Semi-Open Relation Extraction from Scientific Texts

Ruben Kruiper†, Julian F.V. Vincent, Jessica Chen-Burger,
Marc P.Y. Desmulliez, Ioannis Konstas

Heriot-Watt University
Riccarton Campus, EH14 4AS
Edinburgh, United Kingdom
†corresponding author: rk22@hw.ac.uk

Information Extraction (IE) can provide a summary view of a scientific text, which can ease manual analysis and is relevant for downstream tasks [1–6]. However, the focus of narrow IE systems and datasets is too narrow, i.e., they extract a handful of semantic relations, such as 'PART-OF' and 'COMPARISON' [4, 7–10]. And the alternative Open Information Extraction (OIE) paradigm [11, 12] is too inclusive, i.e., these systems extract many uninformative, incoherent and redundant relations [13, 14]. To make things worse, scientific IE is a much harder task than general domain IE. First, sentences in scientific texts are longer and more complex, which (1) leads to reduced performance of systems developed for general domain texts [15, 16], and (2) can affect the quality of pre-computed features [17, 18]. Second, scientific texts contain many unique relations, and creating a specific classifier for each possible relation type is impractical [15]. Third, arguments of relations in scientific text are often keyphrase that contain Multi-Word Expressions (MWE); these MWEs are harder to identify that Named Entities, may refer to a single entity in various surface forms, and are often rare [4, 7, 19]. Ensuring that these rare terms are not treated as Out-Of-Vocabulary (OOV) is important, as they can provide important cues [20].

In this work we combine the output of narrow IE and OIE systems to achieve Semi-Open Relation Extraction[1], a new task that we explore in the Biology domain [22]. We train a strong scientific IE system [10] on the Focused Open Biological Information Extraction (FOBIE) dataset [21] to accurately extract negative correlations between keyphrases. These correlations capture the central information in a subset of biology research articles [23–26]. The trained narrow IE system and a state-of-the-art OIE system [27–30] are both run on a corpus of 10K open-access biological texts. We use the central keyphrases identified through narrow IE to filter the OIE extractions, and automatically discard a significant amount (65%) of erroneous and uninformative OIE extractions. A qualitative comparison of the filtered and unfiltered extractions indicates that Semi-Open Relation Extraction improves the overall informativeness of OIE extractions for a human reader.

Our exploratory results indicate that Semi-Open Relation Extraction provides a balance between the accuracy of narrow IE and the flexibility of OIE. Our simple approach can be improved, and may already be used to aid the collection of a larger and more comprehensive dataset to push the boundaries of scientific IE.

[1] The Focused Open Biological Information Extraction (FOBIE) dataset [21], as well as code to explore the Semi-Open Relation Extraction (SORE) task [22] can be found on https://github.com/rubenkruiper/FOBIE

References

[1] Sonal Gupta and Christopher D Manning. Analyzing the Dynamics of Research by Extracting Key Aspects of Scientific Papers. In *Proceedings of 5th international joint conference on natural language processing*, pages 1–9, 2011.

[2] Chen-Tse Tsai, Gourab Kundu, and Dan Roth. Concept-based analysis of scientific literature. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management - CIKM '13*, pages 1733–1738, New York, New York, USA, 2013. ACM Press.

[3] Mausam. Open Information Extraction Systems and Downstream Applications. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence*, pages 4074–4077, 2016.

[4] Isabelle Augenstein, Mrinal Das, Sebastian Riedel, Lakshmi Vikraman, and Andrew McCallum. SemEval 2017 Task 10: ScienceIE - Extracting Keyphrases and Relations from Scientific Publications. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 546–555, Stroudsburg, PA, USA, 2017. Association for Computational Linguistics.

[5] Chris Quirk and Hoifung Poon. Distant Supervision for Relation Extraction beyond the Sentence Boundary. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, volume 1, pages 1171–1182, Stroudsburg, PA, USA, 2017. Association for Computational Linguistics.

[6] Nanyun Peng, Hoifung Poon, Chris Quirk, Kristina Toutanova, and Wen-tau Yih. Cross-Sentence N-ary Relation Extraction with Graph LSTMs. *Transactions of the Association for Computational Linguistics*, 5:101–115, 12 2017.

[7] Louise Delėger, Robert Bossy, Estelle Chaix, Mouhamadou Ba, Arnaud Ferrė, Philippe Bessières, and Claire Nėdellec. Overview of the Bacteria Biotope Task at BioNLP Shared Task 2016. In *Proceedings of the 4th BioNLP Shared Task Workshop*, pages 12–22, Stroudsburg, PA, USA, 2016. Association for Computational Linguistics.

[8] Estelle Chaix, Bertrand Dubreucq, Abdelhak Fatihi, Dialekti Valsamou, Robert Bossy, Mouhamadou Ba, Louise Delėger, Pierre Zweigenbaum, Philippe Bessières, Loïc Lepiniec, and Claire Nėdellec. Overview of the Regulatory Network of Plant Seed Development (SeeDev) Task at the BioNLP Shared Task 2016. In *Proceedings of the 4th BioNLP Shared Task Workshop*, pages 1–11, 2016.

