On tense interpretation in Slavic: A corpus study and computational model

Damir Cavar, Zoran Tiganj

NLP-Lab

Indiana University

NLP Lab Team

- The following graduate students and colleagues at Indiana University at Bloomington contributed to TIE-ML, corpora, and the computational and theoretical frameworks:

Ali Abdulaziz Aljubailan, Soyoung Kim, Billy Dickson, Andrew Davis, Matthew Fort, Ludovic Mompelat, Yuna Won

- See for more details:

http://nlp-lab.org/

Events and Tense

- Interaction of temporal and event properties in complex sentences
- Scope relations between clauses
  - Interpretation of tense associated with a clause level predicate.
    - see sequence of tense puzzle in Kiparsky (2002)

- Interpretation of embedded predicates:
  - Reuters reported… \
    [that Apple merged with Alphabet]
  - Reuters will report… /

Motivation

- Research on event and temporal logic in different genres (e.g., dialog, news, manual)
- Narrowed down:
    - Event sequencing
    - Event duration
    - Scope relation & grammatical restrictions
- Causality
    - Sequencing of events and temporal correlation
- Quantitative and Qualitative Study

Example: Event Sequencing

- Sequencing of events into sub-events
    1. *Narežite korijen celera na prutiće*
    2. *i kuhajte oko 7 minuta u kipućoj vodi – neka omekša,*
    3. *ocijedite na papirnatom ručniku*
    4. *pa uvaljajte u brašno,*
    5. *zatim u razmućena jaja,*
    6. *pržite ga na vrućem ulju dok ne dobije zlatnosmeđu boju*
    7. *pa izvadite na papirnati ručnik.*
- Variation:

Prije nego što pržite korijen celera 6.

…

narežite ga na prutiće… 1.

Example: Event Sequencing

- Sub-event sequencing impacts our interpretation of causality
  - Exposition of events in sequence leads to default causal relation interpretations
    - X was a health worker
    - X received the vaccine
    - X died a month later
    - = vaccine might have caused the death

  - Deceptive narrative, propaganda, and “fake news” utilize default causality interpretation tendency
  - Medical report narrative leading to detection of adverse drug reactions, etc.

Example: Common Sense Duration

- Event Deviation
  - *pržite korijen celera na vrućem ulju…*
  - *#pržite korijen celera na vrućem ulju dva sata*
  - *#pržite korijen celera na vrućem ulju dva sata*

- Common sense interpretation of event durations for
  - Detection of deception
  - Abnormality detection (modes of normal behavior vs. abnormal behavior)
  - General event classification or detection (cooking longer or shorter implying other variables)
  - etc.

Factivity

- Factivity of events
  - Past tense predicates strongly indicate that the described event occurred
  - Here: Gazprom and Lukoil are now a single organization

- Past tense implies factivity or a positive truth value:

(1) Газпром объединился с Лукойлом.

Factivity

- Scope effects
  - Simple matrix clause with a past tense predicate does not change the default factivity interpretation

(2) Reuters подтвердил, что Газпром объединился с Лукойлом.

- Still default: Gazprom and Lukoil are now a single organization

Factivity

- Altering the tense of the matrix clause:
  - Affects the interpretation of the temporal properties of the embedded predicate significantly:

(3) Reuters Завтра подтвердит, что Газпром объединился с Лукойлом.

- No longer certain: that Gazprom indeed merged with Lukoil at speaker time.
- Not possible to exclude the merger prior to speaker time neither.
- Future tense of the matrix clause provides a new time frame that affects the past tense interpretation of the embedded clause.

Tense Agreement and Selection

- Adjunct clauses modifying a predicate
  - Agree with respect to tense with the modified predicate

- Selected clauses (controlled by a predicate), as with *report* (“izjaviti”) and the subordinate clause

  (4) [Kada smo bili u Parizu] Reuters je bio izjavio [da je Gazprom preuzeo Lukoil]

- (5) in contrast to (4) is deviant and semantically problematic, if not completely ungrammatical

  (5) * [Kada smo bili u Parizu] Reuters će izjaviti [da je Gazprom preuzeo Lukoil]

Tense Agreement and Selection

- Contrast

  (4) [Kada smo bili u Parizu] Reuters je bio izjavio [da je Gazprom preuzeo Lukoil]

  (5) *[Kada smo bili u Parizu] Reuters će izjaviti [da je Gazprom preuzeo Lukoil]

- The ungrammaticality of (5) is due to the mismatch between the tense in the modifier headed by *biti* and the matrix clause head predicate *izjaviti*.
- We observe obligatory tense agreement constraints with adjunct clauses and scope-based interpretation of tense in selected clauses.

