Event and Temporal Relation Extraction from Croatian Newspaper Texts

Mladen Marović, Jan Šnajder, Goran Glavaš

Eighth Language Technologies Conference (LTC)
Ljubljana, October 8th, 2012

Motivation and goal

take[lab];

- **Event extraction** (EE) and **temporal relation extraction** (TRE) – non-trivial information extraction tasks
- Important role in various NLP applications
- EE: event identification or event classification
- TRE: classification of temporal relations between extracted pairs of events

Our goal

Develop and evaluate **EE** and **TRE** from Croatian newspaper text using **supervised machine learning** with simple features

Motivation and goal

take[lab];

- **Event extraction** (EE) and **temporal relation extraction** (TRE) – non-trivial information extraction tasks
- Important role in various NLP applications
- EE: event identification or event classification
- TRE: classification of temporal relations between extracted pairs of events

Our goal

Develop and evaluate **EE** and **TRE** from Croatian newspaper text using **supervised machine learning** with simple features

Outline

take[lab];

- Sketch of related work
- Corpus annotation
- Models and features
- Results
- Conclusion

Related work – event extraction

take[lab];

- [Vendler, 1957] – *states*, *activities*, *accomplishments*, *achievements*
- [Pustejovsky, 1991] – structural event hierarchy
- [Siegel & McKeown, 2000] – machine learning for determining the aspectual properties of verbs
- [Pustejovsky et al., 2003a] – *TimeML*, eight classes of events
- [Pustejovsky et al., 2003b] – *TimeBank* manually annotated for events
  - [Saurí et al., 2005], [Boguraev & Ando, 2005], [Bethard & Martin, 2006]

Related work – temporal relations

take[lab];

- [Allen, 1983] – interval temporal algebra
- [Pustejovsky et al., 2003a] – *TimeML*, eight labels for relations
- [Pustejovsky et al., 2003b] – *TimeBank* annotated with low inter-annotator agreement
- [Mani et al., 2006] – expanded TimeBank using a temporal closure algorithm
- [Lapata & Lascarides, 2004, Lapata & Lascarides, 2006] – probabilistic models for inserting temporal connectives (*during*, *after*, . . .) into sentences
- [Verhagen et al., 2007], and [Verhagen et al., 2010]– *TempEval* and *TempEval-2* evaluation exercises

Corpus

take[lab];

- 230 newspaper articles from the Croatian newspaper *Vjesnik* spanning years 1999—2009
- Avg. article length: 500 tokens (including words and punctuation marks)
- Topics: daily news, sports, politics, and culture
- Opinionated text (i.e., columns, reviews) was not considered
- 102,830 words, 26,095 word-form types, 10,963 lemma types

Events and event classes

- Because we focus on **news corpora**, we introduce three modifications to TimeML guidelines:
    1. only **realis** events
    2. no generic events
    3. no states (only state changes)
- Events considered on a single word basis
- 7 event classes

Event annotation summary

take[lab];

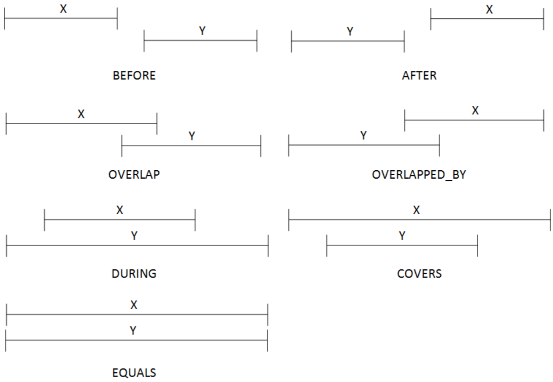
- 5 annotators
- Inter-annotator agreement: $F_1 = 0.7951$

Event Class	Frequency	IAA
OCCURRENCE	6,867	0.6537
REPORTING	1,303	0.8207
I_ACTION	1,124	0.3341
HALF_GENERIC	642	0.2080
STATE_CHANGE	348	0.2349
ASPECTUAL	301	0.4272
PERCEPTION	58	0.3383
Total	10,643	

Temporal relations

take[lab];

