Final Report:
Information Consolidation: A New Paradigm in Knowledge Search

December 14, 2021

1 General Information

DFG reference number

Grant **GU 798/17-1** and **DIP grant DA 1600/1-1**

Applicants

Iryna Gurevych, Prof. Dr
Chair: Ubiquitous Knowledge Processing (UKP)
Technische Universität Darmstadt, Computer Science Department
Hochschulstraße 10, 64289 Darmstadt
Web page: `https://www.ukp.tu-darmstadt.de`
Ido Dagan, Prof. Dr.
Chair: Natural Language Processing Lab
Bar Ilan University
Ramat Gan, 52900, Israel
Web page: `http://u.cs.biu.ac.il/~dagan/`

Topic/title of the project

English

Title: Information Consolidation: A New Paradigm in Knowledge Search

Topic: Combination of methods from information retrieval, natural language processing and information science with the aim of going a step further in the field of information access by means of the development of an automated information consolidation approach, in order to substantially automate the process of knowledge search.

Deutsch

Titel: Informationskonsolidierung als neues Paradigma der Wissenssuche

Thema: Die Kombination von Methoden aus Informationsgewinnung, maschineller Sprachverarbeitung und der Informatik, um einen Ansatz zur Informationskonsolidierung zu entwickeln, der den Prozess der Wissenssuche signifikant automatisiert.

Period covered by the report, overall funding period

Covered by the report

1.12.2014 - 01.05.2021

Overall funding period

1/12/2014 - 30/11/2017 (initial approved period)
1/12/2017 - 01/05/2021 (second funding period)

1.1 Selected Publications in Acknowledgement to DIP

Overall, the project members have produced 66 publications in acknowledgement to DIP. All of our publications were either published at or accepted to venues with scientific quality assurance, in standard formats. Throughout the report, citations to project publications are marked in bold.

Joint DIP publications

- Tobias Falke, Gabriel Stanovsky, Iryna Gurevych, Ido Dagan. 2016. **Porting an Open Information Extraction System from English to German**. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP '16), pp:892-898. Austin, Texas, U.S.A.
- Ivan Habernal, Maria Sukhareva, Fiana Raiber, Ann Shtok, Oren Kurland, Hadar Ronen, Judit Bar-Ilan and Iryna Gurevych. 2016. **New Collection Announcement: Focused Retrieval Over the Web**. In: Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '16), pp:701-704. Pisa, Italy.
- Omer Levy, Ido Dagan, Gabriel Stanovsky, Judith Eckle-Kohler, Iryna Gurevych. 2016. **Modeling Extractive Sentence Intersection via Subtree Entailment**. In: Proceedings of the 26th International Conference on Computational Linguistics (COLING '16), pp: 2891-2901. Osaka, Japan.
- Iryna Gurevych, Judith Eckle-Kohler, Michael Matuschek. 2016. **Linked Lexical Knowledge Bases: Foundations and Applications**. Synthesis Lectures on Human Language Technologies. Editor: Graeme Hirst. Morgan & Claypool Publishers. ISBN: 9781627059749.
- Gabriel Stanovsky, Judith Eckle-Kohler, Yevgeniy Puzikov, Ido Dagan, Iryna Gurevych. 2017. **Integrating Deep Linguistic Features in Factuality Prediction over Unified Datasets**. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL '17), pp:352-357. Vancouver, Canada.
- Rachel Wities, Vered Shwartz, Gabriel Stanovsky, Meni Adler, Ori Shapira, Shyam Upadhyay, Dan Roth, Eugenio Martínez-Cámara, Iryna Gurevych and Ido Dagan. 2017. **A Consolidated Open Knowledge Representation for Multiple Texts**. In: Proceedings of the Workshop Linking Models of Lexical, Sentential and Discourse-level Semantics (LSDSem 2017) in (EACL '17), pp:12-24. Valencia, Spain.
- Eugenio Martínez Cámara, Vered Shwartz, Iryna Gurevych, Ido Dagan. 2017. **Neural Disambiguation of Causal Lexical Markers Based on Context**. In: Proceedings of the 12th International Conference on Computational Semantics (IWCS 2017), Volume 2: Short papers, Montpellier, France,
- Gabriel Stanovsky, Judith Eckle-Kohler, Yevgeniy Puzikov, Ido Dagan, Iryna Gurevych. 2017. **Integrating Deep Linguistic Features in Factuality Prediction over Unified Datasets**. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL 2017), pp. 352-357, Vancouver, Canada.
- Shany Barhom, Vered Shwartz, Alon Eirew, Michael Bugert, Nils Reimers and Ido Dagan. 2019. **Revisiting Joint Modeling of Cross-document Entity and Event Coreference Resolution**. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL '19), pp: 4179-4189. Florence, Italy.
- Yevgeniy Puzikov, Claire Gardent, Ido Dagan, and Iryna Gurevych. 2019. **Revisiting the Binary Linearization Technique for Surface Realization**. In: The 12th International Conference on Natural Language Generation (INLG 2019), pp. 268-278, Tokyo, Japan,
- Florian Böhm, Yang Gao, Christian Meyer, Ori Shapira, Ido Dagan, Iryna Gurevych. 2019. **Better Rewards Yield Better Summaries: Learning to Summarise Without References**. In: The 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP 2019), pp. 3101-3111, Hong Kong, China.
- Michael Bugert, Nils Reimers, Shany Barhom, Ido Dagan and Iryna Gurevych. 2020. **Breaking the Subtopic Barrier in Cross-Document Event Coreference Resolution**. In: Proceedings of Text2Story — Third Workshop on Narrative Extraction From Texts co-located with 42nd European Conference on Information Retrieval (ECIR '20). Lisbon, Portugal.

Publications in acknowledgement to DIP

- Oren Melamud, Ido Dagan and Jacob Goldberger. 2015. **Modeling Word Meaning in Context with Substitute Vectors**. In: Proceedings of the 2015 Conference of the North American Chapter of the

Association for Computational Linguistics: Human Language Technologies (NAACL '15), pp:472-482. Denver, Colorado.

- Oren Melamud, Omer Levy and Ido Dagan. 2015. **A Simple Word Embedding Model for Lexical Substitution**. In: Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing, pp:1-7. Denver, Colorado.
- Vered Shwartz, Omer Levy, Ido Dagan and Jacob Goldberger. 2015. **Learning to Exploit Structured Resources for Lexical Inference**. In: Proceedings of the Nineteenth Conference on Computational Natural Language Learning (CoNLL '15), pp:175-184. Beijing, China.
- Shwartz, Vered, Yoav Goldberg, and Ido Dagan. **Improving Hypernymy Detection with an Integrated Path-based and Distributional Method.** In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 2389-2398. 2016. Harvard. **(won the Outstanding Paper award)**
- Elinor Brondwine, Anna Shtok and Oren Kurland. 2016. **Utilizing Focused Relevance Feedback**. In: Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '16), pp:1061-1064. Pisa, Italy.
- Sukhareva, Maria ; Eckle-Kohler, Judith ; Habernal, Ivan ; Gurevych, Iryna. 2016. **Crowdsourcing a Large Dataset of Domain-Specific Context-Sensitive Semantic Verb Relations.** In: Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016), pp. 2131-2137, European Language Resources Association (ELRA), Portoroz, Slovenia.
- Eckle-Kohler, Judith. 2016 **Verbs Taking Clausal and Non-Finite Arguments as Signals of Modality – Revisiting the Issue of Meaning Grounded in Syntax.** In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016), pp. 811-822, Association for Computational Linguistics, Berlin, Germany.
- Ficler, Jessica, and Yoav Goldberg. **Improved Parsing for Argument-Clusters Coordination.** In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pp. 72-76. 2016.
- Ficler, Jessica, and Yoav Goldberg. **Coordination Annotation Extension in the Penn Tree Bank.** In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 834-842. 2016.
- Ficler, Jessica, and Yoav Goldberg. **A Neural Network for Coordination Boundary Prediction.** In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pp. 23-32. 2016.
- Shwartz, Vered, Gabriel Stanovsky, and Ido Dagan. **Acquiring predicate paraphrases from news tweets.** Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (* SEM 2017). 2017.
- Shapira, Ori, Hadar Ronen, Meni Adler, Yael Amsterdamer, Judit Bar-Ilan, and Ido Dagan. **Interactive abstractive summarization for event news tweets.** In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pp. 109-114. 2017.
- Bugert, Michael ; Puzikov, Yevgeniy ; Rücklé, Andreas ; Eckle-Kohler, Judith ; Martin, Teresa ; Martínez Cámara, Eugenio ; Sorokin, Daniil ; Peyrard, Maxime ; Gurevych, Iryna. 2017. **LSDSem 2017: Exploring Data Generation Methods for the Story Cloze Test.** In: Proceedings of the 2nd Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics, pp. 56-61, Association for Computational Linguistics, The 2nd Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics, Valencia, Spain.
- Ficler, Jessica, and Yoav Goldberg. **Improving a Strong Neural Parser with Conjunction-Specific Features.** In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, pp. 343-348. 2017. Harvard
- Ficler, Jessica, and Yoav Goldberg. **Controlling Linguistic Style Aspects in Neural Language Generation.** In Proceedings of the Workshop on Stylistic Variation, Association for Computational Linguistics, pp. 94-104. 2017.
- Puzikov, Yevgeniy ; Gurevych, Iryna. 2018. **E2E NLG Challenge: Neural Models vs. Templates.** In: Proceedings of the 11th International Conference on Natural Language Generation (INLG 2018), pp. 463-471.

