# A Consolidated Open Knowledge Representation for Multiple Texts

[1]**Rachel Wities**, [1]Vered Shwartz, [1]Gabriel Stanovsky, [1]Meni Adler, [1]Ori Shapira,
[2]Shyam Upadhyay, [2]Dan Roth, [3]Eugenio Martinez Camara, [3]Iryna Gurevych and [1]Ido Dagan
**rachelvov@gmail.com**

[1]Bar-Ilan University, Israel;
[2]University of Illinois at Urbana-Champaign, IL, USA;
[3]Technische Universitat Darmstadt, Germany

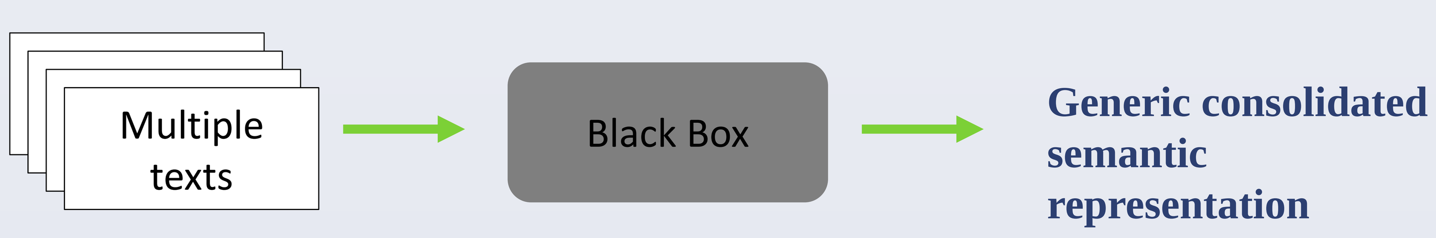
## Our goal – "out of the box" consolidated semantic representation

- Many semantic applications require multiple texts consolidation.
  For example:

  *3 people dead in shooting in Wisconsin.*
  *Man kills three in spa shooting.*
  *Shooter was identified as Radcliffe Haughton, 45.*

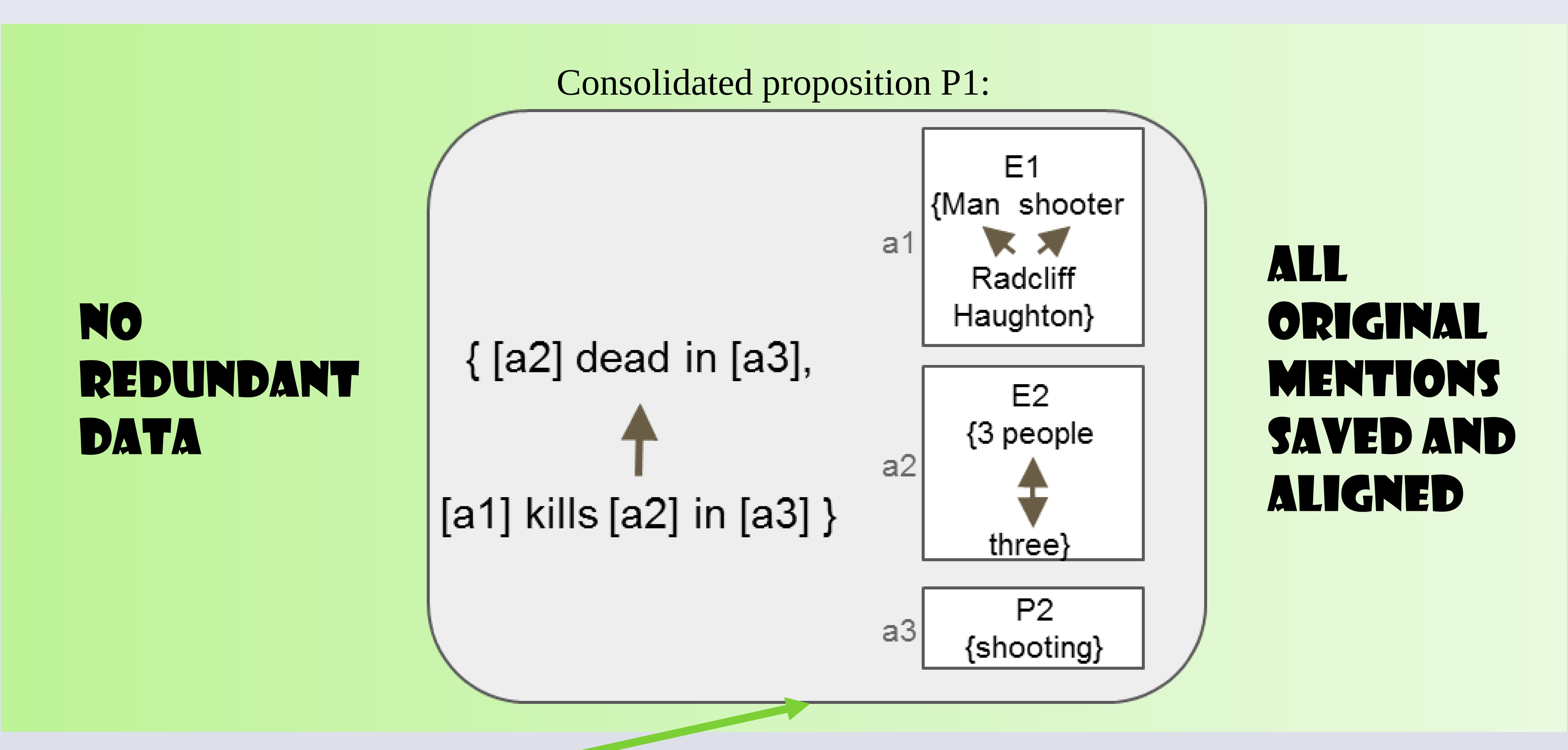
  - Question answering:
    - *How many people did Radcliffe Haughton shoot? 3 people*
  - Absractive summarization:
    - *Radcliffe Haughton, 45, kills three in spa shooting in Wisconsin.*
  - And more…