- Based on Allen's relation, but with some conflations
- 8 relations



Temporal relation annotation summary

- 4 annotators
- Inter-annotator agreement: $\kappa = 0.5855$

Relation Type	Frequency	IAA
BEFORE	4,860	0.7660
AFTER	3,500	0.8676
EQUALS	1,880	0.4968
COVERS	1,597	0.5847
DURING	1,341	0.5775
NON-DETERMINABLE	763	0.1813
OVERLAP	46	0.0000
OVERLAPPED_BY	24	0.0833
Total	14,011	

Classifiers and features

take[lab];

- Naive Bayes (NB), k-nearest neighbors (k-NN) with $k = 3$, and support vector machine (SVM)
- Baseline: word-conditioned majority class (events) and majority class (temporal relations)
- Event extraction features:
  - word, lemma, stem, POS tag, case, number, modality, auxiliary words, verb form, verb valence class (from CROVALLEX), negation, surrounding words
- Temporal relation extraction features:
  - word, lemma, stem, POS tag, modality, auxiliary words, verb valence class (from CROVALLEX), event class, binary feature vector for words between events

Experimental evaluation

take[lab];

- Event extraction: two experiments
  - binary classification (event identification)
  - multiclass classification (event classification)
- Temporal relation extraction: classification of relations between all pairs of events within the same sentence
- Performance estimates: ten-fold cross-validation
- Results: macro-averaged F_1 scores averaged over ten folds

Results – event extraction

	Baseline	NB	k-NN	SVM
Event identification	5.82 ± 0.69	68.85 ± 0.63	71.07 ± 1.31	$\mathbf{77.40 \pm 0.80}$
Event classification	0.84 ± 0.20	33.56 ± 1.27	43.63 ± 2.93	$\mathbf{48.04 \pm 3.21}$
OCCURRENCE	3.10 ± 0.76	55.40 ± 2.04	53.33 ± 2.14	$\mathbf{62.34 \pm 1.23}$
REPORTING	0.69 ± 0.87	75.28 ± 1.09	76.44 ± 3.32	$\mathbf{79.91 \pm 2.36}$
ASPECTUAL	0.39 ± 1.24	12.25 ± 2.31	58.42 ± 7.37	$\mathbf{59.21 \pm 5.48}$
PERCEPTION	0.00 ± 0.00	24.66 ± 8.65	50.80 ± 15.64	$\mathbf{56.32 \pm 18.18}$
I_ACTION	0.99 ± 0.90	$\mathbf{28.63 \pm 1.88}$	24.21 ± 4.26	24.28 ± 2.40
STATE_CHANGE	0.72 ± 0.93	$\mathbf{25.11 \pm 3.62}$	23.18 ± 8.04	23.17 ± 6.49
HALF_GENERIC	0.00 ± 0.00	13.59 ± 2.11	18.99 ± 5.70	$\mathbf{31.04 \pm 6.15}$

Discussion – event extraction

take[lab];

- Comparison difficult due to differences in comparison schemes
- [Saurí et al., 2005]
  - Slightly higher event identification results (80%), based on word chunking, whereas we consider only single words
  - Much better results for event classification (86%), but with a different set of event classes, unfair performance estimate [Bethard & Martin, 2006]
- [Bethard & Martin, 2006] – F_1 scores of up to 76% for event identification, 58% for event classification
- [Verhagen et al., 2010] – F_1 scores for Spanish and English – 88% and 80% for event identification, 66% and 79% for event classification

Results – temporal relations

take[lab];

	Baseline	Bayes	k-NN	SVM
Temp. relation classif.	6.44 ± 0.01	38.77 ± 1.87	32.17 ± 1.86	$\mathbf{51.16 \pm 2.94}$
BEFORE	51.43 ± 0.05	63.63 ± 1.74	59.61 ± 3.47	$\mathbf{73.12 \pm 0.85}$
AFTER	—	59.35 ± 2.18	56.16 ± 3.83	$\mathbf{71.08 \pm 1.46}$
OVERLAP	—	$11,88 \pm 11.70$	0.00 ± 0.00	$\mathbf{32.07 \pm 19.44}$
OVERLAPPED_BY	—	2.64 ± 2.37	0.00 ± 0.00	$\mathbf{20.67 \pm 23.82}$
DURING	—	55.59 ± 3.14	46.16 ± 4.26	$\mathbf{60.41 \pm 2.89}$
COVERS	—	36.51 ± 2.52	24.49 ± 3.72	$\mathbf{50.83 \pm 3.49}$
EQUALS	—	36.91 ± 3.54	33.87 ± 2.30	$\mathbf{46.01 \pm 3.43}$
NON-DETERM.	—	43.63 ± 3.71	37.05 ± 7.22	$\mathbf{55.11 \pm 8.20}$

