Report from Dagstuhl Seminar 19182

Multi-Document Information Consolidation

Edited by

Ido Dagan[1], Iryna Gurevych[2], Dan Roth[3], and Amanda Stent[4]

1 **Bar-Ilan University – Ramat Gan, IL,** dagan@cs.biu.ac.il
2 **TU Darmstadt, DE,** gurevych@ukp.informatik.tu-darmstadt.de
3 **University of Pennsylvania – Philadelphia, US,** danroth@seas.upenn.edu
4 **Bloomberg – New York, US,** amanda.stent@gmail.com

Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 19182 "Multi-Document Information Consolidation". At this 5-day Dagstuhl seminar, an interdisciplinary collection of leading researchers discussed and develop research ideas to address multi-documents in machine learning and NLP systems. In particular, the seminar addressed four major topics: 1) how to represent information in multi-document repositories; 2) how to support inference over multi-document repositories; 3) how to summarize and visualize multi-document repositories for decision support; and 4) how to do information validation on multi-document repositories. General talks as well as topic-specific talks were given to stimulate the discussion between the participants, which lead to various new research ideas.

Seminar April 28–May 3, 2019 – http://www.dagstuhl.de/19182
2012 ACM Subject Classification Information systems → Information retrieval, Computing methodologies → Machine learning
Keywords and phrases Information Consolidation, Multi-Document, NLP
Digital Object Identifier 10.4230/DagRep.9.4.124
Edited in cooperation with Nils Reimers

1 Executive Summary

Ido Dagan (Bar-Ilan University – Ramat Gan, IL)
Iryna Gurevych (TU Darmstadt, DE)
Dan Roth (University of Pennsylvania – Philadelphia, US)
Amanda Stent (Bloomberg – New York, US)

Today's natural language processing (NLP) systems mainly work on individual text pieces like individual sentences, paragraphs, or documents. For example, most question answering systems require that the answer to a user's questions is provided in a single document, ideally in a single sentence. If the information is scattered across documents, most systems will fail. The capability of current systems to link information across multiple documents is often limited.

This is in strong contrast to how humans answer difficult questions or make complex decisions. We usually read multiple documents on a topic and then infer the answer to the question or we make a decision based on the evidence we found. In most cases, we consolidate the information across multiple sources. Further, considering only one document can create a biased or incomplete view on a topic. Many aspects in our life are open for

multiple interpretations and each author must limit which and how to present information in a document. By reading multiple documents, we are able to identify overlaps, differences, and opposing views between authors. Considering and merging these possible opposing views can be a crucial step in everydays decision making. For example, when booking a hotel, one might read multiple user reviews and create an internal understanding of positive and negative aspects of the hotel.

At this 5-day Dagstuhl Seminar, an interdisciplinary collection of leading researchers discussed and develop research ideas that will lead to advanced multi-document information consolidation systems and enable modern NLP systems to profit from a multi-document perspective.

The seminar was centered around four major themes: 1) how to represent information in multi-document repositories; 2) how to support inference over multi-document repositories; 3) how to summarize and visualize multi-document repositories for decision support; and 4) how to do information validation on multi-document repositories. Questions of semantics, pragmatics (author perspectives, argumentation), representation, and reasoning (including spatio-temporal reasoning and entailment) arose across these themes.

Information Representations and Inference are the theoretical foundation that allows systems to extract information from multiple documents and to infer new knowledge. The challenge is to find a representation that can broadly be used. Multiple documents are likely to bring up multiple perspectives and identifying the relations between them is at the heart of multi-document inference.

A connection to real applications, used in actual user scenarios, is critical for the advancement of the multi-document information consolidation field. Multi-document systems are especially useful in situations where users must make complex decisions. In such situations, users often search for sources that provide information or arguments for or against certain decisions. Hence, one working group focused on Multi-Document Systems in User Decision Scenarios. In order to provide value to users, the systems must return true statements (accurate syntheses) given all the available context. Otherwise, the user lose their trust in the system. However, the internet is full of statements that are intentionally or unintentionally misleading. So how do we identify these misleading statements and avoid that those are presented to a user without the necessary context? This research question was addressed by a working group focusing on Information Validation for Multi-Document Scenarios.