- Puzikov, Yevgeniy ; Gurevych, Iryna. 2018. **BinLin: A Simple Method of Dependency Tree Linearization.** In: Proceedings of the Multilingual Surface Realization Workshop 2018 (ACL 2018), pp. 13-28, Melbourne, Australia, Surface Realization Shared Task 2018, Melbourne, Australia.
- Michael, Julian, Gabriel Stanovsky, Luheng He, Ido Dagan, and Luke Zettlemoyer. **Crowdsourcing Question-Answer Meaning Representations.** In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), pp. 560-568. 2018.
- Stanovsky, Gabriel, and Ido Dagan. **Semantics as a foreign language.** In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 2412-2421. 2018.
- Aharoni, Roee, and Yoav Goldberg. **Split and Rephrase: Better Evaluation and Stronger Baselines.** In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pp. 719-724. 2018. Harvard
- Glockner, Max, Vered Shwartz, and Yoav Goldberg. **Breaking NLI Systems with Sentences that Require Simple Lexical Inferences.** In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pp. 650-655. 2018. Harvard
- Shapira, Ori, David Gabay, Hadar Ronen, Judit Bar-Ilan, Yael Amsterdamer, Ani Nenkova, and Ido Dagan. **Evaluating multiple system summary lengths: A case study.** In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 774-778. 2018.
- Elazar, Yanai, and Yoav Goldberg. **Adversarial Removal of Demographic Attributes from Text Data.** In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 11-21. 2018.
- Amrami, Asaf, and Yoav Goldberg. **Word Sense Induction with Neural biLM and Symmetric Patterns.** In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 4860-4867. 2018.
- Shwartz, Vered, and Ido Dagan. **Paraphrase to Explicate: Revealing Implicit Noun-Compound Relations.** In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 1200-1211. 2018.
- Stanovsky, Gabriel, Julian Michael, Luke Zettlemoyer, and Ido Dagan. **Supervised open information extraction.** In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pp. 885-895. 2018.
- Sheetrit, Eilon. **Utilizing Inter-Passage Similarities for Focused Retrieval.** In The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, pp. 1453-1453. 2018.
- Sheetrit, Eilon, Anna Shtok, Oren Kurland, and Igal Shprincis. **Testing the cluster hypothesis with focused and graded relevance judgments.** In The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, pp. 1173-1176. 2018.
- Shapira, Ori, David Gabay, Yang Gao, Hadar Ronen, Ramakanth Pasunuru, Mohit Bansal, Yael Amsterdamer, and Ido Dagan. **Crowdsourcing Lightweight Pyramids for Manual Summary Evaluation.** In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp. 682-687. 2019.
- Shwartz, Vered, and Ido Dagan. **Still a Pain in the Neck: Evaluating Text Representations on Lexical Composition.** Transactions of the Association for Computational Linguistics 7 (2019): 403-419.
- Moryossef, Amit, Yoav Goldberg, and Ido Dagan. **Improving Quality and Efficiency in Plan-based Neural Data-to-text Generation.** In Proceedings of the 12th International Conference on Natural Language Generation, pp. 377-382. 2019.
- Moryossef, Amit, Yoav Goldberg, and Ido Dagan. **Step-by-Step: Separating Planning from Realization in Neural Data-to-Text Generation.** In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp. 2267-2277. 2019.
- Rozen, Ohad, Vered Shwartz, Roee Aharoni, and Ido Dagan. **Diversify Your Datasets: Analyzing Generalization via Controlled Variance in Adversarial Datasets.** In Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL), pp. 196-205. 2019.

- Sheetrit, Eilon, and Oren Kurland. **Cluster-based focused retrieval.** In Proceedings of the 28th ACM International Conference on Information and Knowledge Management, pp. 2305-2308. 2019.
- Elazar, Yanai, and Yoav Goldberg. **Where's My Head? Definition, Data Set, and Models for Numeric Fused-Head Identification and Resolution.** Transactions of the Association for Computational Linguistics 7 (2019): 519-535.
- Reimers, Nils ; Gurevych, Iryna. 2019. **Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks.** In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP 2019), Hong Kong, China.
- Roit, Paul, Ayal Klein, Daniela Stepanov, Jonathan Mamou, Julian Michael, Gabriel Stanovsky, Luke Zettlemoyer, and Ido Dagan. **Controlled Crowdsourcing for High-Quality QA-SRL Annotation.** In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 7008-7013. 2020.
- Meged, Yehudit, Avi Caciularu, Vered Shwartz, and Ido Dagan. **Paraphrasing vs Coreferring: Two Sides of the Same Coin.** In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings, pp. 4897-4907. 2020.
- Bornstein, Ari, Arie Cattan, and Ido Dagan. "CoRefi: A Crowd Sourcing Suite for Coreference Annotation." In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pp. 205-215. 2020.
- Pyatkin, Valentina, Ayal Klein, Reut Tsarfaty, and Ido Dagan. **QADiscourse-Discourse Relations as QA Pairs: Representation, Crowdsourcing and Baselines.** In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 2804-2819. 2020.
- Barkan, Oren, Avi Caciularu, and Ido Dagan. **Within-Between Lexical Relation Classification Using Path-based and Distributional Data.** In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 3521-3527. 2020.
- Reimers, Nils ; Gurevych, Iryna. 2020. **Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation.** pp. 4512-4525, Association for Computational Linguistics, The 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020).
- Mesgar, Mohsen ; Bücker, Sebastian ; Gurevych, Iryna. 2020. **Dialogue Coherence Assessment Without Explicit Dialogue Act Labels.** pp. 1439-1450, The 58th annual meeting of the Association for Computational Linguistics (ACL 2020).
- Şahin, Gözde Gül ; Kementchedjhieva, Yova ; Rust, Phillip ; Gurevych, Iryna. 2020. **PuzzLing Machines: A Challenge on Learning From Small Data.** pp. 1241-1254, The 58th annual meeting of the Association for Computational Linguistics (ACL 2020).
- Simpson, Edwin ; Gurevych, Iryna. 2020. **Scalable Bayesian Preference Learning for Crowds.** In: Machine Learning, 109, pp. 689-718. Springer.
- Klein, Ayal, Jonathan Mamou, Valentina Pyatkin, Daniela Stepanov, Hangfeng He, Dan Roth, Luke Zettlemoyer, and Ido Dagan. **QANom: Question-Answer driven SRL for Nominalizations.** In Proceedings of the 28th International Conference on Computational Linguistics, pp. 3069-3083. 2020.
- Sheetrit, Eilon, Anna Shtok, and Oren Kurland. **A passage-based approach to learning to rank documents.** Information Retrieval Journal 23, no. 2 (2020): 159-186.
- Eirew, Alon, Arie Cattan, and Ido Dagan. **WEC: Deriving a Large-scale Cross-document Event Coreference dataset from Wikipedia.** In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 2498-2510. 2021.
- Shapira, Ori, Ramakanth Pasunuru, Hadar Ronen, Mohit Bansal, Yael Amsterdamer, and Ido Dagan. **Extending Multi-Document Summarization Evaluation to the Interactive Setting.** In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 657-677. 2021.
- Bugert, Michael ; Gurevych, Iryna. 2021. **Event Coreference Data (Almost) for Free: Mining Hyperlinks from Online News.** In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Association for Computational Linguistics, The 2021 Conference on Empirical Methods in Natural Language Processing, virtual Conference and Punta Cana, Dominican Republic.

- Bugert, Michael ; Reimers, Nils ; Gurevych, Iryna. 2021. **Generalizing Cross-Document Event Coreference Resolution Across Multiple Corpora.** In: Computational Linguistics, MIT Press, ISSN 0891-2017, DOI: 10.1162/coli_a_00407.
- Reimers, Nils ; Gurevych, Iryna. 2021 **The Curse of Dense Low-Dimensional Information Retrieval for Large Index Sizes.** In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pp. 605-611.
- Thakur, Nandan ; Reimers, Nils ; Daxenberger, Johannes ; Gurevych, Iryna. 2021. **Augmented SBERT: Data Augmentation Method for Improving Bi-Encoders for Pairwise Sentence Scoring Tasks.** In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 296-310, ACL, 2021 Annual Conference of the North American Chapter of the Association for Computational Linguistics.
- Mesgar, Mohsen ; Simpson, Edwin ; Gurevych, Iryna. 2021. **Improving Factual Consistency Between a Response and Persona Facts**. In: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, pp. 549-562, 16th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2021).
- Pyatkin, Valentina and Roit, Paul and Michael, Julian and Goldberg, Yoav and Tsarfaty, Reut and Dagan, Ido. 2021. **Asking It All: Generating Contextualized Questions for any Semantic Role.**. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP).
- Brook Weiss, Daniela, Paul Roit, Ayal Klein, Ori Ernst, and Ido Dagan. 2021. **QA-Align: Representing Cross-Text Content Overlap by Aligning Question-Answer Propositions.**. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP).

2 Final Progress Report

This section reports the progress of the project: the initial question and objectives of the project (Section 2.1), the performed research, by work packages (Section 2.2), follow-up research (Section 2.3) and economic value (Section 2.4).

2.1 Project's initial questions and objectives

Our proposal targeted what might be one of the next big steps in information access technology: supporting users in identifying and quickly assimilating the factual content of multiple relevant texts, by exploring the set of concrete *statements* found within them. The goal was approached via an interdisciplinary approach, combining methods from information retrieval, natural language processing and information science. Concretely, we proposed developing an automated information consolidation approach that consists of three major processes: (a) automatically **extracting relevant "atomic" statements** from multiple texts; (b) **consolidating the information** in these statements by constructing a statement graph, which specifies the inference relations between all extracted statements; and (c) enabling users to **explore the consolidated information** in the graph via a suitable interactive user interface. The concrete research sub-goals of the project were to develop the required methods for the above processes, as well as their integration. The primary application domain for demonstrating the novel consolidation and exploration scheme has been texts in the educational domain.

2.2 Project Developments

A series of work packages (WPs) were proposed to achieve the above goals. We first outline briefly the WP structure and then describe the actual research progress, achieved under each WP.

WP1 and **WP2** concern with **preparatory** steps: corpus creation and annotation, and setting up an NLP pipeline for various project processes.

WP3 and **WP4** are concerned with the **statement extraction** phase. WP3 deals with the Information Retrieval aspect of extracting relevant passages (sentences) from retrieved documents while WP4 deals with the NLP aspect of extracting atomic statements from given sentences.

WP5 and **WP6** deal with the **statement consolidation** phase. WP5 aims to define a suitable graph-based structure that represents consolidated information from multiple statements. WP6 develops lexical inference methods, which are needed for the consolidation process.

In **WP7** we develop methods for **interactive exploration** of the consolidated information.

WP8 and **WP9** address the **supporting aspects of evaluation and system integration**, respectively.

In the first half of the project we achieved research results for all project components and obtained a first integrated system, while working on texts from the educational and news domains.

These results are described below for each work package. References to the corresponding publications in acknowledgement to DIP appear in bold within the text, where the corresponding authors indicate the project members that were inherently involved in each particular research component.