[9] Kata Gábor, Davide Buscaldi, Anne-Kathrin Schumann, Behrang Qasemizadeh, Hafa Zargayouna, and Thierry Charnois. SemEval-2018 Task 7: Semantic Relation Extraction and Classification in Scientific Papers. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 679–688, 2018.

[10] Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. Multi-Task Identification of Entities, Relations, and Coreference for Scientific Knowledge Graph Construction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3219–3232, Stroudsburg, PA, USA, 8 2018. Association for Computational Linguistics.

[11] Michele Banko, M Cafarella, Stephen Soderland, M J Broadhead, and Oren Etzioni. Open information extraction from the web. *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, pages 2670–2676, 2007.

[12] Oren Etzioni, Anthony Fader, Janara Christensen, Stephen Soderland, and Mausam. Open information extraction: The second generation. In *IJCAI International Joint Conference on Artificial Intelligence*, volume 11, pages 3–10, 2011.

[13] Rudolf Schneider, Tom Oberhauser, Tobias Klatt, Felix A. Gers, and Alexander Löser. Analysing Errors of Open Information Extraction Systems. In *Proceedings of the First Workshop on Building Linguistically Generalizable NLP Systems*, pages 11–18, 7 2017.

[14] Christina Niklaus, Matthias Cetto, André Freitas, and Siegfried Handschuh. A Survey on Open Information Extraction. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3866–3878. Association for Computational Linguistics, 2018.

[15] Paul Groth, Michael Lauruhn, Antony Scerri, and Ron Daniel. Open Information Extraction on Scientific Text: An Evaluation. In *Proceedings of the 27th International Conference on Computational Linguistics*, page 3414–3423. Association for Computational Linguistics, 2 2018.

[16] Arzoo Katiyar and Claire Cardie. Going out on a limb: Joint Extraction of Entity Mentions and Relations without Dependency Trees. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, 2017.

[17] Ryan McDonald and Joakim Nivre. Characterizing the errors of data-driven dependency parsing models. In *EMNLP-CoNLL 2007 - Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 122–131. Association for Computational Linguistics, 2007.

[18] Xiaotian Jiang, Quan Wang, Peng Li, and Bin Wang. Relation Extraction with Multi-instance Multi-label Convolutional Neural Networks. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers.*, pages 1471–1480, 2016.

[19] Dzmitry Bahdanau, Tom Bosc, Stanisław Jastrzębski, Edward Grefenstette, Pascal Vincent, and Yoshua Bengio. Learning to Compute Word Embeddings On the Fly. In *arXiv preprint arXiv:1706.00286*, 2017.

[20] Yi Luan, Mari Ostendorf, and Hannaneh Hajishirzi. Scientific Information Extraction with Semi-supervised Neural Tagging. In *Proceedings of the*

2017 Conference on Empirical Methods in Natural Language Processing, page 2641–2651, Copenhagen, Denmark, 2017.

[21] Ruben Kruiper, Julian F V Vincent, Jessica Chen-Burger, Marc P Y Desmulliez, and Ioannis Konstas. A Scientific Information Extraction Dataset for Nature Inspired Engineering. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, page 2078–2085, Marseille, 2020.

[22] Ruben Kruiper, Julian F V Vincent, Jessica Chen-Burger, Marc P Y Desmulliez, and Ioannis Konstas. In Layman's Terms: Semi-Open Relation Extraction from Scientific Texts. In *arXiv preprint arXiv:2005.07751*, 2020.

[23] S C Stearns. Trade-Offs in Life-History Evolution. *Functional Ecology*, 3(3):259–268, 1989.

[24] Theodore Garland. Trade-offs. *Current Biology*, 24(2):R60–R61, 2014.

[25] Thomas Ferenci. Trade-off Mechanisms Shaping the Diversity of Bacteria. *Trends in Microbiology*, 24(3):209–223, 3 2016.

[26] Julian F. V. Vincent. The trade-off: a central concept for biomimetics. *Bioinspired, Biomimetic and Nanobiomaterials*, 6(2):67–76, 6 2016.

[27] Janara Christensen, Mausam, Stephen Soderland, and Oren Etzioni. An analysis of open information extraction based on semantic role labeling. In *Proceedings of the sixth international conference on Knowledge capture - K-CAP '11*, page 113, 2011.

[28] Harinder Pal and Mausam. Demonyms and Compound Relational Nouns in Nominal Open IE. In *Proceedings of the 5th Workshop on Automated Knowledge Base Construction*, page 35–39, 2016.

[29] Swarnadeep Saha, Harinder Pal, and Mausam. Bootstrapping for Numerical Open IE. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 317–323, 2017.

[30] Swarnadeep Saha and Mausam. Open Information Extraction from Conjunctive Sentences. In *Proceedings ofthe 27th International Conference on Computational Linguistics*, volume 1, pages 2288–2299, Santa Fe, New Mexico, 2018.