Seminar participants, including established experts and promising young researchers from academia and industry, had the opportunity to present research ideas, to outline their vision regarding the future of multi-document information consolidation technologies, and to collaborate in discussion groups led by the seminar organizers.

Each seminar participant joined two themes with regular cross-theme meetings. As the topics are quite novel in the research community, no established terminology and task definition exists. Hence, participants discussed how these tasks can be defined such that these can be scientifically studied. For example, what does it mean to validate a claim? The participants discussed issues with existing approaches and proposed new research topics, that could be the content of a Ph.D. thesis.

The last day of the seminar was used to summarize results and to create collaborations for future research projects. In total, 12 joint research ideas were proposed. For most of the ideas, this is a new collaboration.

2 Table of Contents

3 Invited Talks

3.1 Question-drive Information Consolidation

Jonathan Berant (Tel Aviv University, IL)

Humans often have complex information needs when performing activities such as learning about a new topic, performing research, or planning a future activity. Such scenarios invariably lead to questions that require deep understanding of questions and consolidation of information across multiple information sources. In this talk, I will present two lines of work focusing on the problem of answering complex questions, which require on-the-fly information consolidation. In the first thread, complex questions are handled by decomposing them into simpler questions and consolidating the information through symbolic operations. I will briefly describe past and ongoing work on building both models and datasets for question decomposition and question understanding. In the second thread, I will describe ongoing work on differentiable graphs, where information is represented with a graph structure, and information consolidation is performed with an end-to-end differentiable model over this graph. I will also discuss use cases in which these two opposing approaches are suitable.

3.2 Claim Validation by Humans and Machines: Where We Are and the Road Ahead

Iryna Gurevych (TU Darmstadt, DE)

Claim validation is a highly demanding expert task which prevents the proliferation of misinformation. In the recent years, we have seen a rapidly increasing interest in this problem domain. This interest is due to both the task significance and the impressive advances in AI-/NLP-based approaches. The talk will present novel datasets and problem definitions and experimental results related to automated claim validation. Information consolidation is an important, but yet untapped research direction for claim validation. Systems presenting just raw lists of evidences are insufficient to support humans in the challenging tasks of validating claim. We will conclude by outlining some open challenges for future research.

3.3 Consolidating Social, Behavioral and Textual Information

Dan Goldwasser (Purdue University – West Lafayette, US)

In this talk I will describe ongoing work, aiming to consolidate textual information, consisting of many interconnected documents, as well as social and behavioral information, capturing how these documents are shared and the reactions their contents receive. Formulating a broad definition of information consolidation which takes into account both aspects, would

allow us to answer questions about the social and behavioral context in which documents appear (i.e., "how to combine documents by the same author, to capture their perspective on a topic"), as well as exploit this structure to derive a supervision signal for identifying patterns in textual information (i.e., "how to exploit social information to identify that documents contain inconsistent information"). I will discuss our current efforts focusing on political discourse analysis on social media, online debate networks and partisan news analysis.

3.4 Multi-Document Summarization: from state-of-the-art to open research questions

Giuseppe Carenini (University of British Columbia – Vancouver, CA)

In essence, a multi-document summarizer is a system that takes as input a set of documents and generates a summary as output. Given this high-level view, we can start envisioning a design space for multi document summarization (MDS) by identifying key properties of the possible inputs, of the possible outputs and of the summarization process itself. In this talk, I will characterize such a design space, so that both the state of the art and open research questions in MDS can be better framed, discussed and understood.

