Constructing Coherent Event Hierarchies from News Stories

Goran Glavaš and **Jan Šnajder**
University of Zagreb, Faculty of Electrical Engineering and Computing
Text Analysis and Knowledge Engineering Lab
Unska 3, 10000 Zagreb, Croatia
{goran.glavas,jan.snajder}@fer.hr

Abstract

News describe real-world events of varying granularity, and recognition of internal structure of events is important for automated reasoning over events. We propose an approach for constructing coherent event hierarchies from news by enforcing document-level coherence over pairwise decisions of spatiotemporal containment. Evaluation on a news corpus annotated with event hierarchies shows that enforcing global spatiotemporal coreference of events leads to significant improvements (7.6% F_1-score) in the accuracy of pairwise decisions.

1 Introduction

Although real-world events have exact spatiotemporal extent, event mentions in text are often spatially and temporally vague. Moreover, event mentions typically denote real-world events of varying granularity (e.g., *summit* vs. *conversation*). If not addressed, these issues hinder event-based inference.

Research efforts in event extraction have focused on either extracting temporal relations (Pustejovsky et al., 2003a; UzZaman et al., 2013) or recognizing spatial relations (Mani et al., 2010; Roberts et al., 2013) between events. Apart from being difficult to recognize, temporal and spatial containment – when considered in isolation – do not suffice to infer that one event is a part of another. Temporally, an event may happen *during* another event and not be a part of it, as in (1).

(1) *In the midst of the World **War** II, the Argentinian government **reduced** rents.*

In this case, *"the reduction of rents in Argentina"* happened *during "the World War II,"* but was not part of it. Conversely, an event may occur *within* the spatial extent of another event and not be a part of it, as shown by (2).

(2) *The fire **destroyed** 60% of London after almost 30,000 people died from **plague**.*

The spatial extent of *"destruction by fire"* is contained *within* the extent of *"people dying from plague,"* but the former is not a part of the latter. An event e_1 is a part of event e_2 if and only if e_1 is spatially *and* temporally contained within e_2.

In previous research (Chambers and Jurafsky, 2008; Jans et al., 2012), news narratives were modeled as *chains* of events involving the same participants. Such script-like representations, however, do not account for the non-linear (hierarchical) nature of events. In contrast, in this work we model the structure of events in a narrative via relations of *spatiotemporal containment* (STC) between event mentions, effectively inducing a hierarchy of events. We construct directed acyclic graphs of event mentions, in which edges denote STC relations between events. We call this structure an *event hierarchy directed acyclic graph* (EHDAG).

We propose a two-step approach for constructing EHDAGs from news. We first detect the STC relations between pairs of event mentions in a supervised fashion, building on our previous approach (Glavaš et al., 2014). We then enforce structural coherence over local predictions, framing the task as a constrained optimization problem, which we solve using Integer Linear Programming (ILP).

2 Related Research

Introduction of the TimeML standard (Pustejovsky et al., 2003a) and the TimeBank corpus (Pustejovsky et al., 2003b) triggered a surge of research on extraction of temporal relations, much of which within TempEval campaigns (Verhagen et al., 2010; UzZaman et al., 2013). More recently, following the emergence of the SpatialML standard (Mani et al., 2010), Roberts et al. (2013) have proposed an annotation scheme and the supervised model for extracting spatial relations between events.

The abovementioned approaches, however, do

Proceedings of TextGraphs-9: the workshop on Graph-based Methods for Natural Language Processing, pages 34–38,
October 29, 2014, Doha, Qatar. ©2014 Association for Computational Linguistics

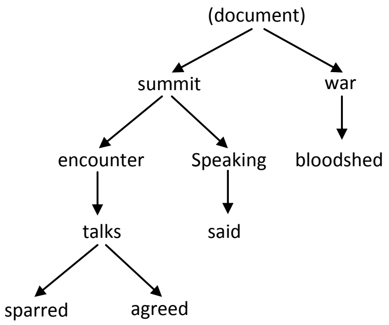

Figure 1: An example of an EHDAG for a narrative

not account for global narrative coherence. Chambers and Jurafsky (2008) consider narratives to be chains of temporally ordered events linked by a common protagonist. Limiting a narrative to a sequence of protagonist-sharing events can often be overly restrictive. E.g., an *"encounter between Merkel and Holland"* may belong to the same "summit" narrative as a *"meeting between Obama and Putin,"* although they share no protagonists.