Work Package 1: Corpus creation and annotation

In this work package we created corpora with two different types of annotations, which correspond to different tasks: (1) annotating passage relevance judgements, corresponding to WP3, and (2) annotating (jointly) statement extraction and consolidation, corresponding to WP4, WP5 and WP6.

The main outcomes for the first task are a set of user needs-motivated queries, a specification of the annotation process, and a publicly available corpus annotated with information about sentences relevant to the queries. Specific details can be found in **(Habernal et al., 2016)**.

The source document corpus is category A of the English ClueWeb12 collection which contains about 733 million documents. Forty-nine short keyword queries, accompanied by descriptions of the information need, were created.

The queries are from the educational domain and are of topical nature. They represent various information needs of parents, chosen as our *target group*. To ensure high variability of queries, as well as their pertinence to the target group (parents), we combined two approaches for query compilation. The first approach relied on exploiting existing Question-Answering (QA) Web portals and the second approach utilized a user questionnaire.

Sentences in documents highly ranked in response to the queries by a state-of-the-art learning-to rank method were annotated for relevance, as well as their ambient documents. For each annotated document, we computed observed annotation agreement using the DKPro Agreement package (Meyer et al., 2014). The minimal units for agreement computation were sentences, each with two categories per annotator (relevant or non-relevant). Average agreement over all documents judged is 0.725 and standard deviation is 0.175 (where agreement ranges between 0 and 1).

Overall, the documents retrieved for 49 queries were annotated. Per query, about 98 documents on average were annotated on a sentence level. On average, about 87 documents and about 4,618 sentences per query were judged relevant. Overall, about 89% of the annotated documents are relevant and about 36% of the sentences are relevant.

The aim for the second task has been the creation of annotation guidelines, and an annotated corpus, for the annotation of statement extraction from input sentences and their consolidation. The results of this effort are described in detail in **(Wities et al., 2017a)**. The annotation guidelines were designed to implement the Open Knowledge Representation (OKR) scheme, which was developed as part of WP5 (see further below), aiming to capture the consolidated information expressed jointly in a set of texts. Following the OKR scheme **(Wities et al., 2017a)**, we annotated a corpus with 1257 sentences.

For this first corpus, we chose to annotate clusters of news-related tweets taken from the Twitter Event Dataset (McMinn et al., 2013). These texts were chosen because they present high redundancy, which our representation aims to address, making manual annotation easier. In later efforts, we annotated additional texts from the education domain, extracted from the English ClueWeb12 corpus, assessing the generality of the OKR scheme and its applicability for the education domain.

Our dataset is available at `http://u.cs.biu.ac.il/~nlp/resources/downloads/twitter-events`. Detailed annotation guidelines, the annotation tool and the baseline implementations are available at `https://github.com/vered1986/OKR`.

In another line of work, we adopted and extended the Question-Answer driven Semantic Role Labeling paradigm (QASRL) (Fitzgerald et al., 2018), for attaining more fine grained propositions in the level of predicate-argument relations.

Importantly, QASRL reliance on naturally sounding QAs enables it to be easily annotated by laymen, e.g. using crowdsourcing (Fitzgerald et al., 2018). To alleviate a substantial coverage issue in the prior constructed dataset, and to facilitate strict parser evaluation and comparison for future advancements, we crowdsourced a new evaluation set for QASRL **(Roit et al., 2020)**. Our 'controlled crowdsourcing' methodology, which we developed here and re-use in all forthcoming resource construction works, consists of pre-selecting and training crowd-workers for the task, and assuring data quality through an additional annotation consolidation step, undertaken by the best performing annotators. The evaluation dataset and scripts are available at (`https://github.com/plroit/qasrl-gs`).

The original QASRL works tackle only verbal predicates. In practice, however, many events in text are

manifested through nominalizations. We addressed this gap by three-step process, where we first (1) filter candidate common nouns that have a derivationally related verb, using lexical resources and heuristics; (2) classified, using crowd-workers, whether each candidate noun occurrence has a verbal meaning in context; and (3) if so, collected QASRL-like question-answer pairs for that deverbal noun, capturing its predicate-argument structure in a uniform label space with verbal QASRL. The resulting dataset, dubbed QANom **(Klein et al., 2020)**, consists of over 10K sentences and 26K QA pairs (`https://github.com/kleinay/QANom`).

In another work, we have devised a QA-based representation for propositional discourse relations, dubbed QADiscourse **(Pyatkin et al., 2020)**. In this paper, we target extracting relations which hold between two propositions, expressed through QA-pairs. We created question templates suitable to ask about discourse relation senses, similar to the relation senses expressed in the Penn Discourse Tree Bank (PDTB). For example, when asking about a concession-relation, one would start the question with "Despite what..?", while when asking about a contrast-relation, one would start with "What is contrasted with. . . ?". Our collected dataset contains more than 16K QA pairs, expressing 16 different discourse relation senses, and it can be found with the following link `https://github.com/ValentinaPy/QADiscourse`.

Complementary to the efforts described above, we are working on soliciting essential information conveyed by adjectives (work in progress).

Literature review revealed that existing corpora do not accurately reflect the interconnectedness of events found in news, since their annotated event clusters are highly self-contained in small sets of documents. Furthermore, their annotation was laborious and costly, requiring several months of commitment from trained expert annotators.

Improving model prediction quality requires larger, more diverse CDCR corpora, as well as novel annotation and curation techniques to produce such corpora at scale. We developed several such approaches:

To overcome the issue of self-contained clusters and annotation time, we developed a first crowdsourcing approach to CDCR in which crowd annotators mark sentences as mentioning a subset of predefined events. Here, restricting the set of events and granularity of mention spans allows annotating a large number of documents without having to cluster the documents a priori, leading to annotation of coreference links even between dissimilar documents. Using this technique, we produced the Football Coreference Corpus (FCC) consisting of 500 documents from the sports news domain annotated with football matches and events (Bugert et al., 2020a). Its creation took crowd annotators only three weeks, while reaching similar annotation quality to NLP experts (0.68 Krippendorff's alpha between crowd and experts). We later extended this corpus with additional mentions and token-level annotations using expert annotation (see (Bugert et al., 2020b)).

WEC: Wikipedia Event Coreference (Eirew et al., 2021), is an efficient methodology for gathering a large scale cross-document event coreference (CDEC) dataset from Wikipedia. We applied this methodology to the English Wikipedia, extracting WEC-Eng. Our method is generic and can be applied with rather few adjustments to other Wikipedia languages. We made both code and WEC-Eng dataset publicly available.

Work Package 2: Linguistic Annotation

The objective of WP2 has been technical – to provide a pipeline of linguistic annotation tools for other WPs. This WP ended up being rather simple, since we managed to base our definitions of standard representations and standard tools on UIMA-based annotations and the corresponding DKPro Core framework – a collection of software components for natural language processing which was developed in recent years at UKP (the German partner of the project).[1]

Work Package 3: Segment Retrieval

The goal of this work package was developing improved segment (passage) retrieval methods, addressing the corresponding project sub-goal. A fundamental, somewhat unexplored, question in the realm of passage retrieval is the difference between relevant and non-relevant passages in documents that are considered *relevant* to a query (that is, contain some relevant parts).

We addressed several challenges in using passage (segment)-based information, either for passage retrieval or for document retrieval. First, we developed novel ad hoc document retrieval methods that utilize focused relevance feedback — i.e., feedback with regard to which passages of documents are relevant. We showed that using non-relevant information from documents deemed relevant can actually be of substantial merit **(Brondwine et al., 2016)**. We also developed and executed cluster hypothesis tests for focused retrieval **(Sheetrit et al., 2018)**; that is, we tested the association between segments based on the percentage

[1] `https://dkpro.github.io/`

of relevant information they include. In **(Sheetrit et al., 2018)** we presented the first (to the best of our knowledge) study of cluster-based passage retrieval approaches. (**(Sheetrit, 2018)** discusses, in general, the use of inter-passage similarities for passage retrieval.) In **(Sheetrit et al., 2020)** we presented document retrieval methods, based on learning-to-rank, which utilize highly effective ranking (produced by learning-to-rank) of document passages. Another contribution of our work is a newly created dataset for passage retrieval reported in **(Habernal et al., 2016)**.

In **(Reimers & Gurevych, 2019)** we showed how pre-trained transformer networks can be tuned for text embeddings and retrieval, leading to significant improvements over previous approaches. The resulting software `sentence-transformers`[2] is widely adapted by the community with over 5 million downloads since the release[3]. In follow-up work, we extended the approach to multiple languages **(Reimers & Gurevych, 2020)**. As the domain of this project is highly specialized and getting sufficient amount of training data is difficult, we proposed new state-of-the-art methods to learn retrieval for low resource domains **(Thakur et al., 2021; Wang et al., 2021)**. Further, we addressed the issue of dense retrieval methods on large datasets and showed that these profit from higher dimensional vector spaces **(Reimers & Gurevych, 2021)**.

Work Package 4: Extracting atomic statements

In WP4 we aim to extract individual statements, or *propositions*, from given input sentences. We started addressing this task by developing PropS **(Stanovsky et al., 2016)**, a deterministic conversion which simplifies dependency trees by conveniently marking a wide range of predicates (e.g, verbal, adjectival, nonlexical) and positioning them as direct heads of their corresponding arguments. From the PropS representation it is easy to read out individual propositions, in an Open Information Extraction (Open IE) style.

In addition to showing several applications of PropS for the English language, in **(Falke et al., 2016)** we show that the deterministic rules of PropS' can also be effectively adapted to the German language to create a first German Open IE system.

Following, we explored novel linguistic and semantic aspects of proposition extraction which are required for obtaining *correct* and *minimal* propositions. These projects, reported below, investigated several different such aspects, while we aim to integrate them in follow up research in order to improve extraction quality.

Coordination To correctly extract atomic propositions from coordination structures, it is needed to properly recognize their internal structure. Despite being a prevalent and complex phenomenon, certain aspects of coordination were often overlooked in previous efforts.

For example, consider the coordination (denoted by brackets) in the sentence: *"Some [students and parents] [believe this kind of education cannot nurture a student thoroughly] and [choose to go to an alternative school]"*.