Tense Agreement and Selection

- Adjunct clauses modifying a predicate
    - Agree with respect to tense with the modified predicate

- Subordinate clauses selected by a predicate
    - Do not agree with respect to tense with the selecting predicate

- The tense and event properties of selected clauses depend on the tense and event properties of the selecting verb.

Corpus Study

- Goals:
  - Capture quantitative and qualitative aspects of
    - Tense sequencing in narratives
    - Scope relations and effects on factivity and event variables
    - Temporal duration and common sense values for prototypical events
    - Dependency / Selection effects between tenses in adjunct and complement clauses
    - Intra- and cross-linguistic variation
  - Corpus development and annotation
    - Annotation standards and approaches
    - Theoretical background assumptions
  - Engineering of Computational Algorithms
    - Mapping of event sequences on the time axis
    - Identification of event time and temporal durations of events
    - Factivity checks, deception detection, anomaly detection, …

Sequence of Tense

- Reichenbachian theory of tense and aspect (Reichenbach 1947)
  - Temporal Intervals
    - **E** (event time)
    - **R** (reference time, the time to which for example temporal reference items refer)
    - **S** (speaker time) (**P** = perspective time in Kiparsky, 2002)
  - Tenses
    - Simple Present (E,R,S where R = now) (*I see Ross now.*)
    - Simple Past (E,R_S where R = yesterday) (*I saw Ross yesterday.*)
    - Simple Future (S_E,R where R = tomorrow) (*I will see Ross tomorrow.*)
    - Present Perfect (E_S,R where R = now) (*I have seen Ross now.*)
    - Past Perfect (E_R_S where R = yesterday) (*I had seen Ross yesterday.*)
    - Future Perfect (S_E_R where R = tomorrow) (*I will have seen Ross tomorrow.*)

Questions

- Topology of predicate tense and derived tense in complex predicate structures.
  - Matrix tense impacts embedded tense:
    - Shift of event or reference time of embedded clauses

- Among others:
  - Which of the sub-variables undergoes what kind of shift under specific circumstances?
  - How does the tense interpretation interact with factivity?
  - How can common sense interpretation of event durations and sequencing be derived/computed?

Approach

- Corpus Annotation – Automatic
  - syntactic scope relations (dominance and precedence at least),
  - the tense of the particular clauses, and
  - the semantic relations between clauses in terms of selection vs. modification.
  - Using parsers and language models.
- Corpus Annotation – Manual
  - Sequencing of sub-events
  - Duration of events (incl. overt temporal markers)

Existing Standards

- TimeML and Annotation Standards (Pustejovsky et al. 2003)
  - XML-based markup language and metadata standard
  - Annotating events and temporal expressions in natural language or time information in general
  - Most detailed and theoretically grounded framework
- Elements
  - Four core annotation tags
    - EVENT: encodes events that are punctual or that have a duration associated with them
    - TIMEX3: encoding temporal functions and reference points
    - SIGNAL: used to mark up function words with a temporal reference
    - LINK: encodes relationships between events

Complexity

- Issues with existing sophisticated standards:
  - Training time is excessive, introduction to event semantics and temporal logic, and language-specific peculiarities
  - Annotation time per complex sentence can consume significant time
  - Annotator agreement evaluation is complex given many detailed annotation tags and variations
  - Annotation errors increase with higher complexity of annotation standards
- Solution
  - Simplification of annotation standard
  - Simplification of annotation tasks

TIE-ML Standard

- Temporal Information Event Markup Language (TIE-ML) (Cavar et al. 2021)
    - Simplified temporal annotation schema
    - Focuses on event sequencing annotation and clause level temporal properties of main predicates
    - Goal
        - improve upon previous markup strategies' accuracy and productivity via simplification
        - increasing the production of *good data* with the event and temporal properties annotated will
            - facilitate the development of computational linguistic, AI, machine learning models for applications that can benefit from specific semantic analytics

TIE-ML

- Annotation formats
  - JSON
  - XML
  - Simple text-based
  - Using technologies like INCEpTION (web-based corpus annotation), see https://inception-project.github.io/

- Extension of existing corpora and annotations
  - Syntactic treebanks (providing dominance and hierarchical relations, as well as functional structures and annotations)
  - Discourse corpora providing some semantic properties for utterances