3.5 Knowledge Base Population

Heng Ji (Rensselaer Polytechnic Institute – Troy, US)

Traditional Information Extraction techniques pull information from individual documents in isolation. However, in many real applications such as disaster management, intelligence analysis and scientific discovery, users might need to gather information that's scattered among multiple documents from a variety of sources. Complicating matters, these facts might be redundant, complementary, incorrect, or ambiguously worded; the extracted information might also need to augment an existing Knowledge Base (KB), which requires the ability to link events, entities, and associated relations to KB. This problem is called Knowledge Base Population (KBP). In this talk, I will introduce the state-of-the-art techniques for two core tasks in KBP: entity discovery and linking and slot filling, and discuss the remaining challenges and potential solutions. Then I will present several new research directions, including (1) moving from entity-centric KBP to event-centric KBP which requires event actuality extraction and truth finding across documents; (2) extend KBP to a multi-media multi-lingual paradigm; (3) background knowledge acquisition to enhance the quality of KBP capabilities.

4 Working Group – Information Representation

Many text-driven applications need to consider information that is consolidated across multiple texts. Such applications may benefit from an intermediate representation that effectively and informatively consolidates such cross-text information, making it more easily and uniformly accessible for downstream applications. The Information Representation workgroup discussed various aspects of developing such useful representation frameworks. The discussion covered challenges related to logical constructions, semantic phenomena and learning approaches, as well as potential tasks and datasets that could drive future research in this relatively unexplored space.

Throughout the sessions, four participants presented a short pitch related to multi-text information consolidation. Dipanjan Das explored "requirements" of a distributed representation for a human-computer conversation scenario. Keeping track of the multiple previous speech acts, together with their joint meaning, seem to be key aspects for delivering useful answers for user's open-domain questions. Sebastian Arnold suggested a vector space approach for representing local "hotspots" of selected aspects (e.g. topics or named entities) coherently over long documents, building on existing sentence embeddings and aligning them with the context of the document using distant supervision. Ivan Titov presented a recently proposed method for learning interpretable classification models, and speculated how it may be integrated with graph convolutional neural networks (GCNs), which are effective for integrating information across documents while relying on structured representations (e.g., coreference chains). Finally, Omri Abend specified challenges and insights raised in cross-language information consolidation. Since both the cross-language and the cross-document settings deal with linguistic realization diversity, i.e. different ways to express the same content, both confront similar phenomena, e.g. lexical differences, grammatical differences, different social connotations and different narrative styles.

Several important features for an explicit symbolic representation of multi-text information were brought up in discussions. Predicate-argument relationships were proposed consensually as a backbone of such semantic representations. Nevertheless, other layers of representation were deemed crucial. Cross- and intra- document Coreference for entities and events is a key component for identifying overlapping and complementary information. Temporal links between mentioned events, or a tidy timeline alignment of which, along with a set of discourse relations as causality and conditionality, are also essential for capturing the information conveyed by a set of texts. Aside from explicit denotation of specific semantic aspects, a notable core principle of information representation for multi-text consolidation was considered to be decomposability, that is, the breakdown of sentences into smaller meaning units, allowing for fine-grained cross-document alignment of "minimal" information units.

A major topic of discussion regarded the fundamental dichotomy of distributed (continuous) vs. explicit (symbolic) meaning representations. While contemporary contextualized vector-space representations have demonstrated great utility for many natural language understanding tasks, the multiple-text setting might benefit from the advantages of explicit representations. Specifically, the group enumerated several phenomena for which explicit representations would be desirable. These include logical aspects, such as quantification of entities and set membership (do "several blue and green pillows" correspond to "a dozen of colorful pillows"); capturing implicit entailed relations and arguments ("ex-wife" entailing a "divorce" event or status); and explicitly maintaining inference relations, such as entailment, equivalence or contradiction.

4.1 Talk – Representations for Open-Domain Conversation

Dipanjan Das (Google – New York, US)

We speculate about a scenario where a human is interacting with a system that can return answers to questions in a conversational scenario. In this talk, we explore "requirements" of a distributed representation that could serve as a "memory" for enabling this system.

4.2 Talk – Challenges in Cross-linguistic Information Consolidation

Omri Abend (The Hebrew University of Jerusalem, IL)

The talk discussed challenges that come up in cross-linguistic information consolidation. Many translation divergences (different ways of expressing similar content in different languages) also show up when consolidating information within a single language, underscoring the importance of this perspective. Examples discussed include lexical differences, grammatical differences, different social connotations and different narrative styles.