Indeed, Current state-of-the-art syntactic parsers often fail on such constructions.

In **(Ficler & Goldberg, 2016b)** we improved parsing accuracy on one sub-type of coordination construction, called argument-cluster coordination. In **(Ficler & Goldberg, 2016a)**, we created a large annotated corpus of annotated coordination structures, together with the complete syntactic trees of the sentences.

In **(Ficler & Goldberg, 2016c)** we build upon this newly created resource and propose a state-of-the-art model for automatic disambiguation of the boundary of coordinated elements in coordination structures.

Non-restrictive modification Using minimal propositions[4] was shown to be extremely useful in various settings (e.g., in knowledge base population (Angeli et al., 2015), word similarity **(Stanovsky et al., 2015)** or Open Information Extraction (Del Corro & Gemulla, 2013)). In our setting, we aim to represent minimal propositions, allowing to perform more substantial information consolidation across such minimal units of information. Specifically, in **(Stanovsky & Dagan, 2016a)** we explore the reduction of propositions through the removal of *non-restrictive modifications*.

Compare, for example, the restrictive (obligatory) modifier in example (1), versus the non-restrictive (parenthetical) modifier in example (2): (1) She wore the necklace *that her mother gave her*; (2) The speaker thanked president Obama *who just came into the room*.

This reduction often needs to take into account subtle lexical-syntactic and semantic cues, such as the word introducing the modifier, or whether the modifier clause is preceded by a comma. In **(Stanovsky & Dagan, 2016a)** we develop a state-of-the-art CRF model that uses these features to classify the restrictiveness of modifiers.

[2] `https://github.com/UKPLab/sentence-transformers`

[3] `https://pepy.tech/project/sentence-transformers`

[4] For example, using the argument span "Barack Obama" instead of "Barack Obama, the former U.S. president".

Factuality Detection Identifying the commitment of the author towards the assertions in the text (also known as *factuality detection*) is a crucial stage in populating our Open Knowledge Representation (OKR) graphs. Specifically, we would like to admit factual statements (*"John went home"*) into the knowledge base, while hypothetical or negated ones should be considered only with the appropriate context (*"Mary denied that John might have gone home"*). Previous models for factuality prediction were trained and tested against a specific annotated dataset, subsequently limiting the generality of their results. In **(Stanovsky et al., 2017)** we propose an intuitive method for mapping three previously annotated corpora (Minard et al., 2016; Lee et al., 2015; Saurí & Pustejovsky, 2009) onto a single factuality scale, thereby enabling models to be tested across these corpora. In addition, we design a novel factuality model, shown to perform well across corpora, by extending a previous rule-based factuality prediction system (Lotan et al., 2013), and then using the output of this system within a supervised classifier.

Large gold standard corpus for Open Information Extraction As we described above, PropS provides Open IE style extractions via a deterministic rule-based approach, similarly to the vast majority of other Open-IE systems (e.g., (Mausam et al., 2012; Fader et al., 2011)). While this contributes to robustness across domains, it also limits performance by prohibiting the use of latest advancements in machine learning models for NLP. The lack of such learning models is due to the fact that Open Information Extraction was missing an independent large gold standard corpus. In the context of this project, this also inhibits an objective comparison of alternative proposition extraction models. In **(Stanovsky & Dagan, 2016b)**, we therefore created a first large Open IE corpus. We demonstrated that Question Answering Driven Semantic Role Labeling (QA-SRL) (He et al., 2015), a recent variation on Semantic Role Labelling (SRL), can be effectively and deterministically converted to Open IE notation. In a follow up work (currently under review), we leverage our recently created corpus and propose a novel supervised Open IE method. We model Open IE as a word transducion task while borrowing and adapting ideas from recent work for neural SRL (Roth & Lapata, 2016; Zhou & Xu, 2015).

QASem As an extension to QA-SRL (He et al., 2015; Fitzgerald et al., 2018), we developed question and answer pairs representations to decompose and represent propositions, using natural language. Such a soft annotation approach has benefits for facilitated, large-scale crowdsourcing, or interpretability **(Roit et al., 2020)**. We developed a scheme to represent deverbial nominalizations and their semantic arguments using QA-pairs **(Klein et al., 2020)**. Similarly, we proposed to represent discourse relations as QA-pairs between two clauses (e.g. one clause is part of the question and the other part of the answer) **(Pyatkin et al., 2020)**. Currently we are working on extending the scheme for adjectives. We aim to have a comprehensive and unified representation of all propositions in text, using QA-pairs. We showed the benefits of QASem in our paper on generating information-seeking questions, where the QA-SRL and QANom format and datasets helped in creating prototypical role-questions **(Pyatkin et al., 2021)**.

Work Package 5: Consolidating Statements

This work package aims to merge a set of "atomic" statements (propositions), extracted in WP4, into a graph-based structure that represents the consolidated information expressed jointly by the original statements. In our original proposal, we envisioned the consolidated structure to take the form of a statement entailment graph, whose nodes represent clusters of original statements with equivalent meaning, while edges represent (non-equivalence) semantic relations between these statements, such as entailment or contradiction.

While investigating our original approach, we realized that beyond capturing the original extracted statements, we should represent also additional statements that may be entailed from the given ones. This led us to examine the operations of sentence union and intersection (Marsi & Krahmer, 2005), which generate such entailed statements, having in mind generating a graph that is based on these operations.

Starting with investigating intersection operation, in **(Levy et al., 2016)**, we identified several issues with existing literature. In particular, we found that existing intersection datasets are either too small or are not reliably annotated. This situation seems to stem mainly from the the costly expert annotation which is currently required for the fairly complex and subtle task guidelines. Our work attempted to mitigate this dependence on trained annotators by proposing a new method for crowd-annotating sentence intersection, which essentially breaks its annotation into several smaller, crowdsourceable tasks. Further, this decomposed annotation methodology lays the ground for analogous automated algorithms for the task. We published a corpus of 1,764 sentence pairs, annotated for intersection via crowdsourcing using our novel methodology, and provided useful analysis of the intersection phenomenon.

The above initial investigation of sentence intersection was applied to a corpus of sentence pairs. How-

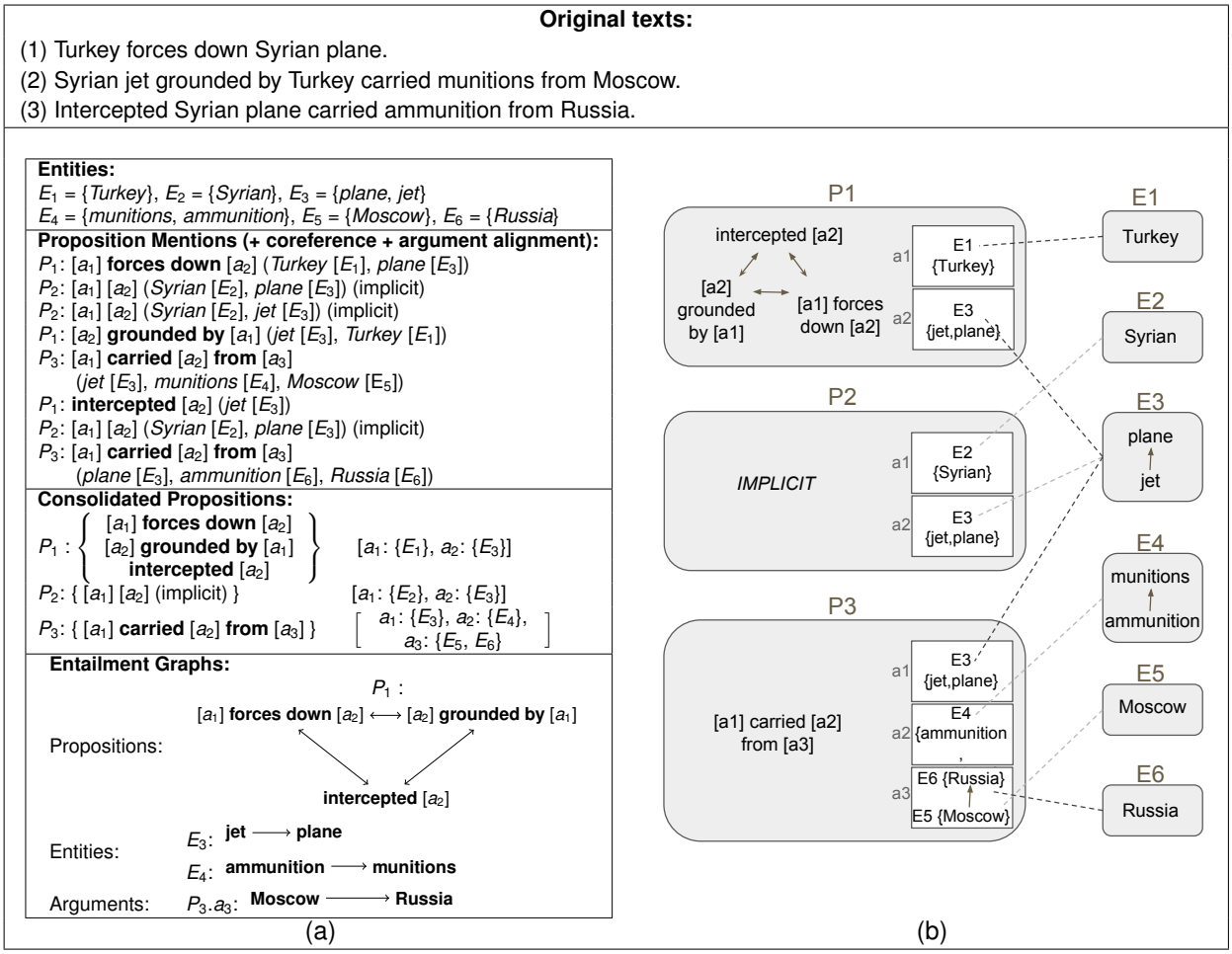

Figure 1: An illustration of our OKR formalism **(a)**, with a corresponding graphical view of the consolidated structure **(b)**. In (b), dashed lines connect entities to their instantiation within arguments, while allowing graph-traversal inferences such as: what is the relation between *Turkey* and *Russia*? *Turkey* intercepted a plane that carried ammunition from *Russia* (the path from E_1 to E_6 via the darker dashed lines).

ever, when annotating multiple texts describing the same event, we found that this annotation methodology would not be scalable, as it should be applied to every pair of original statements.

