Unsupervised Semantic Frame Induction using Triclustering

Dmitry Ustalov[†], Alexander Panchenko[‡], Andrei Kutuzov[⋆],
Chris Biemann[‡], and Simone Paolo Ponzetto[†]

[†] Data and Web Science Group, University of Mannheim, Germany
[‡] Universität Hamburg, Department of Informatics, Language Technology Group, Germany
[⋆] University of Oslo, Norway

Presented at the 56th Annual Meeting of the Association for Computational Linguistics (ACL 2018) in Melbourne, Australia.
Paper: `https://aclweb.org/anthology/P18-2010`.

Summary

- We use dependency triples automatically extracted from a Web-scale corpus to perform unsupervised semantic frame induction.
- We cast the frame induction problem as a *triclustering* problem that is a generalization of clustering for *triadic* data.
- Our replicable benchmarks demonstrate that the proposed graph-based approach, *Triframes*, shows state-of-the-art results on this task on a FrameNet-derived dataset and performs on par with competitive methods on a verb class clustering task.

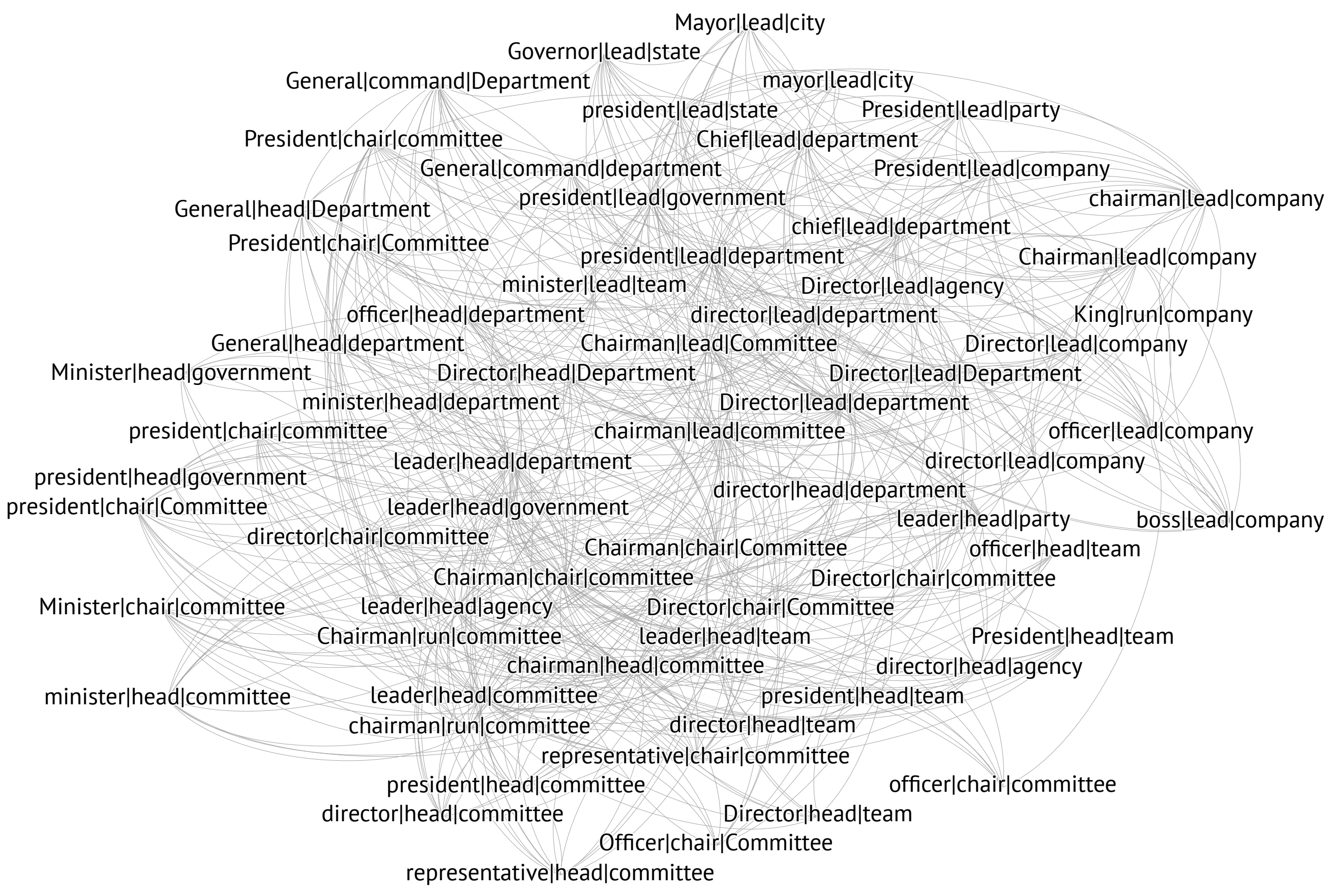
Triframes Algorithm

We use the Watset meta-algorithm by Ustalov et al. (ACL 2017) for fuzzy clustering of the dependency triple graph. Watset creates an intermediate representation of the input graph that naturally reflects the "ambiguity" of its nodes. Then, it uses hard clustering to discover clusters in this intermediate graph.