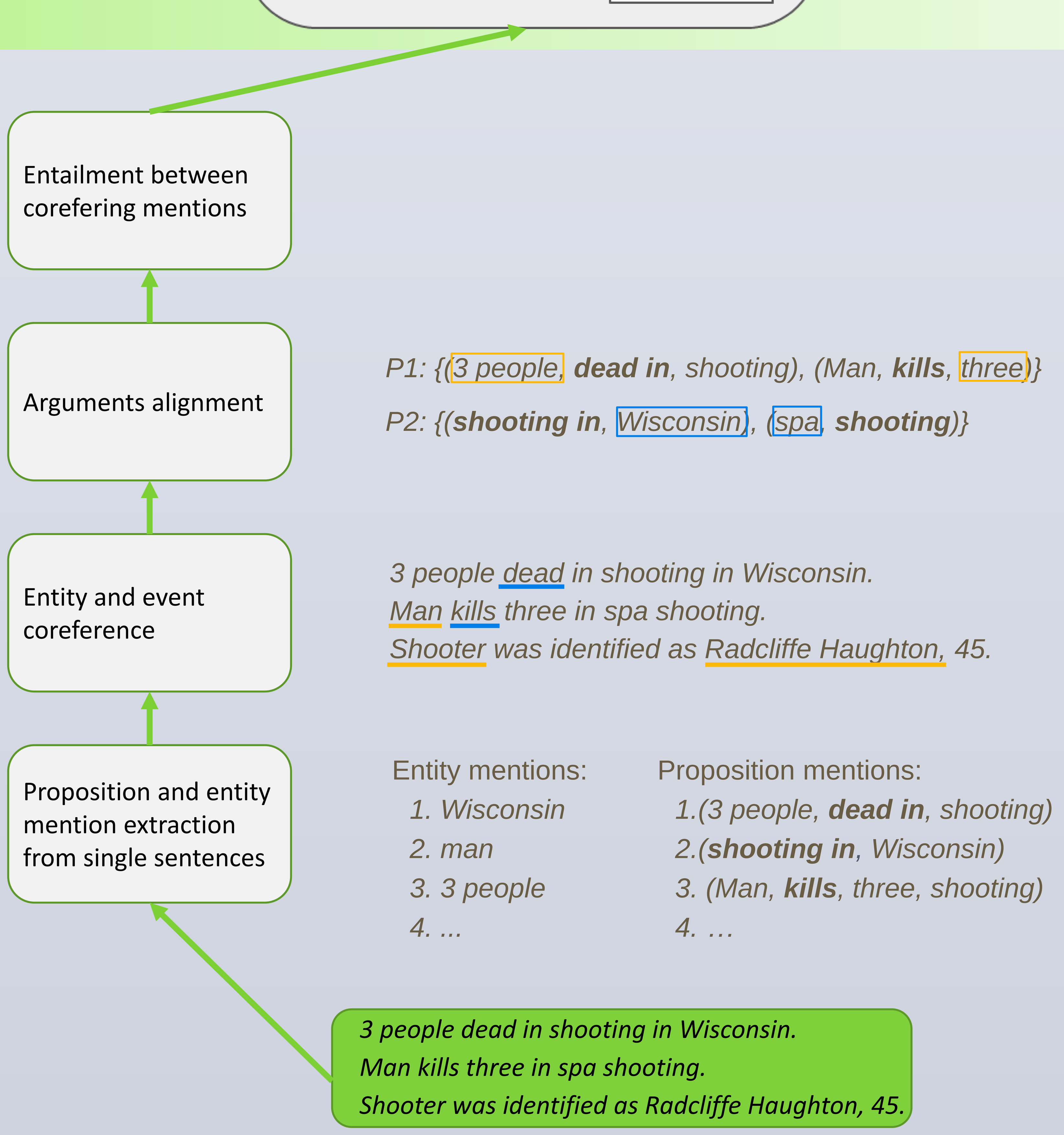
- Current predicate-argument semantic structures (Open IE, SRL, AMR, etc.) are at a single sentence level.
- Therefore, the consolidation is currently done at the application level, to various partial extent.
- We propose: a generic consolidated semantic structure, for all applications!