Yet, during our attempts to annotate intersection for sets of statements, we figured out that a consolidated representation may be feasibly generated, both in manual annotation and automatically, by considering the individual lexical elements within the given statements. Consequently, we developed a representation termed Open Knowledge Representation (OKR), described in detail in **(Wities et al., 2017a)**, Here, we consolidate information at the level of entities and predicates, based on identifying coreference and lexical entailment links between them. More specifically, we first extract proposition mentions, each composed of a single predicate and an arbitrary number of arguments. We then merge these mentions based on proposition coreference (an extended notion of event coreference). This process yields consolidated propositions, each corresponding to a single fact, or assertion, in the described scenario. Similarly, entity coreference links are used to establish reference to real-world entities. Taken together, our proposed representation encodes information about events and entities in the real world, similarly to what is expected from structured knowledge representations. Yet, being an open text-based representation, we record the various lexical terms used to describe the scenario. Furthermore, we model information redundancy and containment among these terms through lexical entailment. The entailment relation particularly lends itself to interactive text exploration (in WP7), by imposing a specific-to-generic ordering between mentions of consolidated entities and predicates. This allows users to *drill down* into more specific information regarding the explored information elements. The resulting structure is illustrated in Figure 1 (see **(Wities et al., 2017a)** for detail).

The OKR structure we defined successfully consolidates the complete information about all propositions,

across all their mentions in the multiple input texts, while keeping track of the entailment relations between different information elements. It thus provides a suitable basis for the interactive exploration phase, described in WP7. As described earlier for WP1, the OKR representation proved to allow feasible human annotation, resulting in a first medium-size annotated corpus. In addition, as detailed in **(Wities et al., 2017a)**, we implemented and evaluated initial automation of the component tasks involved in OKR generation, based on certain modules developed WP4 and WP6, . as well as additional available and implemented tools. Recently, we have integrated these components to yield a fully automated pipeline that generates OKR structures.

Open-domain CDCR systems need to be sufficiently general to be applied on arbitrary sets of documents. To better understand the current progress, we surveyed existing CDCR models and performed deep analysis on three established corpora with compatible event definition (ECB+ (Cybulska & Vossen, 2014), GVC (Vossen et al., 2018), and FCC). By applying an interpretable system using a wide range of handcrafted features relating to the action, participants, time and location of events, we measured the importance of these four aspects for resolving CDCR on the three corpora in practice. We found that due to their structure and annotation design decisions, neither corpus alone is sufficient to test a system's skills on all four aspects, leading us to recommend evaluation across as many available corpora as possible for future work in this area.

Additionally, we focused on improving the coreference resolution component. In **(Barhom et al., 2019)** we propose a neural architecture for cross-document coreference resolution. Inspired by (Lee et al., 2012), we jointly model entity and event coreference. We represent an event (and similarly entity) mention using its lexical span, surrounding context, and relation to entity (event) mentions via predicate-arguments structures. Our joint model achieves state-of-the-art results, outperforming the previous state-of-the-art event coreference model on the ECB+ dataset, with a gap of 10.5 CoNLL F_1 points. Our work is also the first to provide entity coreference results on this corpus.

We further focused on the development of more expressive CD coreference annotation guidelines and tooling to extend the applicability and consistency of coreference resolution in consolidating raw text statements across multiple documents. To ensure cheaper and more efficient annotation, we developed CoRefi, a web-based coreference annotation suite, oriented for crowdsourcing. Beyond the core coreference annotation tool, CoRefi provides guided onboarding for the task as well as a novel algorithm for a reviewing phase. CoRefi is open source and directly embeds into any website, including popular crowdsourcing platforms. Our experiments demonstrated that CoRefi's automatic onboarding is effective at augmenting controlled crowdsourcing **(Roit et al., 2020)**. Overall, we demonstrated that non-expert annotators can be trained to effectively perform and review coreference annotations, allowing for cost-effective empirical experimentation to refine the description of otherwise complex decisions.

Continuing this line of work, we next focus on identifying and aligning information that should be consolidated, i.e. information that is repeated in multiple different texts. Specifically in a multi-text setting, we often encounter redundant information that is longer than single word entities or events, but rather full propositions. To tackle this, we define a new task focused on aligning predicate-argument relations (i.e propositions) in a multi-text setting **(Brook Weiss et al., 2021)**. This task addresses pairs of input sentences that contain overlapping information (express similar content), and aligns predicate-argument relations that express the same meaning. This is done by aligning QA-SRL question-answer sets across the two sentences that express the same content. We use Mechanical Turk to train a small pool of workers and to collect our initial "gold" set for this task. We find that although semantically challenging, the task is learnable by lay-man workers (although trained using **(Roit et al., 2020)**'s methodology) and achieves a 70 F1 pairwise agreement between workers (not sure if this is needed). We compare our methodology to similar research lines such as event and entity coreference, as well as previous similar alignment works that utilized the predicate-argument structures. In addition, we have also worked on creating a dataset for a sentence fusion task, which takes as input multiple similar sentences, and creates a fused sentence that expresses a loose intersection of the input. We hope that once we have a baseline model for this task, the predicate-argument alignment will be utilized for this fusion dataset, demonstrating an extrinsic utility for these alignments and shedding light on the information consolidation task.

We observe that research on cross-document (CD) coreference has been lagging behind the impressive strides made in within-document coreference. As the time seems ripe to promote advances in CD coreference modeling as well, we present two steps to facilitate and trigger such systematic research, with respect to proper evaluation methodologies and current modeling approaches. In our first work **(Cattan et al., 2021b)**, we point out that common evaluation practices for cross-document coreference resolution

have been unrealistically permissive in their assumed settings, yielding inflated results. We propose addressing this issue via two evaluation methodology principles. First, as in other tasks, models should be evaluated on predicted mentions rather than on gold mentions. Doing this raises a subtle issue regarding singleton coreference clusters, which we address by decoupling the evaluation of mention detection from that of coreference linking. Second, we argue that models should not exploit the synthetic topic structure of the standard ECB+ dataset, forcing models to confront the lexical ambiguity challenge, as intended by the dataset creators. When evaluated under our realistic evaluation principles, we observe a significant drop in performance (33 F1), better pointing at weaknesses that future modelling work could explore. In our parallel work **(Cattan et al., 2021a)**, we develop the first end-to-end model for CD coreference resolution from raw text, which extends the prominent model for within-document coreference (Lee et al., 2017) to the CD setting. Our model achieves competitive or state-of-the-art results for event and entity coreference resolution on gold mentions. More importantly, we set first baseline results, on the standard ECB+ dataset, for CD coreference resolution over predicted mentions. Further, our model is simpler and more efficient than recent CD coreference resolution systems, while not using any external resources.

Work Package 6: Acquiring Domain-Specific Entailment Rules

This work package is intended to support the inferences required in WP5, by providing information about *entailment relations* (Dagan et al., 2013) between mentions of consolidated elements (predicates and entities). Recognizing entailment serves two purposes: first, it allows constructing the entailment graphs for entities and predicates, which are part of the OKR representation (see WP5); second, recognizing the (context dependent) lexical entailment relations between text elements from different documents can greatly help in the task of *cross-document coreference*, for both entities and predicates, as entailing elements are also likely to co-refer. For example, consider the following co-referring and often entailing entities: “The Russian government”, “Kremlin”, “Moscow”.

Aiming to improve available techniques, we investigated various aspects of lexical entailment models:

Analyzing lexical inference One seemingly promising approach for acquiring lexical inference knowledge is the relatively recent introduction of neural word embeddings, such as word2vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014). We explored these advances by contrasting them to the previously well-studied word vectors from distributional semantics, and found that the two approaches are highly-related **(Levy et al., 2015a)**. Specifically, we found that neural algorithms tend to incorporate new hyperparameters that boost their performance on a collection of word similarity benchmarks. We show that these hyperparameters can also be adapted to the traditional methods, yielding similar improvements. We also explored how word embeddings, as well as traditional distributional representations, can be used for lexical inference. In particular, we analyzed a set of supervised methods for lexical inference that are based on word vectors as inputs **(Levy et al., 2015b)**. This analysis showed some inherent issues with existing supervised methods, and demonstrated that they are not, in fact, learning to recognize lexical inference. Finally, due to its important role in textual inference, we studied the hypernymy relation in great detail and performed a survey of unsupervised hypernymy detection measures **(Shwartz et al., 2017a)**.

Novel models for lexical inference In **(Shwartz et al., 2015)**, we presented a novel supervised model for lexical inference which achieves high precision by integrating information from high quality external knowledge bases (e.g., DBPedia (Lehmann et al., 2014) or Wikidata (Vrandečić & Krötzsch, 2014)). In a follow up line of work **(Shwartz et al., 2016; Shwartz & Dagan, 2016b; 2016c)**, we took a corpus-based approach, and developed several closely related neural models, which integrate path-based and distributional models. We show that this integration yields state of the art results on several benchmarks, including a recent shared task on classifying lexical semantic relations (Shwartz & Dagan, 2016b).

Lexical inference in context – datasets and models Language is context-sensitive, and this can greatly impact any inference method’s ability to perform in a real-life scenario. To address this issue, which did not receive significant attention in previous literature, we constructed a dataset of fine-grained lexical inference relations in context **(Shwartz & Dagan, 2016a)**, and developed a series of unsupervised *context-sensitive* inference methods **(Melamud et al., 2015a; 2015b; 2016)**. Predicates are perhaps even more sensitive to context than other linguistic items, making them an appropriate benchmark for context-sensitive inference methods. To that end, we created a new dataset of over 16,000 examples of predicate inference in context **(Levy & Dagan, 2016)**. This dataset was collected by emulating a question-answering system, while crowdsourcing the actual inference decisions. The process avoids biases that existed in previous datasets, and provides a much more challenging and realistic benchmark. In a closely related effort, we explore semantic relations between predicates, also including temporal and causal relations **(Sukhareva et al., 2016)**. To

Figure 2: Our knowledge exploration system's initial view of a summary covering 109 tweets. Ten generated sentences cover the information throughout these tweets, and are ordered along the event timeline.

this end, we built a dataset of verbal predicates, consisting of 12,403 examples of semantic verb relations in context. To annotate this dataset, we developed a new semantic verb relation scheme and designed a multi-step annotation approach for crowdsourcing. The resulting annotated dataset serves as a challenging benchmark for identifying temporal or causal relations between verbs.