```
Input:  an embedding model $v \in V \rightarrow \vec{v} \in \mathbb{R}^d$,
        a set of SVO triples $T \subseteq V^3$,
        the number of nearest neighbors $k \in \mathbb{N}$.
Output: a set of triframes $F$.
$S \leftarrow \{t \rightarrow \vec{t} \in \mathbb{R}^{3d} : t \in T\}$
$E \leftarrow \{(t, t') \in T^2 : t' \in \mathrm{NN}_k^S(\vec{t}), t \neq t'\}$
$F \leftarrow \emptyset$
for all $C \in \textsc{Watset}(T, E)$ do
    $f_s \leftarrow \{s \in V : (s, v, o) \in C\}$
    $f_v \leftarrow \{v \in V : (s, v, o) \in C\}$
    $f_o \leftarrow \{o \in V : (s, v, o) \in C\}$
    $F \leftarrow F \cup \{(f_s, f_v, f_o)\}$
end for
return $F$
```

As the input, we use the standard Google News word embeddings and dependency triples from the DepCC corpus (Panchenko et al., LREC 2018).

Triple Relationships within a Triframe Cluster



Frame # 848	
Subjects:	Company, firm, company
Verbs:	buy, supply, discharge, purchase, expect
Objects:	book, supply, house, land, share, company, grain, which, item, product, ticket, work, this, equipment, House, it, film, water, something, she, what, service, plant, time

Frame # 849	
Subjects:	student, scientist, we, pupil, member, company, man, nobody, you, they, US, group, it, people, Man, user, he
Verbs:	do, test, perform, execute, conduct
Objects:	experiment, test