4.3 Talk – Distributed Representation of Local Information in Long Documents

Sebastian Arnold (Beuth Hochschule für Technik Berlin , DE)

This pitch talk introduces our vision of a neural document representation for multi-document passage retrieval. The challenge is to represent local "hotspots" of selected aspects (e.g. topics or named entities) coherently over long documents. Our current work on SECTOR utilizes existing sentence embeddings and aligns them with the context of a document using distant supervision. This allows us to retain the vector space of the embedding and retrieve coherent passages across multiple documents.

4.4 Talk – Towards Interpretability in Multi-Document Question Answering

Ivan Titov (University of Edinburgh, GB)

Graph convolutional neural networks (GCNs) are effective tools for integrating information across documents while relying on structured representations (e.g., coreference chains). Unfortunately, predictions of GCNs are hard to interpret and validate. In contrast, for classification

problems, we have recently proposed a method for effectively learning interpretable / sparse models. I speculate how merging the two ideas can lead to interpretable and also effective models for multi-document QA.

5 Working Group – Inference

The goal of this working group was to discuss problems, formulations, and possible approaches, that pertain to inference with respect to natural language text and, in particular, inference that arises in the context of dealing with multiple documents or multiple information sources. The presentations and discussions allowed us to develop better understanding of the keys issues involved in inference with multiple texts, develop important working examples, learn about existing research efforts, and identify research directions. In particular, we discussed and presented some of the existing datasets that could help drive future research in these directions.

5.1 Talk – Multi-passage Summarization for Query-specific Article Summarization

Laura Dietz (University of New Hampshire – Durham, US)

The TREC Complex Answer Retrieval track (TREC CAR) is a shared task about responding to web search requests with machine-constructed comprehensive articles. Such articles can be in the style of a Wikipedia page, how-stuff-works article, or grade-school textbook chapter. The purpose of this article is to inform the user about different important facets of the query. So far, our work is focused on the IR-side: (1) retrieving paragraph-length passages on the topic, (2) identifying which concepts/entities are central to the topic, and (3) arranging paragraphs into an outline.

I would like to use the opportunity of this Dagstuhl seminar focus on the multi-paragraph summarization aspects of this work — I am hoping to solicit some help/ideas/advice from the community. In return I can provide lots and lots of train/test data and intermediate results from the IR stage. While the shared task at TREC is focused on IR (ranking and selection of paragraphs), a holistic solution needs to also address challenges in multi-document summarization.

The pitch talk mostly focuses on the problem – not a solution. We have some initial data to demonstrate the difficulty of the problem. For example, ROUGE/ROUGE-SU is not a useful metric here; the word overlap of similar content is negligible; at the same time multiple subtopics are present, but difficult to extract and identify.

5.2 Talk – Inference in the age of DL?

Yoav Goldberg (Bar-Ilan University – Ramat Gan, IL)

I focus on the machine-learning sense of 'Inference', in which we are looking to solve an argmax problem over a large and somewhat structured space. This has been a major research area in structured prediction. Is this still needed in the deep learning era? I will take the provocative view that this is not needed, and that good enough networks model the inference as part of their learning process. The talk hopes to initiate discussion on this issue.

5.3 Talk – Top-down and bottom-up success in computational semantics

Alexander Koller (Universität des Saarlandes, DE)

A quick history lesson

- Back in the old days, when we did "computational semantics", inference meant "logical inference". The idea was to map sentences to formulas of predicate logic or some such (what is today called "semantic parsing") and then run a sound and complete theorem prover to perform the inference.
- Around 2000 there was much talk in the computational semantics community about shared tasks. The idea was repeatedly rejected because people didn't think about end-to-end tasks, but about mapping from language to specific semantic representations, and couldn't agree on a type of representation.
- Then Ido came along with "textual entailment", which was a shared task that computational semanticists should have been able to handle. But it turned out that the coverage issues were so severe that the old-school systems were useless, and these methods fell out of fashion very quickly.