To predict and rank predicate paraphrases, in **(Meged et al., 2020)** we first annotated Chirps **(Shwartz et al., 2017b)** - a predicate paraphrases resource, using distant supervision based on a gold annotation of event coreference resolution dataset, then we extracted features from the Chirps resources and trained a scorer based on the labels and the extracted features. The new scoring gained more than 18 points in average precision upon their ranking by the original scoring method. Next, we integrated the features into an event coreference-resolution model, which improved the state-of-the-art results by 0.5 points.

Work Package 7: Knowledge Exploration

In this work package we designed and implemented an interactive knowledge exploration scheme, with a corresponding GUI, which allows users to navigate the knowledge encoded in the OKR structure. Summary sentences are generated, and are presented in a bullet-style format, presenting the consolidated information from the original texts. Several interaction modes are supported by our scheme, providing an "interactive abstractive summary".

Figure 2 depicts the GUI of our exploration system. Only the most salient summary sentences are shown, while the rest are hidden until unfolded (drilled into) by the user. An additional interaction mode is *concept expansion*, where coreferring terms of entities and predicates throughout the texts are shown when hovering over a concept, providing additional information about it.

Supporting tweets from which summary sentences were generated can be displayed by request.

In addition, each sentence, as well as the currently displayed summary, are associated with a "knowledge coverage" score, displayed graphically (circles on the right).

A paper describing this our novel exploration scheme, its implementation and supporting usability studies (see WP8) was accepted as a publication **(Shapira et al., 2017)**. The initial exploration system served as a prototype. Next, we seeked to formally define the task of interactive summarization to enable orderly evaluation and advancement of this line of research. As interactive summarization consists of incrementally growing text being presented to a user, we researched and released a paper on the use of existing summarization datasets for the evaluation of the different length summaries **(Shapira et al., 2018)**. Furthermore, as a manual summary evaluation method, we found that the effective Pyramid method (Nenkova & Passonneau, 2004) can be simplified and made crowdsourcable, at a slight cost in reliability, as a way for measuring the information gain in the growing summaries in the interactive setting **(Shapira et al., 2019)**. Finally, building upon the previous findings, we formally defined the task of expansion-based interactive summarization, and accordingly designed an evaluation framework that is relatively attainable by researchers interested in

Task	Entity Ment.	Entity Co-reference				Prop. Mentions		Proposition Co-Reference								Entailment	
						Pred.	Arg.	Predicate				Argument				Entity	Prop
	avg. acc	MUC	B^3	CEAF	CoNLL F_1	avg. acc	avg. acc	MUC	B^3	CEAF	CoNLL F_1	MUC	B^3	CEAF	CoNLL F_1	F_1	F_1
Pred	.58	.84	.89	.81	.85	.41 (.73, .25)†	.37	.47	.67	.56	.56	.93	.97	.94	.95	.44	.56

Table 1: Predicted performance for the OKR pipeline modules: (1) Entity mention extraction (F1 score) (2) Entity co-reference (standard coreference performance measures) (3) Proposition Extraction (predicate identification and argument detection) (4) Proposition Co-reference (predicate coreference and argument alignment), and (5) Entailment graphs (entity and proposition entailment). † Numbers in parenthesis denote verbal vs. non-verbal predicates, respectively.

working in the field **(Shapira et al., 2021)**.

Work Package 8: System Evaluation

We conducted two primary evaluations, accompanied with corresponding analyses: evaluating the performance of our initial automated pipeline for creating OKR structures and a usability study of our interactive knowledge exploration system.

Pipeline Evaluation The modules of the automated pipeline for creating the OKR structures from input sentences (WP5, with modules from WP4 and WP6) were evaluated using our gold standard corpus (WP1).

Table 1 shows the evaluation metrics of the predicted results of the subtasks in the pipeline. The evaluated modules, the results and analyses are described in detail in **(Wities et al., 2017b)**. Overall, performance of our first version of the pipeline is reasonable in the majority of tasks, while leaving room for our forthcoming research to improve upon various modules. From the text exploration point of view, we aim to tune our modules for precision, while sacrificing some recall, relying on the fact that important information elements typically has multiple mentions and hence are likely to be identified at least in some of them.

In our most recent work on interactive summarization **(Shapira et al., 2021)**, the proposed evaluation scheme strives to overcome the complicated evaluation methods applied in interactive systems, often being a bottleneck for the rapid advancement of the field. Instead of requiring expensive and subjective organized expert user studies, we introduce a crowdsourcable approach that provides comparable scores. In so, any work can be evaluated, standalone, by a feasible crowdsourcing procedure, and then compared to other works that similarly applied the schema. At a high level, crowdworkers are given a relatively objective persona around which they use the interactive system to explore a topic. The information gain is computed over the course of each such session, giving both automatic and manual absolute scores.

Knowledge Exploration Usability Study To assess and improve our knowledge exploration schemata (WP7), we conducted usability studies employing standard Information Science methodologies. In **(Shapira et al., 2017)** we ran two studies. In the first, users were asked to perform tasks on the interface and filled the SU Scale (SUS) questionnaire (Brooke, 1996) for subjective usability evaluation. Overall, users found the prototype easy to use and showed willingness to use it frequently. The observation and verbal reports during the test yielded additional requirements that we then implemented to improve our system. A second comparative study examined the relative effectiveness of our system against two baselines, of viewing either the original texts or a comprehensive static summary. After using all interfaces (on different texts), users were asked to complete a USE Questionnaire (Lund, 2001), which measured dimensions of Usefulness, Satisfaction, Ease of Use, Ease of Learning and an additional Knowledge Exploration dimension that we added. Our system consistently received the highest ranks in the dimensions of Usefulness, Satisfaction and most prominently Knowledge Exploration. These results indicated the value of our novel exploration approach and its potential exploitation within information access applications. We similarly analyzed the effectiveness of our IFACETSUM application **(Hirsch et al., 2021)** through two experiments with human subjects. In the first experiment, we conducted a pilot usability study to inspect whether users felt they were able to satisfactorily complete an information seeking task using our system. In the second, we examined whether IFACETSUM is preferred over a standard document-search system to complete the exploration task. Finally, as an initial investigation for designing our interactive summarization evaluation framework (Shapira et al., 2021), we ran a small-scale user study, similar to the first study described above. This facilitated gathering impressions on how to devise a proper crowdsourcing task that would mimic the advantages of frontal user studies, e.g., the controlled environment for high quality assessment, while capitalizing on the relative scalability and accessibility of crowdsourcing.

Evaluation of Deep Neural Networks Over the course of the project, the field switched from shallow classifiers like SVM to deep neural networks, leading to significant performance increases. However, training these neural networks is non-deterministic, which raises new challenges in the evaluation of systems. In **(Reimers & Gurevych, 2017)** we showed that performances can change significantly based on the random number generator. We extended this observation in **(Reimers & Gurevych, 2018)** and showed that previous evaluation methods, as used for shallow classifier, have severe shortcomings for neural networks. Instead, we proposed a new evaluation approach that takes the randomness of the training procedure into account. Finally, in **(Reimers & Gurevych, 2021)** we showed that the performances of dense retrieval approaches (cf. WP 3) degrade quickly for larger corpora, hence, a careful evaluation on realistically sized corpora is necessary.

Work Package 9: System Integration

This work package had a technical (rather than research) nature, and included two major integration efforts. The first was integrating the modules involved in generating the OKR structure, as defined in WP5 (including modules from WP4 and WP6) and evaluated in WP8. The second was importing the generated OKR structures, as well as the original texts, into the interactive exploration system (WP7). Both efforts included defining and implementing relevant data representation formats and software interfaces.

2.3 Conceivable Follow-up Research

A challenge remains that the developed methods are based on large annotated datasets. Extending this to other domains or other languages is labor and cost intensive, as new annotated datasets have to be created and trained from scratch. An important follow-up research direction is reducing the requirement for annotated data. This will enable these technologies to be used more broadly.

Further, the state-of-the-art methods are based on pre-trained transformer networks, which have been pre-trained on large corpora with billions of sentences. Such large corpora are only available for few high resource languages, making the usage of these systems for low resource languages difficult. To make the developed systems accessible for low resource language, more resource efficient pre-training methods must be developed.

Finally, an important research question remains on how to update the developed system with new world knowledge, like the election of a new president in a country. So far, systems learn world knowledge during the pre-training and the supervised fine-tuning. But keeping its pre-training and annotated corpus up-to-date with new world knowledge is difficult and costly: Hence, new approaches are needed that can update existing systems with new world knowledge.

2.4 Economic Value & Exploitation

The goal of the project was to do fundamental research and to share the results with the large scientific audience. Hence, all produced results (publications, datasets, software) are publicly available. Economic exploitation was not a direct goal of the project.

We are happy that several components of our research have been proven to be relevant enough to be adapted by companies. Sentence-Transformers **(Reimers & Gurevych, 2019)** is used by over 1600 public projects, and several smaller and large companies (e.g. T-Systems, Otto Group, Randstad, SeMI Technologies, DeepSet, Vespa.Ai, Zeta Alpha, and several more) use it to enhance their search functionalities.

Our expertise in the field of summarization and interactive systems enabled us to collaborate with Israeli industry requiring relevant components within new products. These projects went beyond the DIP report. In one such project, organized by the Israel Innovation Authority, we collaborated with several companies (including NICE, Microsoft and startups) and academic research groups (from the University of Haifa) to develop a topical Tweet collection summarizer for real time event retrieval. In another project, in collaboration with LiveU, we summarized single and multiple news articles for a social platform for field newscasters seeking live news events to cover. In these projects, our roles included developing relevant summarization components, and advising on language processing considerations for the project requirements.

3 Summary

Our project targets what might be one of the next big steps in information access technology: supporting knowledge-seeking users in quickly exploring and assimilating the primary content expressed in multiple relevant texts, such as many search results or a large set of tweets. This goal should be confronted with the current common practice, by which potentially relevant texts on a topic of interest become easily available, yet technology provides almost no help in sifting through the available information. The project took an

interdisciplinary approach, combining methods from natural language processing, information retrieval and information science.