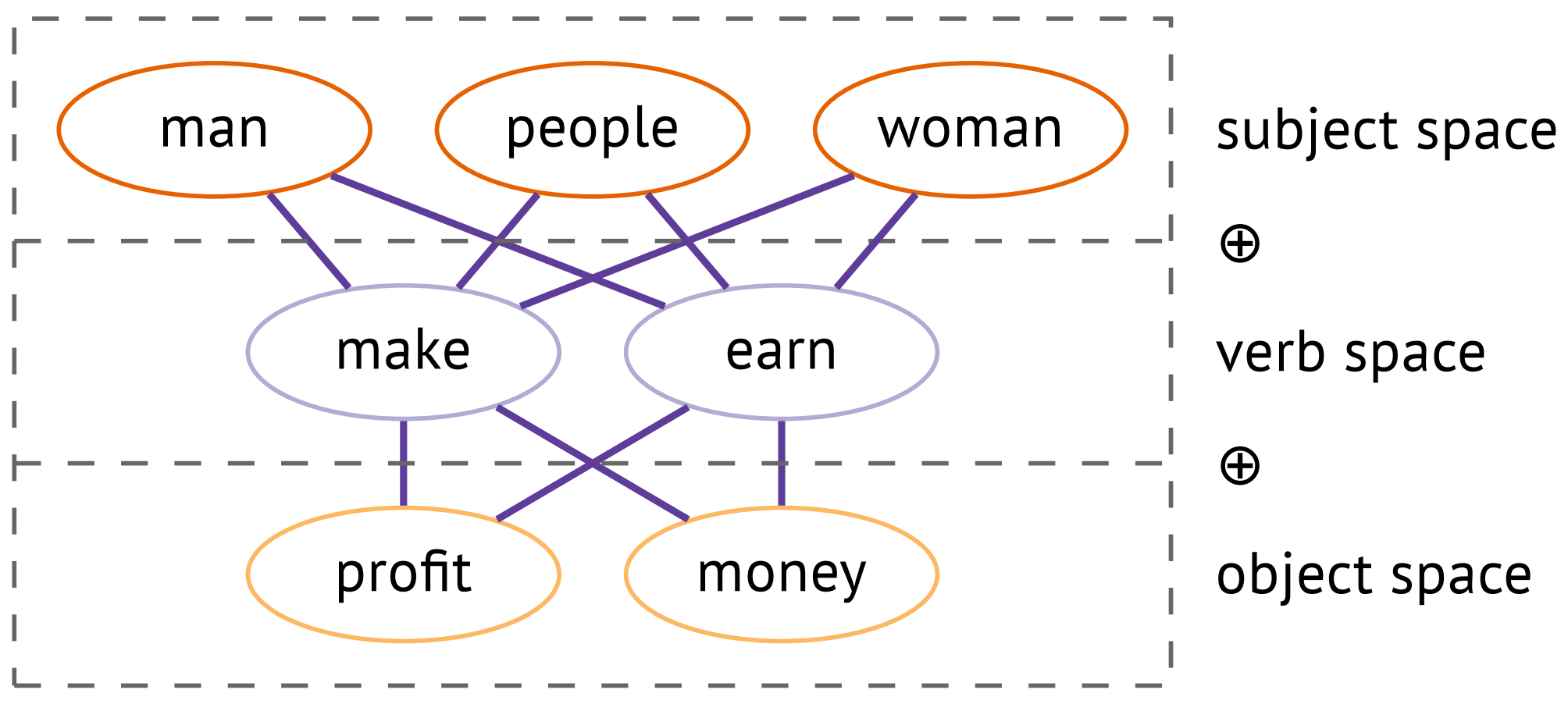
Frame # 3207	
Subjects:	people, we, they, you
Verbs:	feel, seek, look, search
Objects:	housing, inspiration, gold, witness, partner, accommodation, Partner

Evaluation Setup

We use normalized modified purity (nmPU), normalized inverse purity (niPU), and their harmonic mean (F_1) as the evaluation measures.

- In **Verb Classes Evaluation**, we reproduced the experiments by Kawahara et al. (ACL 2014) and compared Triframes to the other approaches only on the polysemous verb classes gold standard dataset by Korhonen et al. (ACL 2003).
- In **Frame Evaluation**, we transform each frame into a set of typed pairs representing frame elements. This allows us to compare frames to each other. As the gold standard, we derived sets of frame elements from the FrameNet-annotated corpus (Bauer et al., LREC 2012).

