We consider the same possibility clues as above, but closer (within a three-token window) to the event anchor;
9. *Distance* between the event anchor and the closest possibility clue.

All features were computed respecting the sentence boundaries. All numeric features (distances from sets of clues) were z-score standardized on the training set.

We used support vector machines (SVM) [3] for both polarity and certainty classification. Because in our case the number of features is much larger than the number of examples (as a result of using predominantly lexical features encoded as sparse binary vectors), we used a linear kernel.

4 EVALUATION

We selected a set of 90 documents from the newspaper corpus *Vjesnik* previously annotated by five annotators for event and temporal relation extraction [6]. The set totals 4596 event

Table 1: *Dataset event factuality statistics*

	Positive	Negative	
Certain	3613	139	3752
Possible	450	39	489
Probable	330	25	355
	4393	203	

mentions, annotated by two annotators (each annotated half of the dataset) for polarity (*positive* or *negative*) and certainty (*certain*, *probable*, or *possible*). As expected for the newspaper genre, the majority of events (78.6%) were labeled as *positive* and *certain* (Table 1).

4.1 Polarity Evaluation

We evaluate the performance of our lexically-based model against two baselines. The first baseline is a simple majority class baseline, predicting every event to be of *positive* polarity. For second baseline we use a simple rule-based method that predicts an event to be of negative polarity if only if one of the negativity clues is found in the immediate left context of the event mention. We estimate the prediction performance using a 10-fold cross validation on the set of annotated events. The results are presented in Table 2.

The difference in performance between the supervised model and the rule-based baseline for both positive and negative polarity classes is not statistically significant at 0.05 level. We credit this to the limited size of the training set in which there is an insufficient number of negative polarity events expressed by lexical units other than the negativity clues (e.g., “Napadač je propustio postići pogodak” – “The attacker failed to score the goal”). Supervised models based on lexical features usually require larger datasets, and we believe that our supervised polarity classifier would benefit from annotating more data. Overall, it seems that both rule-based baseline and the supervised model are capable of recognizing events of negative polarity at satisfactory rates.

4.2 Certainty Evaluation

The performance of our supervised model on the certainty classification task is also evaluated against two baselines. The first baseline is a majority class baseline that predicts every event to be *certain*. The second is a rule-based baseline that predicts every event to be *possible* if one of the *conditional clues* is present in its context, and *probable* if any of the *future clues* is present within its context. As in polarity classification, we estimated the prediction performance using 10-fold cross validation. The results are presented in Table 3.

Our lexically-based model significantly outperforms both baselines. However, it is difficult to expect that the model with F-scores between 40% and 50% for *probable* and *possible* classes can be put to use in real-world applications. The fact that precision is significantly higher than recall indicates that recognizing *possible* and *probable* events requires additional features.

Table 2: Polarity prediction performance

	Positive			Negative			Macro-average		
	P	R	F1	P	R	F1	P	R	F1
Baseline (majority)	95.52	100.0	97.72	–	–	–	47.76	50.00	48.85
Baseline (rule-based)	98.75	99.11	98.93	79.47	73.30	76.26	87.58	86.21	86.88
Supervised model	98.49	99.64	99.06	89.67	67.48	77.22	94.08	83.56	88.51

Table 3: Certainty prediction performance

	Certain			Possible			Probable			Macro-average		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Baseline (majority)	81.44	100.0	89.77	–	–	–	–	–	–	27.15	33.33	29.92
Baseline (rule-based)	88.43	89.56	88.99	35.29	25.77	29.78	37.81	46.76	41.81	53.84	54.03	53.93
Supervised model	88.45	96.08	91.95	54.29	36.20	43.44	61.82	37.16	46.41	68.19	56.48	61.79

5 CONCLUSION

Assessing factuality of events in text is important for many NLP applications. Factuality of events can be defined in terms of their polarity and certainty. In this paper we presented a supervised machine learning approach to recognizing event factuality in Croatian texts. Our model uses lexically-based features, thus it is suitable for resource-poor languages. Our results indicate that while using a lexically-based model to predict the factual polarity of events is feasible and yields satisfactory results, the performance is not statistically significant when compared to a simple rule-based baseline. On the other hand, although it does outperform the baseline, the model still seems insufficient for capturing factual certainty of events, suggesting that this task mandates the use of syntactic (e.g., for capturing the long-distance dependencies) and semantic (e.g., semantic verb classes) features.

Acknowledgments

This work has been supported by the Ministry of Science, Education and Sports, Republic of Croatia under Grant 036-1300646-1986.

References

- [1] ACE. Evaluation of the detection and recognition of ace: Entities, values, temporal expressions, relations, and events. 2005.
- [2] T. Erjavec, C. Krstev, V. Petkevič, K. Simov, M. Tadić, and D. Vitas. The MULTEXT-East morphosyntactic specifications for Slavic languages. In *Proceedings of the EAACL2003 Workshop on Morphological Processing of Slavic Languages*, pages 25–32, 2003.
- [3] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874, 2008.
- [4] A. Hickl and J. Bensley. A discourse commitment-based framework for recognizing textual entailment. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, pages 171–176. Association for Computational Linguistics, 2007.
- [5] L. Karttunen and A. Zaenen. Veridicity. *Annotating, extracting and reasoning about time and events*, (05151), 2005.
- [6] M. Marović, J. Šnajder, and G. Glavaš. Event and temporal relation extraction from Croatian newspaper texts. In *Proc. of the Eighth Language Technologies Conference*. Slovenian Language Technologies Society, 2012. In press.
- [7] J. Pustejovsky, J. Castano, R. Ingria, R. Sauri, R. Gaizauskas, A. Setzer, G. Katz, and D. Radev. Timeml: Robust specification of event and temporal expressions in text. *New Directions in Question Answering*, 2003:28–34, 2003.
- [8] R. Sauri and J. Pustejovsky. Are you sure that this happened? assessing the factuality degree of events in text. *Computational Linguistics*, 38(2):261–299, 2012.
- [9] J. Šnajder, B.D. Bašić, and M. Tadić. Automatic acquisition of inflectional lexica for morphological normalisation. *Information Processing & Management*, 44(5):1720–1731, 2008.
- [10] A. ter Meulen. Temporal reasoning in natural language. *Foundations of Artificial Intelligence*, 1:559–585, 2005.
- [11] M. Verhagen, R. Sauri, T. Caselli, and J. Pustejovsky. Semeval-2010 task 13: Tempeval-2. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 57–62. Association for Computational Linguistics, 2010.