Several approaches enforce coherence of temporal relations at a document level. Bramsen et al. (2006) represent the temporal structure of a document as a DAG in which vertices denote textual segments and edges temporal precedence. Similarly, Do et al. (2012) enforce coherence using ILP for joint inference on decisions from local event–event and event–time interval decisions.

Complementary to Chambers and Jurafsky (2008), who use a linear temporal structure, with EHDAGs we model the hierarchical structure of events with diverse participants. Similarly to Bramsen et al. (2006), we use an ILP formulation of global coherence over local decisions, but consider STC relations between events rather than temporal relations between textual segments.

3 Constructing Coherent Hierarchies

As an example, consider the following news snippet, with the corresponding EHDAG shown in Fig. 1:

(3) *Obama* ***sparred*** *with Vladimir Putin over how to end the* ***war*** *in Syria on Monday during an icy* ***encounter*** *at a G8* ***summit****.* ***Speaking*** *after* ***talks*** *with Obama, Putin* ***said*** *they* ***agreed*** *the* ***bloodshed*** *must end...*

We first use a supervised classifier to determine the STC relations between all pairs of events in a document. In the second step, we induce a spatiotemporally coherent EHDAG by enforcing coherence constraints on the local classification decisions.

3.1 Spatiotemporal Containment Classifier

We first describe the classifier used for predicting local STC relations. The classifier is given a pair of event mentions, (e_1, e_2), where mention e_1 occurs in text before mention e_2. The classifier predicts one of the following relations: (1) e_1 SUBSUPER e_2, denoting that the e_1 (*subevent*) is spatiotemporally contained by event e_2 (*superevent*); (2) e_1 SUPERSUB e_2, denoting that e_1 (*superevent*) spatiotemporally contains e_2 (*subevent*); and (3) NOREL, denoting that neither of the two events spatiotemporally contains the other. We use the following rich set of features for the STC relation classifier.

Event-based features: Word, lemma, stem, POS-tag, and TimeML type of both event mentions. Additionally, we compare the event arguments of three semantic types: AGENT, TARGET, and LOCATION, which we extract automatically from raw text using the rule-based model by Glavaš and Šnajder (2013).

Bag-of-words features: All lemmas in between the two event mentions, with the special status being assigned to temporal signals (e.g., *before*) and spatial signals (e.g., *inside*).

Positional features: The distance between event mentions in the document, both in the number of sentences and the number of tokens. Additionally, we use a feature indicating if the two mentions are adjacent (no mentions occur in between).

Syntactic features: All dependency relations on the path between events in the dependency tree and features that indicate whether one of the features syntactically governs the other. We compute the syntactic features only for pairs of event mentions from the same sentence, using the Stanford dependency parser (De Marneffe et al., 2006).

Knowledge-based features: Computed using WordNet (Fellbaum, 1998), VerbOcean (Chklovski and Pantel, 2004), and CatVar (Habash and Dorr, 2003). We use a feature indicating whether one event mention or any of its derivatives (obtained from CatVar) is a WordNet hypernym of (for nominalized mentions) or entailed from (for verb mentions) the other mention (or any of its derivatives). We use an additional feature to indicate the VerbOcean relation between the event mentions, if such exists. Unlike features from previous groups, knowledge-based features have not been used often for temporal relation classification.

We employ a L2-regularized logistic regression as our pairwise classification model, which is motivated by the high-dimensional feature space spanned by the lexical features Moreover, the global coherence component of the model requires probability distributions for local decisions over relation types. We use the LibLinear (Fan et al., 2008) implementation of logistic regression.

3.2 Global Coherence

The hierarchy of events induced from the independent pairwise STC decisions may be globally incoherent. We therefore need to optimize the set of pairwise STC classifications with respect to the set of constraints that enforce global coherence. We perform exact inference using Integer Linear Programming (ILP), an approach that has been proven useful in many NLP applications (Punyakanok et al., 2004; Roth and Yih, 2007; Clarke and Lapata, 2008). We use the *lp_solve*[1] solver to optimize the objective function with respect to the constraints.