TIE-ML XML Example

```
<tieml>
        <s>
                <c eventid="1">    Danny watched the movie    </c>
                <c eventid="2">    and ate popcorn    </c>.
        </s>
        <s>
                <c eventid="3">    Josh brought the pizza    </c>.
        </s>
</tieml>
```

TIE-ML Example

```
<s>
        <c eventid="1" timeslot="2">Before you fry the vegetables</c>
        <c eventid="2" timeslot="1">chop them into cubes</c>.
</s>
```

Or

```
<s>
        <c e="-1" s="0">Danny watched the movie.</c>
</s>
```

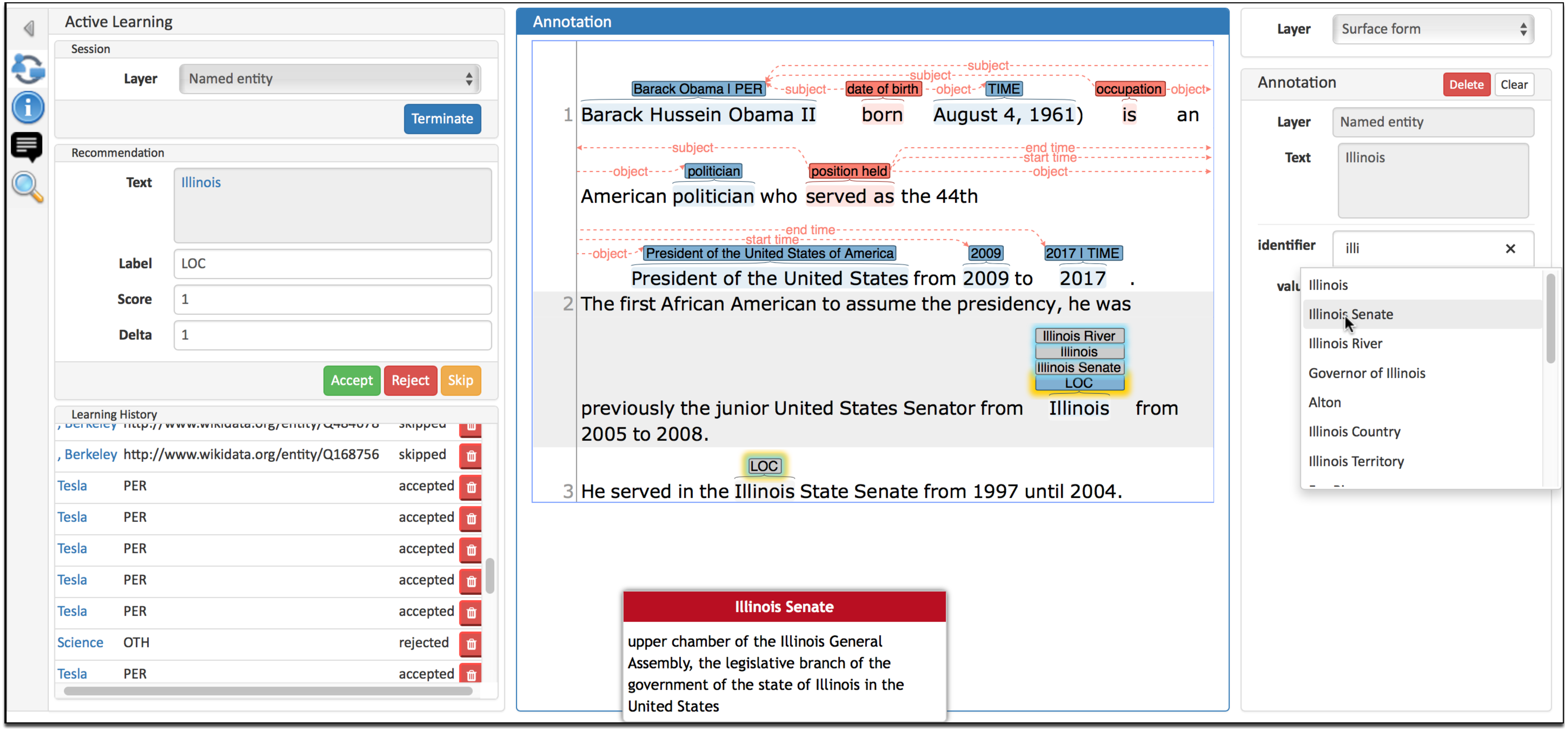
CoNLL Style Syntax

CLAUSE	ID	TS
Which car	1	2
did John say	2	1
that Mary will like?	1	2
She will like the blue car.	1	1

Example: Temporal Scope

- Temporal scope and activity
  - Apple acquired Alphabet.
  - Reuters reported that Apple acquired Alphabet.
  - Reuters will report that Apple acquired Alphabet.