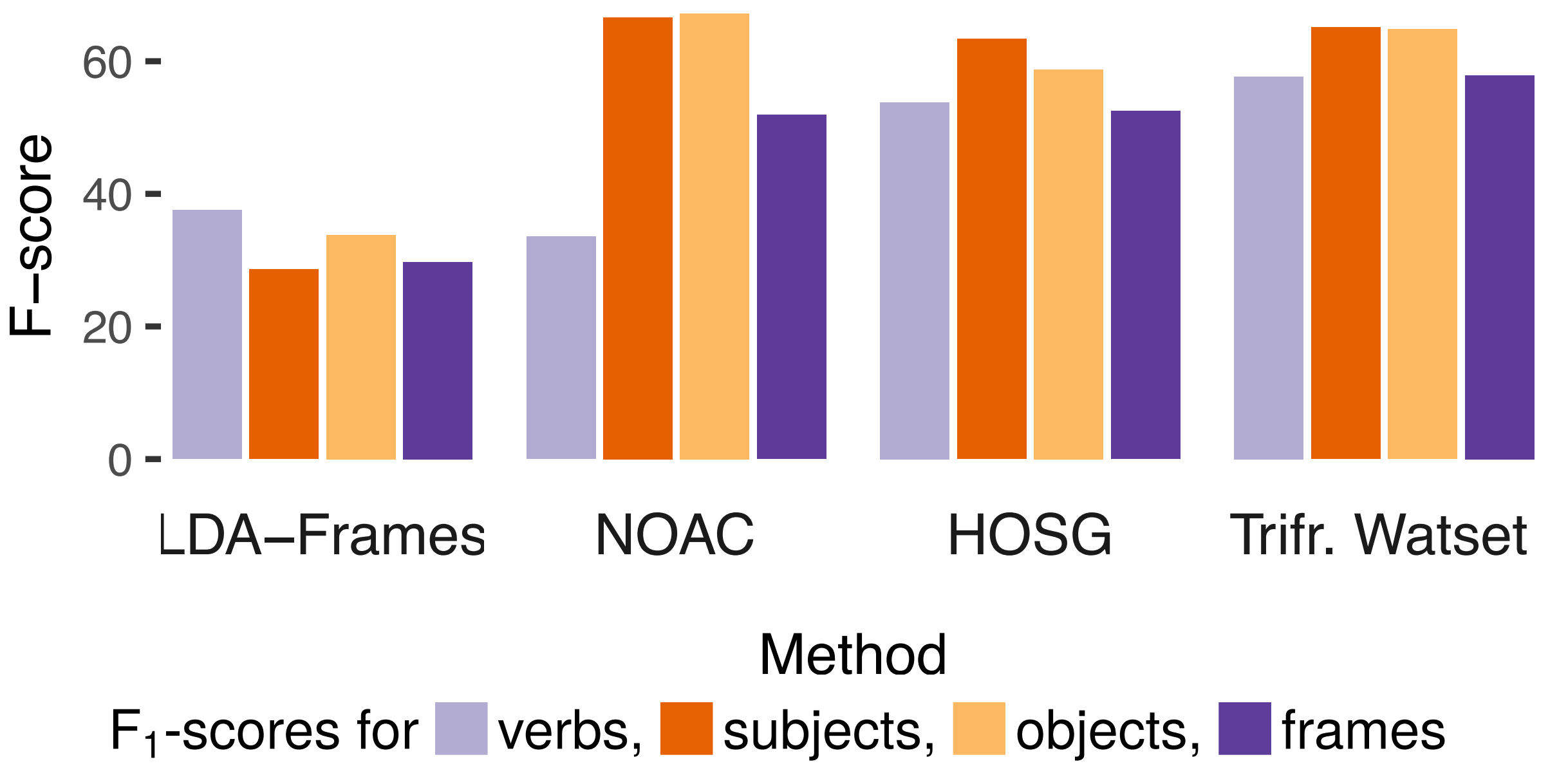
A triple (Freddy : *Predator*, kidnap : *FEE*, kid : *Victim*) is converted to three pairs (Freddy, *Predator*), (kidnap, *FEE*), (kid, *Victim*) during the Frame Evaluation experiment.



Verb Classes Evaluation (Korhonen et al., ACL 2003)

Method	nmPU	niPU	F_1
LDA-Frames	**52.60**	45.84	**48.98**
Triframes Watset	40.05	62.09	48.69
NOAC	37.19	64.09	47.07
HOSG	38.22	43.76	40.80
Triadic Spectral	35.76	38.96	36.86
Triadic k-Means	52.22	27.43	35.96
Triframes CW	18.05	12.72	14.92
Whole	24.14	**79.09**	36.99
Singletons	0.00	27.21	0.00

Verb, Subject, Object, and Frame Evaluation on the FrameNet Corpus (Bauer et al., LREC 2012)