Objective function. Let $M = \{e_1, e_2, \ldots, e_n\}$ be the set of all event mentions in the news story and P be the set of all considered pairs of event mentions, $P = \{(e_i, e_j) \mid e_i, e_j \in M, i < j\}$. Let $R = \{\textsc{SuperSub}, \textsc{SubSuper}, \textsc{NoRel}\}$ be the set of spatiotemporal relation types and let $C(e_i, e_j, r)$ be the probability, produced by the pairwise classifier, of relation r holding between event mentions e_i and e_j. We maximize the sum of local probabilities assigned to all pairs of events (summed over all relation types):

$$\sum_{(e_i,e_j)\in P} \sum_{r\in R} C(e_i, e_j, r) \cdot x_{e_i,e_j,r} \qquad (1)$$

where $x_{e_i,e_j,r}$ is a binary indicator variable that takes the value 1 iff the relation of type r is predicted to hold between events e_i and e_j.

Spatiotemporal constraints. The objective function is a subject to two basic constraints: (i) the constraint that declares $x_{e_i,e_j,r}$ to be binary indicator variables (eq. 2) and (ii) the exclusivity constraint, which allows only one relation to hold between two events (eq. 3).

$$x_{e_i,e_j,r} \in \{0,1\}, \quad \forall (e_i, e_j) \in P,\ r \in R \qquad (2)$$

$$\sum_{r\in R} x_{e_i,e_j,r} = 1, \quad \forall (e_i, e_j) \in P \qquad (3)$$

Following the work of Bramsen et al. (2006) and Do et al. (2012), we also incorporate the transitivity constraints into the model (transitivity is not enforced for NoRel):

$$x_{e_i,e_j,r} + x_{e_j,e_k,r} - 1 \leq x_{e_i,e_k,r}, \qquad (4)$$
$$\forall r \in R, \{(e_i, e_j), (e_j, e_k), (e_i, e_k)\} \subseteq P$$

The transitivity constraint states that, if the same relation r holds for pairs of events (e_i, e_j) and (e_j, e_k), then r must also hold for the pair (e_i, e_k).

Coreference constraints. The constraints presented so far did not consider the coreference of event mentions. However, a truly coherent event structure must account for the different mentions of the same event. More precisely, two different constraints have to be enforced: (i) a pair of coreferent event mentions can only be assigned relation of the NoRel type because coreferent event mentions cannot be part of each other (eq. 5) and (ii) all coreferent mentions of one event must be in the same relation with all coreferent mentions of the other event (eqs. 6–9). Let $\mathit{coref}(e_i, e_j)$ be a predicate that holds iff mentions e_1 and e_2 corefer. The coreference constraints are as follows:

$$x_{e_i,e_j,r} = 1, \qquad (5)$$
$$\forall (e_i, e_j) \in P, r = \textsc{NoRel}, \mathit{coref}(e_i, e_j)$$

$$x_{e_i,e_k,r} - x_{e_j,e_k,r} = 0, \qquad (6)$$
$$\forall (e_i, e_k), (e_j, e_k) \in P, r \in R, \mathit{coref}(e_i, e_j)$$

$$x_{e_i,e_k,r} - x_{e_k,e_j,r^{-1}} = 0, \qquad (7)$$
$$\forall (e_i, e_k), (e_k, e_j) \in P, r \in R, \mathit{coref}(e_i, e_j)$$

$$x_{e_k,e_i,r} - x_{e_j,e_k,r^{-1}} = 0, \qquad (8)$$
$$\forall (e_k, e_i), (e_j, e_k) \in P, r \in R, \mathit{coref}(e_i, e_j)$$

$$x_{e_k,e_i,r} - x_{e_k,e_j,r} = 0, \qquad (9)$$
$$\forall (e_k, e_i), (e_k, e_j) \in P, r \in R, \mathit{coref}(e_i, e_j)$$

In equations (7) and (8), the relation type r^{-1} denotes the inverse of the relation type r. The inverse of SuperSub is SubSuper (and vice versa), whereas NoRel is an inverse to itself.

