

# Facts Inat Matter

## Marco Ponza<sup>1</sup>, Luciano Del Corro<sup>2</sup> and Gerhard Weikum<sup>2</sup>

<sup>1</sup>Department of Computer Science, University of Pisa, Italy <sup>2</sup>Max Planck Institute for Informatics, Saarland Informatics Campus, Germany

## **Problem Definition & Contributions**

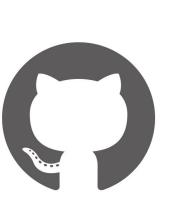


Generating a machine-readable representation of the most prominent information in a text document as a set of **salient open facts** 



Unsupervised and knowledge agnostic, based on Open Information Extraction, PageRank and Clustering

https://github.com/mponza/SalIE



## Experiments

- Sal E outperforms baselines and text summarizers
- Open facts are an effective way to compress information

## SallE - Salient Open Information Extraction

("Abrams", "was 56-years-old native of", "Pittsburgh area")

("Abrams", "had been stabbed to death in", "apartment")

("Apartment", "tending wounds at time of", "murder")

("Cousin of husband", "had gone into", "business")

("Remains", "were discovered beside warehouse at edge of", "cinder-topped soccer field on outskirts of Panama City")

("Abrams", "got more involved in", "real estate")

Open Facts

## Open Information Extraction

(i.e., MinIE (Gashteovski, EMNLP 2017))



("Remains", "were discovered beside warehouse at edge of", "cinder-topped soccer field on outskirts of Panama City")

("Apartment", "tending wounds at time of", "murder")

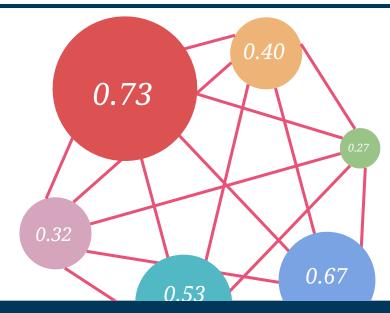
("Cousin of husband", "had gone into", "business")



iversificatior

#### Facts as Nodes

Open facts are mapped into nodes of a complete graph.



#### **Relevance Computation**

PageRank's stationary distribution eventually captures the final **relevance** score of each open fact.

#### Clustering

Open facts are clustered together with respect to the head of their subject.

#### Diversification

The resulting facts are chosen by selecting the most relevant fact for each cluster.

### **Edge Weighting**

Edges' weights are computed with the cosine similarity between centroids of word embeddings in the facts.

GloVe embeddings learned over more than 100M of open facts extracted from Wikipedia (everything is **publicly available**!)

#### Relevance Prior

Author commonly express relevant information at the beginning: PageRank's **teleport vector** is instantiated with respect to facts' positional information.

0.73 - ("Abrams", "had been stabbed to death in", "apartment")

0.32 - ("Abrams", "got more involved in", "real estate")

0.27 - ("Abrams", "was 56-years-old native of", "Pittsburgh area")

0.67 - ("Remains", "were discovered beside warehouse at edge of", "cinder-topped soccer field on outskirts of Panama City")

0.53 - ("Apartment", "tending wounds at time of", "murder")

0.40 - ("Cousin of husband", "had gone into", "business")

#### ("Abrams", "had been stabbed to death in", "apartment")

Salient Open Facts

## **Experiments**

#### **ROUGE-1 ROUGE-L** Method 20.4 24.8 27.8 29.7 21.8 24.4 26.0 13.9 18.1 Position 19.8 21.7 15.3 24.5 13.8 17.2 15.2 19.8 26.1 TextRank 19.2 11.3 22.2 23.3 16.7 18.2 20.3 10.1 14.3 29.7 **30.4** 22.5 **26.**7 8.5 34.1 8.00 18.0 Berkeley 10.0 10.6 21.2 25.4 14.8 19.0 22.5 9.90 30.0 26.8 **17.1** 30.9 24.3 26.0 24.2 **15.3** 25.9 19.1 11.6 **17.9** 21.6 **24.2 15.9**

Experimental results on a real-world dataset (i.e., New York Times) where we evaluate the top-k extracted facts, with  $k \in [1; 5]$ . For each system we report the performance with MinIE's safe (top score of each row) and aggressive (bottom score of each row) modes. More experiments in the paper!

## References

Michele Banko, Michael J Cafarella, Stephen Soderland, Matthew Broadhead and Oren Etzioni. **Open Information Extraction from the Web**. In *Proceedings of the 2017 International Joint Conference* on Artificial Intelligence, IJCAI 2007, pages 2670-2676, 2007.

Luciano Del Corro and Rainer Gemulla. ClausIE: Clause-Based Open Information Extraction. In Proceedings of the 2013 International World Wide Web Conference, WWW 2013, pages 355-366, 2013.

Gabriel Stanovsky, Mausam, and Ido Dagan. Open IE as an Intermediate Structure for Semantic **Tasks**. In *Proceedings of the 2015 Annual Meeting of the Association for Computational Linguistics*, ACL 2015, pages 303-308, 2015.

Kiril Gashteovski, Rainer Gemulla and Luciano Del Corro. MinIE: Minimizing Facts in Open Information Extraction. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, pages 2630-2640, 2017.