Discussion – temporal relations

- Comparison difficult due to differences in relation types and the pairs of events considered
- [Verhagen et al., 2010] – F_1 scores of 58% and 66% for the relevant temporal relation extraction tasks
- Higher results are expected because only specific event pairs are considered

Conclusion

take[lab];

- We've addressed EE and TRE for Croatian
- F_1 scores of 77% for event identification, 48% for event classification, and 51% for temporal relation classification
- Difficult to compare to work of others, but satisfactory given the simplicity of features
- We believe results are indicative for other Slavic languages
- Future work
  - a more detailed analysis of the annotation scheme and guidelines
  - use of more sophisticated features (syntactic functions)
  - relating events to normalized TIMEXes

Conclusion

take[lab];

- We've addressed EE and TRE for Croatian
- F_1 scores of 77% for event identification, 48% for event classification, and 51% for temporal relation classification
- Difficult to compare to work of others, but satisfactory given the simplicity of features
- We believe results are indicative for other Slavic languages
- Future work
  - a more detailed analysis of the annotation scheme and guidelines
  - use of more sophisticated features (syntactic functions)
  - relating events to normalized TIMEXes

References

take[lab];

Allen, J. (1983).
Maintaining knowledge about temporal intervals.
Communications of the ACM, 26(11), 832–843.

Bethard, S. & Martin, J. (2006).
Identification of event mentions and their semantic class.
In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing* (pp. 146–154).: Association for Computational Linguistics.

Boguraev, B. & Ando, R. (2005).
Timebank-driven TimeML analysis.
Annotating, Extracting and Reasoning about Time and Events, (05151).

Lapata, M. & Lascarides, A. (2004).
Inferring sentence-internal temporal relations.
In *Proceedings of HLT-NAACL* (pp. 153–160).

Lapata, M. & Lascarides, A. (2006).
Learning sentence-internal temporal relations.
Journal of Artificial Intelligence Research, 27(1), 85–117.

Mani, I., Verhagen, M., Wellner, B., Lee, C., & Pustejovsky, J. (2006).
Machine learning of temporal relations.
In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics* (pp. 753–760).: Association for Computational Linguistics.

Pustejovsky, J. (1991).
The syntax of event structure.
Cognition, 41(1), 47–81.

Pustejovsky, J., Castano, J., Ingria, R., Sauri, R., Gaizauskas, R., Setzer, A., Katz, G., & Radev, D. (2003a).
TimeML: Robust specification of event and temporal expressions in text.
New Directions in Question Answering, 2003, 28–34.

References

Pustejovsky, J., Hanks, P., Sauri, R., See, A., Gaizauskas, R., Setzer, A., Radev, D., Sundheim, B., Day, D., Ferro, L., et al. (2003b).
The TimeBank corpus.
In *Corpus Linguistics*, volume 2003 (pp.4̃0).

Saurí, R., Knippen, R., Verhagen, M., & Pustejovsky, J. (2005).
Evita: a robust event recognizer for QA systems.
In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing* (pp. 700–707).: Association for Computational Linguistics.

Siegel, E. & McKeown, K. (2000).
Learning methods to combine linguistic indicators: Improving aspectual classification and revealing linguistic insights.
Computational Linguistics, 26(4), 595–628.

Vendler, Z. (1957).
Verbs and times.
The philosophical review, 66(2), 143–160.

Verhagen, M., Gaizauskas, R., Schilder, F., Hepple, M., Katz, G., & Pustejovsky, J. (2007).
Semeval-2007 task 15: Tempeval temporal relation identification.
In *Proceedings of the 4th International Workshop on Semantic Evaluations* (pp. 75–80).: Association for Computational Linguistics.

Verhagen, M., Sauri, R., Caselli, T., & Pustejovsky, J. (2010).
Semeval-2010 task 13: Tempeval-2.
In *Proceedings of the 5th International Workshop on Semantic Evaluation* (pp. 57–62).: Association for Computational Linguistics.

Thanks for your attention!