4 Evaluation

We evaluate several models on the publicly available HiEve corpus (Glavaš et al., 2014), consisting of 100 news stories manually annotated with event hierarchies.

[1] http://lpsolve.sourceforge.net/5.5/

	SuperSub			SubSuper			Micro-averaged		
Model	P	R	F_1	P	R	F_1	P	R	F_1
Memorize baseline	60.3	30.2	40.2	66.8	36.7	47.4	63.8	33.5	43.9
Pairwise-NoKB	58.4	47.2	52.2	72.8	**56.2**	63.4	65.5	51.8	57.8
Pairwise-Full	69.8	51.2	59.1	70.6	54.1	61.3	70.2	52.6	60.1
Coherent	79.6	**60.6**	68.6	73.0	52.0	60.8	76.6	**56.5**	65.0
Coref-Auto	79.5	57.6	66.8	73.0	52.0	60.8	76.3	55.0	63.9
Coref-Gold	**87.2**	58.8	**70.3**	**84.2**	52.7	**64.8**	**85.8**	55.9	**67.7**

Table 1: Model performance for recognizing spatiotemporal containment between events

4.1 Experimental Setup

We leave out 20 news stories from the HiEve corpus for testing and use the remaining 80 documents for training the pairwise STC classifiers. Altogether, we evaluate the following five models.

Pairwise model employs only the pairwise classification and does not enforce coherence across local decisions. We evaluate two classifiers: one with knowledge-based features (Pairwise-Full) and one without (Pairwise-NoKB).

Coherent model enforces document-level spatiotemporal coherence by solving the constrained optimization problem on top of pairwise classification decisions. The model uses the constraints from (2)–(4), but not the coreference-based constraints.

Coref-Gold model uses coreference constraints (6)–(9) in addition to constraints (2)–(4). The model uses hand-annotated coreference relations from the HiEve corpus.

Coref-Auto model uses the same set of constraints as the previous model, but relies on the event coreference resolution model by Glavaš and Šnajder (2013) instead on gold annotations.

As the baseline, we use the Memorize model, which simply assigns to each pair of event mentions in the test set their most frequent label in the training set. The NoRel label is predicted for the pairs of lemmas not observed in the training set. A similar baseline has been proposed by Bethard (2008) for automated extraction of event mentions.

To account for the transitivity of the STC relation, we evaluate the predictions of our models against the transitive closure of gold STC hierarchies from the HiEve corpus.

4.2 Results

Table 1 summarizes the results. We show the performance (precision, recall, and F_1-score) for the SuperSub and SubSuper relations along with the micro-averaged performance. All models significantly outperform the Memorize baseline (with the exception of Pairwise-NoKB's precision), which has been shown competitive on the event extraction task (Bethard, 2008). Overall, the Pairwise-Full model outperforms the Pairwise-NoKB model, confirming the intuition that knowledge-based information is useful for detecting relations between events. However, including KB features decreases the performance on the SubSuper class, which requires further analysis.

Comparison of the Pairwise models and the Coherent model reveals that enforcing global coherence of local relations substantially improves the quality of the constructed hierarchies (4.9% F_1-score; significant at $p<0.01$ using stratified shuffling (Yeh, 2000)). With the introduction of additional reference constraints (model Coref-Gold), the quality improves by additional 2.7% F_1-score (significant at $p<0.05$). The fact that the model Coref-Auto is outperformed by the Coherent model, however, suggests that the automated coreference resolution model is not accurate enough to benefit the global coherence constraints.

5 Conclusion

We addressed the task of constructing coherent event hierarchies based on recognition of spatiotemporal containment between events from their mentions in text. The proposed approach constructs event hierarchies by enforcing document-level coherence over a set of local decisions on spatiotemporal containment between events. The quality of the extracted event hierarchies is improved by enforcing global coherence, and can be improved even further using event coreference-based constraints, provided accurate coreference resolution is available. Our next step will be to incorporate predictions from state-of-the-art temporal and spatial relation extraction models, both as STC classifier features and as additional optimization constraints.

References

S. Bethard. 2008. *Finding Event, Temporal and Causal Structure in Text: A Machine Learning Approach*. Ph.D. thesis, University of Colorado at Boulder.