F_1-scores for ■ verbs, ■ subjects, ■ objects, ■ frames

Method	Verb			Subject			Object			Frame		
	nmPU	niPU	F_1	nmPU	niPU	F_1	nmPU	niPU	F_1	nmPU	niPU	F_1
Triframes Watset	42.84	88.35	**57.70**	54.22	81.40	65.09	53.04	83.25	64.80	55.19	60.81	**57.87**
HOSG	44.41	68.43	53.86	52.84	74.53	61.83	54.73	74.05	62.94	55.74	50.45	52.96
NOAC	20.73	88.38	33.58	57.00	80.11	**66.61**	57.32	81.13	**67.18**	44.01	63.21	51.89
Triadic Spectral	49.62	24.90	33.15	50.07	41.07	45.13	50.50	41.82	45.75	52.05	28.60	36.91
Triadic k-Means	**63.87**	23.16	33.99	**63.15**	38.20	47.60	**63.98**	37.43	47.23	**63.64**	24.11	34.97
LDA-Frames	26.11	66.92	37.56	17.28	83.26	28.62	20.80	90.33	33.81	18.80	71.17	29.75
Triframes CW	7.75	6.48	7.06	3.70	14.07	5.86	51.91	76.92	61.99	21.67	26.50	23.84
Singletons	0.00	25.23	0.00	0.00	25.68	0.00	0.00	20.80	0.00	32.34	22.15	26.29
Whole	3.62	**100.0**	6.98	2.41	**98.41**	4.70	2.38	**100.0**	4.64	2.63	**99.55**	5.12

Source Code and Data

`https://github.com/uhh-lt/triframes`

`mailto:dmitry@informatik.uni-mannheim.de`

We acknowledge the support of DFG under the "JOIN-T" and "ACQuA" projects and thank three anonymous reviewers for their helpful comments. Furthermore, we thank Dmitry Egurnov, Dmitry Ignatov, and Dmitry Gnatyshak for help in operating the NOAC method using the multimodal clustering toolbox. Besides, we are grateful to Ryan Cotterell and Adam Poliak for a discussion and an implementation of the HOSG method. Finally, we thank Bonaventura Coppola for discussions and preliminary work on graph-based frame induction.

References

Daniel Bauer et al. 2012. The Dependency-Parsed FrameNet Corpus. In *Proceedings of the Eight International Conference on Language Resources and Evaluation*, LREC 2012, pages 3861–3867, Istanbul, Turkey. European Language Resources Association (ELRA).

Ryan Cotterell et al. 2017. Explaining and Generalizing Skip-Gram through Exponential Family Principal Component Analysis. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 175–181, Valencia, Spain. Association for Computational Linguistics.

Dmitry Egurnov et al. 2017. Mining Triclusters of Similar Values in Triadic Real-Valued Contexts. In *14th International Conference on Formal Concept Analysis - Supplementary Proceedings*, pages 31–47, Rennes, France.

Daisuke Kawahara et al. 2014. A Step-wise Usage-based Method for Inducing Polysemy-aware Verb Classes. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics Volume 1: Long Papers*, ACL 2014, pages 1030–1040, Baltimore, MD, USA. Association for Computational Linguistics.

Anna Korhonen et al. 2003. Clustering Polysemic Subcategorization Frame Distributions Semantically. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, ACL '03, pages 64–71, Sapporo, Japan. Association for Computational Linguistics.

Jiří Materna. 2012. LDA-Frames: An Unsupervised Approach to Generating Semantic Frames. In *Computational Linguistics and Intelligent Text Processing, Proceedings, Part I*, CICLing 2012, pages 376–387, New Delhi, India. Springer Berlin Heidelberg.

Alexander Panchenko et al. 2018. Building a Web-Scale Dependency-Parsed Corpus from Common Crawl. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, LREC 2018, pages 1816–1823, Miyazaki, Japan. European Language Resources Association (ELRA).

Dmitry Ustalov et al. 2017. Watset: Automatic Induction of Synsets from a Graph of Synonyms. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, ACL 2017, pages 1579–1590, Vancouver, Canada. Association for Computational Linguistics.