Top-down vs bottom-up success

- The common view that old-school computational semantics (OSCS) has failed is an example of top-down thinking about scientific success. OSCS set its aims very high: to be able to understand all language that a human does; let's say, to answer all questions about a text that a human could. This goal is not achieved until it is achieved fully. Thus OSCS is "failed" because it did not achieve the end goal fully.
- The common view that we have recently made tremendous progress in NLU, including with respect to semantics, using neural methods is an example of bottom-up thinking of scientific success. There is a constant stream of new tasks and datasets on which neural methods have improved the state of the art; each of these counts as a success; the question of whether this gets us closer to any end goal is not a major issue.
- Good science needs both perspectives. Without the bottom-up perspective, progress is hard to make and quantify; without the top-down perspective, progress may climb the wrong hill. We need to stay humble and occasionally recalibrate by thinking about the end goal. Don't be too proud of this technological terror you've constructed.

- We need to define and work on tasks that strike a good balance between ambitious and doable, and maybe take more risks regarding the community's ability to solve the task within a year.

Datasets vs tasks

- There is a disturbing trend in recent NLP to define a specific dataset for a task, then train and evaluate models on this dataset, and call it a success if the model performs well on it. This makes a lot of sense, but only if the dataset reflects all the important aspects of the task. Often, though, the dataset is either very restricted (BaBI, Squad), or the distribution of the language in the dataset is disconnected from that in real text. We should be aware of this, make sure not to overinterpret results on such data, and work towards datasets that reflect the underlying task more and more accurately.
- Meaning has a lot of facets. Not all of these will be relevant for each task. Thus, it is really important to think about what task we're looking at before we decide which facets of meaning our formal representation needs to capture.

5.4 Talk – Abstractive Multi-Document Summarization: Opportunities and Challenges

Fei Liu (University of Central Florida – Orlando, US)

Joint work of Fei Liu, Kristjan Arumae, Logan Lebanoff, Kaiqiang Song, Kexin Liao, Sangwoo Cho

Humans can consolidate textual information from multiple sources and organize the content into a coherent summary. Can machines be taught to do the same? The most important obstacles facing multi-document summarization include excessive redundancy in source content, less understood sentence fusion, and the looming shortage of training data. In this talk I present our recent work tackling these issues through decoupling of content selection and surface realization.

We introduce a novel framework guiding extractive summarization (content selection) using question-answering rewards. We argue that quality extractive summaries should contain informative content so that they can be used as document surrogates to answer important questions, thereby satisfying users' information needs. The question-answer pairs can be conveniently developed from human abstracts. The system learns to promote summaries that are informative, fluent, and perform competitively on question-answering.

We further present an initial investigation into an adaptation method enabling an encoder-decoder model trained on single-document summarization data to work with multiple-document input. Parallel data for multi-document summarization are scarce and costly to obtain, therefore a low-cost adaptation method is highly desirable. Experimental results show that our system compares favorably to state-of-the-art extractive and abstractive methods judged by automatic metrics and human assessors.

Finally, we utilize structure-infused copy mechanisms to encourage salient source words and relations to be preserved in the summary, thereby preventing a summary from dramatically changing the meaning of the original text. I conclude the talk with a discussion of the challenges and opportunities associated with abstractive multi-document summarization.

5.5 Talk – Towards Brainstorming with Spoken Dialog Systems

Kentaro Torisawa (NICT – Kyoto, JP)

In this pitch talk, I'll talk about our spoken dialog system WEKDA, which can chat with users using a wide range of knowledge extracted from 4-billion Japanese Web pages. The knowledge extraction is done by our Web-based open-domain QA system WISDOM X, which provides answers to given questions using the 4-billion Web pages and has been publicly available since 2015 (https://wisdom-nict.jp/). WEKDA automatically generates questions for WISDOM X even from non-question inputs and composes responses to users based on WISDOM X's answers. The final goal of the WEKDA project is to enable it to conduct brainstorming with users through spoken dialogs, using knowledge extracted from a large collection of documents and hypotheses generated from the knowledge. As a future research plan, I'll discuss the possibility of using the auto-generated causal hypotheses in the brainstorming dialogs and list several technical problems.

