University of Zagreb
Faculty of Electrical Engineering and Computing
Text Analysis and Knowledge Engineering Lab

take[lab];

Are You for Real? Learning Event Factuality in Croatian Texts

Goran Glavaš, Jan Šnajder, and Bojana Dalbelo Bašić

October 8, 2012 | **Conference on Data Mining and Data Warehouses**

Outline

Event Factuality

take[lab];

- Real-world events are denoted in text by means of *linguistic events* (*event mentions*)
- However, not all event mentions denote real-world events that actually occurred
  - Absence of happening in the real world (e.g., "The president didn't *visit* Cuba.")
  - Uncertainty of real-world event happening (e.g., "He *suspected* the plain had *crashed*.")
- *Event factuality* is the level of information expressing the factual nature of eventualities mentioned in text (Saurí and Pustejovsky, 2012)
  - Polarity of event mentions – action/occurrence or the lack of it
  - Certainty of event mentions – level of confidence expressed about the real-world event occurring

Lexical Approach to Event Factuality

take[lab];

- Event factuality is the result of of an interaction of multiple linguistic elements (Saurí and Pustejovsky, 2012) at
  - Lexical level
  - Syntactical level
  - Discourse level
- Focus on lexical sources of factuality – investigating the feasibility of automated event factuality prediction for resource-poor languages, such as Croatian

Supervised Lexically-based Model

take[lab];

- Two tasks
  - Predicting *polarity* (*positive*, *negative*) of event mentions
  - Predicting *certainty* (*certain*, *possible*, *probable*) of event mentions
- Supervised learning using support vector machines (SVM) with a rich set of purely lexical features
  - Some features were used for both tasks
  - Linear kernel considering the large feature space

Polarity and Certainty Features

take[lab];

- Word, lemma, and stem of the *event anchor* (a word bearing the meaning of the event)
  - Lemmatization performed using semi-automatically acquired morphological lexicon (Šnajder et al., 2008)
  - Simple stemming removing the suffix from the last vowel
- 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)
- Morphosyntactic descriptor (MSD) of the event anchor – the MULTEXT-East MSDs (Erjavec et al., 2003)
- Bag-of-words (BoW) of the left and right context of the event anchor
  - Two features – one for each context side
  - Context is a 5-token window

Polarity and Certainty Features

take[lab];

- Event type – The manually labelled TimeML-based type of the event anchor
- Verbal and deverbal noun – event anchors that are (de)verbal nouns tend to be hypothetical more often
- Interrogative sentence – events mentions of interrogative sentences are more likely to be uncertain
- Argument of another event – feature indicates if an event anchor is an argument of another event anchor
  - Such events tend to be non-factual more often (e.g., "Napadač je propustio postići pogodak")
  - Event e_1 has another event e_2 as its argument if e_1 is of type I_ACTION and e_1 occurs in a two-token left context of e_2

Polarity features

take[lab];

- Negativity clues found in left context
  - The left context of event anchor consists of all sentence tokens preceding the event anchor
  - Negativity clues are inflectional forms of *not to be* and *not to want* plus *no*, *noone*, *nothing*, *never* and *neither*
- Negativity clues found in *immediate* left context
  - Negativity clues occurring within a 3-token left window from the event anchor
- Distance between the event anchor and the closest negativity clue

Certainty Features

take[lab];

- Conditionality, future tense, and possibility clues clues found in left context
  - Conditionality clues are inflectional forms of *would* plus *if* and *whether*
  - Future tense clues are inflectional forms of *will* and the perfective present tense forms of *to be*
  - Possibility clues are inflectional forms of *can/could* plus *maybe* and *possibly*
- Conditionality, future tense, and possibility clues found in *immediate* left context
  - Occurring within a 3-token left window from the event anchor
- Distance between the event anchor and the closest Conditionality/future tense/possibility clue

Dataset

take[lab];

- Selection of 90 documents from the newspaper corpus *Vjesnik*, previously annotated for event and temporal relation extraction (Marović et al., 2012)
  - Total of 4596 events annotated for polarity and certainty by two annotators
- Majority of events (78.6%) labelled as *positive* and *certain*
  - Expected for the newspaper genre

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

Evaluation

take[lab];

- Two baselines for both tasks
  - Majority class baseline (predicting every event to be positive and certain)
  - Simple rule-based baseline
    - Polarity – *negative* if any of the negativity clues if found in its immediate left context
    - Certainty – *possible* if any of the conditionality clues id found in its immediate left context, and *probable* if any of the future clues is found
- 10-fold cross validation was performed and the average performance is reported

Polarity Results

take[lab];

	Positive	Negative	Macro-average
Baseline (majority)	97.72	–	48.85
Baseline (rule-based)	98.93	76.26	86.88
Supervised model	**99.06**	**77.22**	**88.51**

- Difference in performance between the supervised model and the rule-based baseline 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 events of negative polarity expressed by lexical units other than the negativity clues
  - E.g., "Napadač je propustio *postići* pogodak"

Certainty Results

take[lab];

	Certain	Possible	Probable	Macro-av.
Baseline (majority)	89.77	–	–	29.92
Baseline (rule-based)	88.99	29.78	41.81	53.93
Supervised model	**91.95**	**43.44**	**46.41**	**61.79**

- Supervised model significantly outperforms both baselines
- Performance rates of 40% or 50% for *probable* and *possible* classes are not satisfactory for real-world applications
- Precision significantly higher than recall – need for additional syntactic and discourse level based features

Conclusions

take[lab];

- We presented a supervised machine learning approach to recognizing event factuality in Croatian texts
- The model is based purely on lexical features, thus suitable for resource-poor languages
- Predicting event polarity
  - Feasible with the supervised model
  - Not significantly better than a simple rule-based baseline
- Predicting event certainty
  - Not feasible using purely lexical features although the supervised model outperforms a rule-based baseline
- Determining certainty of event mentions mandates the use of syntactic, semantic, and discourse based features

Thank you for your attention!

take[lab];

Text Analysis and Knowledge Engineering Laboratory
www.takelab.hr
info@takelab.hr, takelab@fer.hr