In working to address this goal, our project has made the following key contributions. The first is developing a computational framework, called *Open Knowledge Representation* (OKR), for a consolidated representation of the information which is expressed in multiple texts. This structure somewhat resembles formal knowledge representation paradigms, where "building blocks" correspond to actual entities, concepts or events in the real world. While being convenient for computational processing, our *open* knowledge representation does not require tedious pre-specification of concepts and relations.Key properties of our OKR include conflating redundant information that is mentioned multiple times across the input texts while merging related complementary information that may be originally scattered across these texts. Further, the OKR keeps track of the level of detail, or information containment relationships, for different original text expressions. Taken together, our representation has the potential to be extremely useful for various applications that should consider joint information from multiple texts, such as question answering, multi-document summarization, information extraction and text mining. Towards this goal we also improved coreference resolution, specifically cross-document co-reference, where we achieved state-of-the-art performance and also proposed appropriate evaluation practices.

Our second key contribution is proposing an interactive scheme for text exploration. It is based on generating automatically an OKR structure for the explored set of texts, and then allowing the user to quickly explore the desired information which is embedded in this structure. The information is presented by generating simple natural language sentences, in bullet form, that express selected information from the underlying structure. The user can then interactively explore additional information and drill down into further detail. We implemented a first prototype of our scheme which allows to explore many news headlines about a news story. We also designed an evaluation framework for the task of expansion-based interactive summarization, which is crowdsourceable and provides comparable scores to expert annotators. The multiple usability studies that we ran, showed that users liked our approach and found it particularly useful for text exploration or completing information seeking tasks. We expect that this interactive text exploration approach, based on the consolidated open knowledge representation, would become a dominant direction in future information access technologies.

A third key contribution is the development of QASem, which comprises soft semantic annotations of question and answer pairs, representing propositions. Using these representations, we created and published numerous datasets, which target nominal and verbal predicates and discourse relations. We showed multiple benefits of these QA representations, among which facilitated crowdsourcing of semantic relations, cross-document proposition alignment and the generation of prototypical role questions. We believe that the QASem representations provide another promising way of exploring and consolidating propositions in text.

While working to achieve these key contributions, our research made additional scientific contributions to targeted component problems. Within information retrieval (search) technology, we developed novel methods for effectively retrieving relevant short *passages* within documents, rather than retrieving just full documents. In contrast to previous lexical methods, the developed methods focused on the semantics of the text and are able to bridge the lexical gap. We extended these methods to require less training data, making them more suitable for low resource domains, and made them more robust for domain shifts.

Next, we improved state of the art methods for extracting "atomic" facts, or statements, from complete sentences. Further, we developed novel methods that identify meaning relationships between different words and expressions in natural language. We also developed methods that identify the particular contexts in which a relation between two terms hold, for example, excluding the "key-piano" relationship when "key" refers to a door key. Finally we created a distantly supervised system to predict and rank predicate paraphrases, which showed considerable performance improvements compared to previous ranking systems.

Many of our research efforts developed in a gradual course, largely following the envisioned plans. Major obstacles were encountered, however, while seeking to develop a suitable knowledge representation scheme for consolidating textual information. In fact, our eventual OKR scheme, followed two earlier approaches that we attempted, one based on inference relations between pairs of statements and the other one is based on the intersection of information between such pairs. Both approaches turned out to be inappropriate for our goals, but gradually led us to figure out how to design our current approach, which so far seems feasible and appropriate.

4 Bibliography

Angeli, Gabor, Melvin Johnson Premkumar & Christopher D. Manning (2015). Leveraging linguistic structure for open domain information extraction. In *Proceedings of the 53rd Annual Meeting of the Association for Computational*

Linguistics (ACL 2015).

Barhom, Shany, Vered Shwartz, Alon Eirew, Michael Bugert, Nils Reimers & Ido Dagan (2019). Revisiting Joint Modeling of Cross-document Entity and Event Coreference Resolution. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4179–4189. Florence, Italy: ACL.

Brondwine, Elinor, Anna Shtok & Oren Kurland (2016). Utilizing focused relevance feedback. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '16, pp. 1061–1064. New York, NY, USA: ACM.

Brook Weiss, Daniela, Paul Roit, Ayal Klein, Ori Ernst & Ido Dagan (2021). Qa-align: Representing cross-text content overlap by aligning question-answer propositions. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Brooke, John (1996). *SUS - "A quick and dirty" usability scale. Usability evaluation in industry*. CRC Press.

Bugert, Michael, Nils Reimers, Shany Barhom, Ido Dagan & Iryna Gurevych (2020a). Breaking the subtopic barrier in cross-document event coreference resolution. In *Text2Story@ECIR*, pp. 23–29.

Bugert, Michael, Nils Reimers & Iryna Gurevych (2020b). Cross-document event coreference resolution beyond corpus-tailored systems. *arXiv preprint*.

Cattan, Arie, Alon Eirew, Gabriel Stanovsky, Mandar Joshi & Ido Dagan (2021a). Cross-document coreference resolution over predicted mentions. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 5100–5107. Online: Association for Computational Linguistics.

Cattan, Arie, Alon Eirew, Gabriel Stanovsky, Mandar Joshi & Ido Dagan (2021b). Realistic evaluation principles for cross-document coreference resolution. In *Proceedings of the Joint Conference on Lexical and Computational Semantics*. Association for Computational Linguistics.

Cybulska, Agata & Piek Vossen (2014). Using a sledgehammer to crack a nut? Lexical diversity and event coreference resolution. In *LREC*, pp. 4545–4552.

Dagan, Ido, Dan Roth, Mark Sammons & Fabio Massimo Zanzotto (2013). Recognizing textual entailment: Models and applications. *Synthesis Lectures on Human Language Technologies*, 6(4):1–220.

Del Corro, Luciano & Rainer Gemulla (2013). Clausie: clause-based open information extraction. In *Proceedings of the 22nd international conference on World Wide Web*, pp. 355–366, International World Wide Web Conferences Steering Committee.

Eirew, Alon, Arie Cattan & Ido Dagan (2021). Wec: Deriving a large-scale cross-document event coreference dataset from wikipedia. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2498–2510.

Fader, Anthony, Stephen Soderland & Oren Etzioni (2011). Identifying relations for open information extraction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 1535–1545, Association for Computational Linguistics.

Falke, Tobias, Gabriel Stanovsky, Iryna Gurevych & Ido Dagan (2016). Porting an open information extraction system from english to german.

Ficler, Jessica & Yoav Goldberg (2016a). Coordination annotation extension in the penn tree bank. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 834–842. Berlin, Germany: Association for Computational Linguistics.

Ficler, Jessica & Yoav Goldberg (2016b). Improved parsing for argument-clusters coordination. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 72–76. Berlin, Germany: Association for Computational Linguistics.

Ficler, Jessica & Yoav Goldberg (2016c). A neural network for coordination boundary prediction. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 23–32. Austin, Texas: Association for Computational Linguistics.

Fitzgerald, Nicholas, Julian Michael, Luheng He & Luke S. Zettlemoyer (2018). Large-scale qa-srl parsing. In *ACL*.

Habernal, Ivan, Maria Sukhareva, Fiana Raiber, Anna Shtok, Oren Kurland, Hadar Ronen, Judit Bar-Ilan & Iryna Gurevych (2016). New collection announcement: Focused retrieval over the web. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '16, pp. 701–704. New York, NY, USA: ACM.

He, Luheng, Mike Lewis & Luke Zettlemoyer (2015). Question-answer driven semantic role labeling: Using natural language to annotate natural language. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pp. 643–653.

Hirsch, Eran, Alon Eirew, Ori Shapira, Avi Caciularu, Arie Cattan, Ori Ernst, Ramakanth Pasunuru, Hadar Ronen, Mohit Bansal & Ido Dagan (2021). *iFACETSUM: Coreference-based Interactive Faceted Summarization for Multi-Document Exploration*.

Klein, Ayal, Jonathan Mamou, Valentina Pyatkin, Daniela Stepanov, Hangfeng He, Dan Roth, Luke Zettlemoyer & Ido Dagan (2020). Qanom: Question-answer driven srl for nominalizations. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 3069–3083.

Lee, Heeyoung, Marta Recasens, Angel Chang, Mihai Surdeanu & Dan Jurafsky (2012). Joint entity and event coreference resolution across documents. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp. 489–500.

Lee, Kenton, Yoav Artzi, Yejin Choi & Luke Zettlemoyer (2015). Event detection and factuality assessment with non-expert supervision. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Lisbon, Portugal: Association for Computational Linguistics.

Lee, Kenton, Luheng He, Mike Lewis & Luke Zettlemoyer (2017). End-to-end neural coreference resolution. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 188–197. Copenhagen, Denmark: Association for Computational Linguistics.

Lehmann, Jens, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N. Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick van Kleef, Sören Auer & Christian Bizer (2014). DBpedia - A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia. *Semantic Web Journal*.

Levy, Omer & Ido Dagan (2016). Annotating relation inference in context via question answering. In *The 54th Annual Meeting of the Association for Computational Linguistics*, pp. 249–255.

Levy, Omer, Ido Dagan, Gabriel Stanovsky, Judith Eckle-Kohler & Iryna Gurevych (2016). Modeling extractive sentence intersection via subtree entailment. In *Proceedings of the 26th International Conference on Computational Linguistics (COLING)*, pp. 2891–2901.

Levy, Omer, Yoav Goldberg & Ido Dagan (2015a). Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225.

Levy, Omer, Steffen Remus, Chris Biemann & Ido Dagan (2015b). Do supervised distributional methods really learn lexical inference relations? In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 970–976. Denver, Colorado: Association for Computational Linguistics.

Lotan, Amnon, Asher Stern & Ido Dagan (2013). Truthteller: Annotating predicate truth. In *HLT-NAACL*, pp. 752–757.

Lund, Arnold M. (2001). Measuring usability with the USE questionnaire. *STC Usability SIG Newsletter*, 8(2).

Marsi, Erwin & Emiel Krahmer (2005). Explorations in sentence fusion. In *Proceedings of the European Workshop on Natural Language Generation*, pp. 109–117, Citeseer.