6 Working Group – Information Validation

Multi-document systems often require the compression of information, as we often have millions of documents with different perspectives for a certain topic. However, how can we ensure that the condensed representation is actually true?

We face the challenge that a sheer amount of documents on every topic is available, and some documents will contain information that is intentionally or unintentionally misleading or plain wrong. Assessing the validity of information is a crucial step in multi-document information consolidation systems. Incorporating misleading or wrong information into a representation can have a snowball effect and many false statements could be inferred from this information. Finally, presenting clearly wrong statements to users can destroy the trust of the user into the system.

The working group started with identifying issues in information validation:

- Sources provide conflicting information, potentially with serious consequences
- Wrong facts are not limited to the political domain, but are also present in the medical domain Source may have agendas and motivations, leading to a biased or wrong presentation of information
- It is extremely difficult to differentiate between wrong information and legitimate opposing perspectives on a topic.

The discussion in the working group were accompanied by selected invited talks throughout the 5 days. Iryna Gurevych started with a talk on *Claim Validation by Humans and Machines: Where We Are and the Road Ahead*, which presented recent work on new datasets and problem definitions for claim validation and argument retrieval. Dan Roth presented the *Perspective Dataset*, which contains 1000 claims with different (potentially opposing) perspectives on these claims. Andreas Vlachos presented the *FEVER Shared Task*, with 185k claims verified on Wikipedia. Coreference resolution is a crucial step for find opposing views on a claim across source, hence Nafise Moosavi gave a talk about *More Applicable Coreference Resolvers* and their shortcomings in community question answering scenarios. The final talk was by Ayal Klein on *Minimal Statements in NL-based Semantic Representation*. Statements are

often embedded in long, complex sentence. Mapping those across documents can significantly be simplified, if they are mapped to minimal statements, containing one atomic information.

The working group spent time on defining and discussing future research directions and projects. Hereby, the group identified the following research questions as especially important to advance the field of claim validation:

- Realistic public dataset needed – For example, using BoolQ questions and rephrasing them to claims and using Wikipedia as a source of evidence
- Claim validation in the medical domain – A dataset could be constructed based on PubMed and provide scientific evidence for health-related claims
- Claim classification (e.g., factual, subjective, unverifiable, multi-perspective) – similar to question type classification, could help finding better strategies for claim validation
- Claim decomposition – How can a claim be decomposed into smaller units, which are easier to check?
- Controversial claims – How to design systems that find and presents opposing (but legitimate) views on a given, controversial topic?
- Interpretable results – How should a system reason about the own decision, which statements are credible and which are not?
- Removing (partial) redundancy in paraphrased evidences: this is a fundamental problem since the user wants to have a compact overview of all evidence.
- Evidence sufficiency: when is the set of evidence sufficient to resolve the claim? How to account for the sources' trustworthiness and speaker attribution?

6.1 Talk – Minimal Statements in NL-based Semantic Representation

Ayal Klein (Bar-Ilan University – Ramat Gan, IL)

For identifying the overlap of multi-document information, e.g. in the context of evidence aggregation, we should account for any information conveyed by a sentence. Such minimal information units can be captured by meaning representations that account for the semantic relations between sentence's concepts in neo-Davidsonian style graphs, e.g. AMR, SDP, etc. These formalisms are hard to apply for new domains, as they require supervised models and expert annotations. In this talk, I presented our ongoing effort of constructing a crowdsourcable semantic representation, extending the QA-SRL paradigm in which valuable semantic analysis of the sentence can be retrieved from laymen through simple tasks.