INCEpTION



Active Learning

Session

Layer: Named entity

Recommendation

Text: Illinois

Label: LOC

Score: 1

Delta: 1

Learning History

, Berkeley	http://www.wikidata.org/entity/Q484078	skipped
, Berkeley	http://www.wikidata.org/entity/Q168756	skipped
Tesla	PER	accepted
Tesla	PER	accepted
Tesla	PER	accepted
Tesla	PER	accepted
Tesla	PER	accepted
Science	OTH	rejected
Tesla	PER	accepted

Annotation

1 Barack Hussein Obama II born August 4, 1961) is an American politician who served as the 44th President of the United States from 2009 to 2017 .

2 The first African American to assume the presidency, he was previously the junior United States Senator from Illinois from 2005 to 2008.

3 He served in the Illinois State Senate from 1997 until 2004.

Illinois Senate

upper chamber of the Illinois General Assembly, the legislative branch of the government of the state of Illinois in the United States

Layer: Surface form

Annotation

Layer: Named entity

Text: Illinois

identifier: illi

valu

Alignment

- Existing treebank data linked
  - Via sentence and clause ID

- Pre-processing of texts using Natural Language Processing pipelines
  - Morphological analyzers and part of speech taggers
  - Constituent and Dependency parsers
  - Tense tagging and Clause segmentation

Challenges

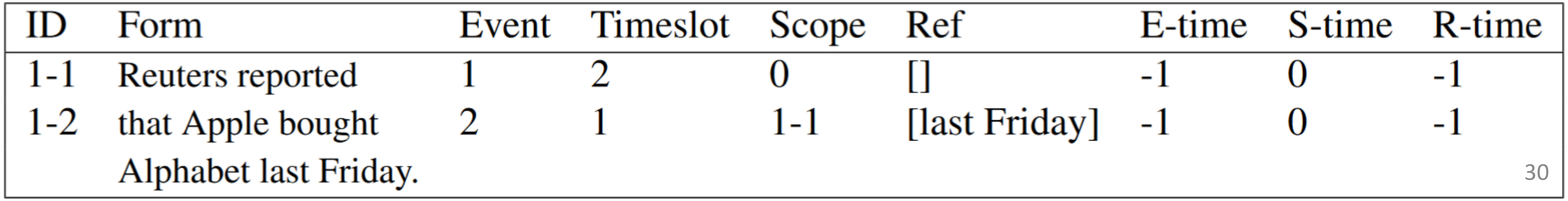
- Annotation effort
  - Complexity & Time
  - Errors
- Complexity of annotation strategy
  - XML-based system with numerous tags and attributes, with complex relation to other entities and elements
  - Education on semantics necessary
- Breaking down:
  - What discourse properties in language get affected?
    - Reference time, Speaker time, Event time

Approaches

- NLP technologies for labeling
  - Event variables and references
    - John called Mary. This upset Susan.
  - Temporal annotation
    - Periphrastic tense
    - Scope effects and contextual variation
- Parsing data sets
  - Manuals, reports
  - Medical

TIE-ML Schema

- Clause level labelling
  - Tense properties : Event, speaker, and reference times (Reichenbach 1947)
  - Temporal scope relations and reference
- Sentence level labelling
  - Event sequencing and duration
- Annotation implementation using INCePTION:
  - CoNLL (tsv) format



ID	Form	Event	Timeslot	Scope	Ref	E-time	S-time	R-time
1-1	Reuters reported	1	2	0	[]	-1	0	-1
1-2	that Apple bought Alphabet last Friday.	2	1	1-1	[last Friday]	-1	0	-1

Results

- Models for NLP of tense and event labeling
- Data sets covering numerous languages
- Annotation tools and data processing environments
- Graph-based models of events and temporal unfolding

Availability

- The corpora, samples, and scripts are made available at the public TIE-ML GitHub repository:

  https://github.com/dcavar/tieml

- More documentation and information about the project can be found at the website of the NLP-Lab:
  - https://nlp-lab.org/timeevents/