## OKR representation

- OKR structure is built using a pipeline of well-known NLP tasks:




## Annotated dataset

OKR Annotation of 1257 news-related tweets from 27 event clusters, collected from the Twitter Event Detection Dataset (McMinn et al., 2013). We used QA-SRL paradigm (He et al., 2015) to annotate semantic roles.

- Annotation example:

*Topic #1 tweet #1: Turkey forced a plane to land*
*Topic #1 tweet #2 : The grounded jet landed*

*Entities:*
*E1 : Turkey*
*E2 : jet➔ plane*

*Propositions:*
*P1 : A1 landed⇔A1 to land*
*A1 - what did land? E2*

*P2: grounded A1➔ A2 forced A1 to A3*
*A1 - what was forced to do something?: E2*
*A2 - what did force something to do something?: E1*
*A3 - what was something forced to do?:P1*

**DOWNLOAD OUR ANNOTATED CORPUS!**

http://u.cs.biu.ac.il/~nlp/resources/downloads/twitter-events

## Baseline & results

We created a simple baseline consisting of the separate OKR pipeline tasks, implemented individually:

- **Entity extraction** – spaCy NER and annotating all nouns as entities.
- **Proposition extraction** - Open IE propositions extracted from PropS (Stanovsky et al., 2016).
- **Proposition and Entity coreference** – simple lexical similarity metrics (e.g., lemma matching and Levenshtein distance).
- **Argument alignment** – align all mentions of the same entity.
- **Entity Entailment** - knowledge resources (Shwartz et al., 2015) and a pre-trained model for HypeNET (Shwartz et al., 2016).
- **Predicate Entailment** - rules extracted by Berant et al. (2012).

**Baseline results (compared to inner-annotator agreement):**

| | Entity Extraction *(avg. accuracy)* | Entity Coref. *(CoNNL F1)* | Proposition extraction *(avg. accuracy)* Predicates | | | Proposition extraction Arguments | Predicate coreference *(CoNNL F1)* | Entailment *(F1)* Entities | Entailment Predicates |
|---|---|---|---|---|---|---|---|---|---|
| Agreement | .85 | .90 | .74 | Verb. .93 | Non verb. .72 | .85 | .83 | .70 | .82 |
| Predicted | .58 | .85 | .41 | Verb. .73 | Non verb. .25 | .37 | .56 | .44 | .56 |

- Main challenges:
  – Recognize arguments for nominal predicates - current systems are verb-centric (well known).
  – Distinguish entity nouns from predicate nouns (organization vs. elections).
  – Entity entailment is hard for multi-word expressions.
  – Predicate coreference is harder than entity corefernce.

## References


- Jonathan Berant, Ido Dagan, and Jacob Goldberger. 2012. **Learning entailment relations by global graph structure optimization**. Computational Linguistics, 38(1):73–111.
- Andrew J. McMinn, Yashar Moshfeghi, and Joemon M. Jose. 2013. **Building a large-scale corpus for evaluating event detection on twitter**. In Proceedings of the 22Nd ACM conference, pages 409–418.
- Luheng He, Mike Lewis, and Luke Zettlemoyer. 2015**. Question-answer driven semantic role labeling: Using natural language to annotate natural language**. In Proceedings of the 2015 EMNLP conference, pages 643–653.
- Vered Shwartz and Ido Dagan. 2016**. Path-based vs. distributional information in recognizing lexical semantic relations**. In Proceedings of the 5th Workshop on Cognitive Aspects of the Lexicon (CogALex - V), pages 24–29.
- Vered Shwartz, Omer Levy, Ido Dagan, and Jacob Goldberger. 2015. **Learning to exploit structured resources for lexical inference.** In Proceedings of the Nineteenth CoNLL conference, pages 175–184.
- Gabriel Stanovsky, Jessica Ficler, Ido Dagan, and Yoav Goldberg. 2016**. Getting more out of syntax with PropS**. arXiv preprint.