6.2 Talk – More Applicable Coreference Resolvers

Nafise Sadat Moosavi (TU Darmstadt, DE)

Coreference resolution has been recognized as an essential step for various tasks like question answering, summarization and fact checking. In order to benefit from coreference resolution in downstream tasks, we need to (1) discriminate coreference relations which would have more impact on target tasks, and (2) develop more generalizable systems since we do not have coreference annotations for downstream datasets. In this presentation, I briefly present our work in these two directions.

6.3 Talk – Perspective Dataset

Dan Roth (University of Pennsylvania – Philadelphia, US)

We construct PERSPECTRUM, a dataset of claims, perspectives and evidence, making use of online debate websites to create the initial data collection, and augmenting it using search engines in order to expand and diversify our dataset. We use crowdsourcing to filter out noise and ensure high-quality data. Our dataset contains 1k claims, accompanied by pools of 10k and 8k perspective sentences and evidence paragraphs, respectively. We provide a thorough analysis of the dataset to highlight key underlying language understanding challenges, and show that human baselines across multiple subtasks far outperform machine baselines built upon state-of-the-art NLP techniques.

6.4 Talk – FEVER Shared Task

Andreas Vlachos (University of Cambridge, GB)

Fact checking is the task of verifying a claim against sources such as knowledge bases and text collections. While this task has been of great importance for journalism, it has recently become of interest to the general public as it is one of the weapons against misinformation. In this talk, I will first discuss the task and what should be the expectations from automated methods for it. Following this, I will present our approach for fact checking simple numerical statements which we were able to learn without explicitly labelled data. Then I will describe how we automated part of the manual process of the debunking website emergent.info, which later evolved into the Fake News Challenge with 50 participants. Finally, I will present the Fact Extraction and Verification shared task, which took place in 2018 and our upcoming plans for the second edition.

7 Working Group – User Decision Support Systems

Decision makers frequently need to synthesize information across many documents for decision support. In NLP, these syntheses are typically static text summaries, however, there is increasing interest in interactive multimedia "summaries", such as timelines, graphs, or spatial visualizations, or extended information exploration dialogs. This group will focus on a taxonomy of, and best practices for, interactive decision support systems over multi-document repositories.

7.1 Talk – MultiConVis: A Visual Text Analytics System for Exploring a Collection of Online Conversations

Giuseppe Carenini (University of British Columbia – Vancouver, CA)

In this talk, I present MultiConVis, a visual text analytics system designed to support the exploration of a collection of online conversations. The system tightly integrates NLP techniques for topic modeling and sentiment analysis with information visualizations, by considering the unique characteristics of online conversations. The resulting interface supports the user exploration, starting from a possibly large set of conversations, then narrowing down to the subset of conversations, and eventually drilling-down to the set of comments of one conversation. Our evaluations through case studies with domain experts and a formal user study with regular blog readers illustrate the potential benefits of our approach, when compared to a traditional blog reading interface.

7.2 Talk – Real-time Twitter Analysis for Disaster Management

Kentaro Torisawa (NICT – Kyoto, JP)

We give demos of two large-scale NLP systems, DISAANA and D-SUMM, which were developed to help disaster victims and rescue workers in the aftermath of large-scale disasters. Immediately after disasters, much useful information is transmitted into cyberspace, especially for such social media as Twitter. Nevertheless, because most people are overwhelmed by the huge amount of information, they are unable to make proper decisions and much confusion ensued. DISAANA provides a list of answers to questions such as "What is in short supply in City X?" and displays locations related to each answer on a map (e.g., locations where food is in short supply) in real time using Twitter as an information source. D-SUMM summarizes the disaster reports from a specified area in a compact format and enables rescue workers to quickly grasp the whole situations from a macro perspective. We also show how the systems are used in actual disaster situations by Japanese local governments and how we are going to extend the whole framework by introducing so-called "chatbots" on chat apps.

8 Open problems

The group brainstormed about open research challenges in the four respective working areas.