Mausam, Michael Schmitz, Stephen Soderland, Robert Bart & Oren Etzioni (2012). Open language learning for information extraction. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp. 523–534. Jeju Island, Korea: Association for Computational Linguistics.

McMinn, Andrew J., Yashar Moshfeghi & Joemon M. Jose (2013). Building a large-scale corpus for evaluating event detection on twitter. In *Proceedings of the 22Nd ACM International Conference on Information & Knowledge Management*, CIKM '13, pp. 409–418. San Francisco, California, USA: Association for Computing Machinery.

Meged, Yehudit, Avi Caciularu, Vered Shwartz & Ido Dagan (2020). Paraphrasing vs coreferring: Two sides of the same coin. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 4897–4907. Online: Association for Computational Linguistics.

Melamud, Oren, Ido Dagan & Jacob Goldberger (2015a). Modeling word meaning in context with substitute vectors. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 472–482. Denver, Colorado: Association for Computational Linguistics.

Melamud, Oren, Jacob Goldberger & Ido Dagan (2016). context2vec: Learning generic context embedding with bidirectional lstm. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pp. 51–61. Berlin, Germany: Association for Computational Linguistics.

Melamud, Oren, Omer Levy & Ido Dagan (2015b). A simple word embedding model for lexical substitution. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pp. 1–7. Denver, Colorado: Association for Computational Linguistics.

Meyer, Christian M., Margot Mieskes, Christian Stab & Iryna Gurevych (2014). Dkpro agreement: An open-source java library for measuring inter-rater agreement. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: System Demonstrations*, pp. 105–109. Dublin, Ireland: Dublin City University and Association for Computational Linguistics.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Gregory S. Corrado & Jeffrey Dean (2013). Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pp. 3111–3119.

Minard, Anne-Lyse, Manuela Speranza, Ruben Urizar, Begona Altuna, Marieke van Erp, Anneleen Schoen & Chantal van Son (2016). Meantime, the newsreader multilingual event and time corpus. *Proceedings of LREC2016*.

Nenkova, Ani & Rebecca Passonneau (2004). Evaluating content selection in summarization: The pyramid method. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pp. 145–152. Boston, Massachusetts, USA: Association for Computational Linguistics.

Pennington, Jeffrey, Richard Socher & Christopher Manning (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543. Doha, Qatar: Association for Computational Linguistics.

Pyatkin, Valentina, Ayal Klein, Reut Tsarfaty & Ido Dagan (2020). Qadiscourse-discourse relations as qa pairs: Representation, crowdsourcing and baselines. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2804–2819.

Pyatkin, Valentina, Paul Roit, Julian Michael, Yoav Goldberg, Reut Tsarfaty & Ido Dagan (2021). Asking it all: Generating contextualized questions for any semantic role. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Reimers, Nils & Iryna Gurevych (2017). Reporting score distributions makes a difference: Performance study of LSTM-networks for sequence tagging. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 338–348. Copenhagen, Denmark: Association for Computational Linguistics.

Reimers, Nils & Iryna Gurevych (2018). Why comparing single performance scores does not allow to draw conclusions about machine learning approaches. *arXiv preprint arXiv:1803.09578*.

Reimers, Nils & Iryna Gurevych (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Reimers, Nils & Iryna Gurevych (2020). Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Reimers, Nils & Iryna Gurevych (2021). The curse of dense low-dimensional information retrieval for large index sizes. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pp. 605–611. Online: Association for Computational Linguistics.

Roit, Paul, Ayal Klein, Daniela Stepanov, Jonathan Mamou, Julian Michael, Gabriel Stanovsky, Luke Zettlemoyer & Ido Dagan (2020). Controlled crowdsourcing for high-quality QA-SRL annotation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7008–7013. Online: Association for Computational Linguistics.

Roth, Michael & Mirella Lapata (2016). Neural semantic role labeling with dependency path embeddings. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1192–1202. Berlin, Germany.

Saurí, Roser & James Pustejovsky (2009). Factbank: a corpus annotated with event factuality. *Language resources and evaluation*, 43(3):227.

Shapira, Ori, David Gabay, Yang Gao, Hadar Ronen, Ramakanth Pasunuru, Mohit Bansal, Yael Amsterdamer & Ido Dagan (2019). Crowdsourcing lightweight pyramids for manual summary evaluation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 682–687. Minneapolis, Minnesota: Association for Computational Linguistics.

Shapira, Ori, David Gabay, Hadar Ronen, Judit Bar-Ilan, Yael Amsterdamer, Ani Nenkova & Ido Dagan (2018). Evaluating multiple system summary lengths: A case study. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 774–778. Brussels, Belgium: Association for Computational Linguistics.

Shapira, Ori, Ramakanth Pasunuru, Hadar Ronen, Mohit Bansal, Yael Amsterdamer & Ido Dagan (2021). Extending multi-document summarization evaluation to the interactive setting. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 657–677.

Shapira, Ori, Hadar Ronen, Meni Adler, Yael Amsterdamer, Judit Bar-Ilan & Ido Dagan (2017). Interactive abstractive summarization for event news tweets. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 109–114. Copenhagen, Denmark: Association for Computational Linguistics.

Sheetrit, Eilon (2018). Utilizing inter-passage similarities for focused retrieval. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pp. 1453–1453.

Sheetrit, Eilon, Anna Shtok & Oren Kurland (2020). A passage-based approach to learning to rank documents. *Information Retrieval Journal*, 23(2):159–186.

Sheetrit, Eilon, Anna Shtok, Oren Kurland & Igal Shprincis (2018). Testing the cluster hypothesis with focused and graded relevance judgments. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pp. 1173–1176.

Shwartz, Vered & Ido Dagan (2016a). Adding context to semantic data-driven paraphrasing. In *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*, pp. 108–113. Berlin, Germany: Association for Computational Linguistics.

Shwartz, Vered & Ido Dagan (2016b). Cogalex-v shared task: Lexnet - integrated path-based and distributional method for the identification of semantic relations. In *Proceedings of the 5th Workshop on Cognitive Aspects of the Lexicon (CogALex - V)*, pp. 80–85. Osaka, Japan: The COLING 2016 Organizing Committee.

Shwartz, Vered & Ido Dagan (2016c). Path-based vs. distributional information in recognizing lexical semantic relations. In *Proceedings of the 5th Workshop on Cognitive Aspects of the Lexicon (CogALex - V)*, pp. 24–29. Osaka, Japan: The COLING 2016 Organizing Committee.

Shwartz, Vered, Yoav Goldberg & Ido Dagan (2016). Improving hypernymy detection with an integrated path-based and distributional method. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2389–2398. Berlin, Germany: Association for Computational Linguistics.

Shwartz, Vered, Omer Levy, Ido Dagan & Jacob Goldberger (2015). Learning to exploit structured resources for lexical inference. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, pp. 175–184. Beijing, China: Association for Computational Linguistics.

Shwartz, Vered, Enrico Santus & Dominik Schlechtweg (2017a). Hypernyms under siege: Linguistically-motivated artillery for hypernymy detection. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, p. to appear. Valencia, Spain: Association for Computational Linguistics.

Shwartz, Vered, Gabriel Stanovsky & Ido Dagan (2017b). Acquiring predicate paraphrases from news tweets. In *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (*SEM 2017)*, pp. 155–160. Vancouver, Canada: Association for Computational Linguistics.

Stanovsky, Gabriel & Ido Dagan (2016a). Annotating and predicting non-restrictive noun phrase modifications. In *Proceedings of the 54rd Annual Meeting of the Association for Computational Linguistics (ACL 2016)*.

Stanovsky, Gabriel & Ido Dagan (2016b). Creating a large benchmark for open information extraction. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Austin, Texas: Association for Computational Linguistics.

Stanovsky, Gabriel, Ido Dagan & Mausam (2015). Open ie as an intermediate structure for semantic tasks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics (ACL 2015)*.

Stanovsky, Gabriel, Judith Eckle-Kohler, Yevgeniy Puzikov, Ido Dagan & Iryna Gurevych (2017). Integrating deep linguistic features in factuality prediction over unified datasets. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL 2017)*. Vancouver, Canada.

Stanovsky, Gabriel, Jessica Ficler, Ido Dagan & Yoav Goldberg (2016). Getting more out of syntax with props. *CoRR*, abs/1603.01648.

Sukhareva, Maria, Judith Eckle-Kohler, Ivan Habernal & Iryna Gurevych (2016). Crowdsourcing a large dataset of domain-specific context-sensitive semantic verb relations. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, pp. 2131–2137. Portoroz, Slovenia: European Language Resources Association (ELRA).

Thakur, Nandan, Nils Reimers, Johannes Daxenberger & Iryna Gurevych (2021). Augmented SBERT: Data augmentation method for improving bi-encoders for pairwise sentence scoring tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 296–310. Online: Association for Computational Linguistics.

Vossen, Piek, Filip Ilievski, Marten Postma & Roxane Segers (2018). Don't Annotate, but Validate: a Data-to-Text Method for Capturing Event Data. In Nicoletta Calzolari (Conference chair), Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis & Takenobu Tokunaga (Eds.), *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Paris, France: European Language Resources Association (ELRA).

Vrandečić, Denny & Markus Krötzsch (2014). Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85.

Wang, Kexin, Nils Reimers & Iryna Gurevych (2021). Tsdae: Using transformer-based sequential denoising auto-encoderfor unsupervised sentence embedding learning. *arXiv preprint arXiv:2104.06979*.

Wities, Rachel, Vered Shwartz, Gabriel Stanovsky, Meni Adler, Ori Shapira, Shyam Upadhyay, Dan Roth, Eugenio Martinez Camara, Iryna Gurevych & Ido Dagan (2017a). A consolidated open knowledge representation for multiple texts. *LSDSem 2017*, p. 12.

Wities, Rachel, Vered Shwartz, Gabriel Stanowsky, Meni Adler, Ori Shapira, Shyam Upadhyay, Dan Roth, Eugenio Martínez Cámara, Iryna Gurevych & Ido Dagan (2017b). A consolidated open knowledge representation for multiple texts. In *Proceeddings of the Workshoop Linking Models of Lexical, Sentential and Discourse-level Semantics*, pp. 12–24, EACL.

Zhou, Jie & Wei Xu (2015). End-to-end learning of semantic role labeling using recurrent neural networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 1: Long Papers*, pp. 1127–1137.