P. Bramsen, P. Deshpande, Y. K. Lee, and R. Barzilay. 2006. Inducing temporal graphs. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP '06)*, pages 189–198. ACL.

N. Chambers and D. Jurafsky. 2008. Unsupervised learning of narrative event chains. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics (ACL 2008)*, pages 789–797.

T. Chklovski and P. Pantel. 2004. VerbOcean: Mining the web for fine-grained semantic verb relations. In *In Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP '04)*, pages 33–40.

J. Clarke and M. Lapata. 2008. Global inference for sentence compression: An integer linear programming approach. *Journal of Artificial Intelligence Research (JAIR)*, 31:399–429.

M. C. De Marneffe, B. MacCartney, and C. D. Manning. 2006. Generating typed dependency parses from phrase structure parses. In *Proceedings of 5th International Conference on Language Resources and Evaluation (LREC 2006)*, volume 6, pages 449–454.

Q. X. Do, W. Lu, and D. Roth. 2012. Joint inference for event timeline construction. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 677–687. ACL.

R. E. Fan, K. Chang, C. J. Hsieh, X. R. Wang, and C. J. Lin. 2008. LibLinear: A library for large linear classification. *The Journal of Machine Learning Research*, 9:1871–1874.

C. Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

G. Glavaš and J. Šnajder. 2013. Exploring coreference uncertainty of generically extracted event mentions. In *Proceedings of the Conference in Intelligent Text Processing and Computational Linguistics CICLing 2013*, pages 408–422. Springer.

G. Glavaš, J. Šnajder, P. Kordjamshidi, and M.-F. Moens. 2014. HiEve: A corpus for extracting event hierarchies from news stories. In *Proceedings of 9th Language Resources and Evaluation Conference (LREC 2014)*, pages 3678–3683.

G. Glavaš and J. Šnajder. 2013. Recognizing identical events with graph kernels. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL 2013)*, pages 797–803. Springer.

N. Habash and B. Dorr. 2003. A categorial variation database for English. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 17–23. ACL.

B. Jans, S. Bethard, I. Vulić, and M. F. Moens. 2012. Skip n-grams and ranking functions for predicting script events. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 336–344. ACL.

I. Mani, C. Doran, D. Harris, J. Hitzeman, R. Quimby, J. Richer, B. Wellner, S. Mardis, and S. Clancy. 2010. SpatialML: Annotation scheme, resources, and evaluation. *Language Resources and Evaluation*, 44(3):263–280.

V. Punyakanok, D. Roth, W.-t. Yih, and D. Zimak. 2004. Semantic role labeling via integer linear programming inference. In *Proceedings of the 20th international conference on Computational Linguistics*, page 1346. ACL.

J. Pustejovsky, J. Castano, R. Ingria, R. Sauri, R. Gaizauskas, A. Setzer, G. Katz, and D. Radev. 2003a. TimeML: Robust specification of event and temporal expressions in text. *New Directions in Question Answering*, 3:28–34.

J. Pustejovsky, P. Hanks, R. Sauri, A. See, R. Gaizauskas, A. Setzer, D. Radev, B. Sundheim, D. Day, L. Ferro, et al. 2003b. The TimeBank corpus. In *Corpus Linguistics*, volume 2003, pages 647–656.

K. Roberts, M. A. Skinner, and S. M. Harabagiu. 2013. Recognizing spatial containment relations between event mentions. In *10th International Conference on Computational Semantics*.

D. Roth and W.-t. Yih. 2007. Global inference for entity and relation identification via a linear programming formulation. *Introduction to statistical relational learning*, pages 553–580.

N. UzZaman, H. Llorens, L. Derczynski, M. Verhagen, J. Allen, and J. Pustejovsky. 2013. SemEval-2013 Task 1: TempEval-3: Evaluating time expressions, events, and temporal relations. In *Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval 2013)*. ACL.

M. Verhagen, R. Sauri, T. Caselli, and J. Pustejovsky. 2010. SemEval-2010 Task 13: TempEval-2. In *Proc. of the SemEval 2010*, pages 57–62.

A. Yeh. 2000. More accurate tests for the statistical significance of result differences. In *Proceedings of the 18th conference on Computational linguistics-Volume 2*, pages 947–953. ACL.