8.1 Multi-Document Representations

- Research challenge: create challenge data sets and probe symbolic and distributed representations for handling of phenomena including: quantification/set membership; implicit relations/arguments; factuality; uncertainty; attribution
- Research challenge: from reports of sports games/controversies from different perspectives, or scientific findings reported in the media, or chains of reporting on an ongoing event – create a consolidated objective report

- Research challenge: representations that incorporate or produce discrete representations
- Research challenge: modeling coreference as part of "self-supervised" learning of representations
- Research challenge: (use wikipedia hyperlinks to) build a dataset that has pairs of paragraphs and a "hypothesis" that can be inferred from the consolidated paragraphs but not from the individual ones

8.2 Multi-Document Inference

- Research challenge: construct a multi-faceted summary to convey the information from the document repository to readers
- Research challenge: construct a summary to achieve complete understanding of a topic or event described in a document repository
- Research challenge: construct an update/timeline summary
- Research challenge: construct a deep abstract of source content without hallucination (where "deep" means neural?)
- S Combine symbolic and continuous semantics: because they are complementary How to combine? (1) use symbolic to represent input structure and continuous to represent nodes; (2) use symbolic to form loss functions; (3) use symbolic structure to enforce constraints over continuous; (4) convert continuous to symbolic to show a user / edit / perform symbolic inference later; (5) combine graph embeddings with text embeddings; (6) reason with symbolic, compute with continuous

8.3 Multi-Document Information Validation

- Research challenge: given a corpus, derive the probability for a claim to be true and present evidence/perspectives which rationalize the probability
- Research challenge: create realistic public dataset for fact checking
- Research challenge: claim validation in the medical domain
- Research challenge: claim validation annotation – existing data sets are either synthetic or have inconsistent annotations. How can we collect and annotate good data for this task?

8.4 Multi-Document User Decision Support Systems

- Research challenge: categorizing user intents, goals, and tasks
- Research challenge: Implementation issues
  - Selecting and consolidating the content (including communicating what the system is not showing/telling and supporting serendipity)
  - Learning from users (machine in the loop, including user intent refinement)
  - Explainability and sourcing
  - Supporting interaction from high level overview of repository to individual documents
- Research challenge: Evaluation
  - The NLP community should be open to a variety of evaluation methods for interactive tasks (automatic over corpus is not always feasible or best)
  - Some components may be susceptible to automatic evaluation

Participants

- Omri Abend
The Hebrew University of Jerusalem, IL
- Sebastian Arnold
Beuth Hochschule für Technik Berlin , DE
- Timothy Baldwin
The University of Melbourne, AU
- Jonathan Berant
Tel Aviv University, IL
- Giuseppe Carenini
University of British Columbia – Vancouver, CA
- Ido Dagan
Bar-Ilan University – Ramat Gan, IL
- Dipanjan Das
Google – New York, US
- Daniel Deutsch
University of Pennsylvania, US
- Laura Dietz
University of New Hampshire – Durham, US
- Yoav Goldberg
Bar-Ilan University – Ramat Gan, IL
- Dan Goldwasser
Purdue University – West Lafayette, US
- Iryna Gurevych
TU Darmstadt, DE
- Heng Ji
Rensselaer Polytechnic Institute – Troy, US
- Ayal Klein
Bar-Ilan University – Ramat Gan, IL
- Alexander Koller
Universität des Saarlandes, DE
- Chin-Yew Lin
Microsoft Research – Beijing, CN
- Fei Liu
University of Central Florida – Orlando, US
- Nafise Sadat Moosavi
TU Darmstadt, DE
- Barbara Plank
IT University of Copenhagen, DK
- Nils Reimers
TU Darmstadt, DE
- Dan Roth
University of Pennsylvania – Philadelphia, US
- Steve S. Skiena
Stony Brook University, US
- Gabriel Stanovsky
University of Washington – Seattle, US
- Amanda Stent
Bloomberg – New York, US
- Ivan Titov
University of Edinburgh, GB
- Kentaro Torisawa
NICT – Kyoto, JP
- Gisela Vallejo
TU Darmstadt, DE
- Andreas Vlachos
University of Cambridge, GB
- Yue Zhang
Westlake University – Hangzhou, CN